



# THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

# Learning Visually Grounded Meaning Representations

*Carina Silberer*



Doctor of Philosophy  
Institute for Language, Cognition and Computation  
School of Informatics  
University of Edinburgh  
2015

# Abstract

Humans possess a rich semantic knowledge of words and concepts which captures the perceivable physical properties of their real-world referents and their relations. Encoding this knowledge or some of its aspects is the goal of computational models of semantic representation and has been the subject of considerable research in cognitive science, natural language processing, and related areas. Existing models have placed emphasis on different aspects of meaning, depending ultimately on the task at hand. Typically, such models have been used in tasks addressing the simulation of behavioural phenomena, e.g., lexical priming or categorisation, as well as in natural language applications, such as information retrieval, document classification, or semantic role labelling. A major strand of research popular across disciplines focuses on models which induce semantic representations from text corpora. These models are based on the hypothesis that the meaning of words is established by their distributional relation to other words (Harris, 1954). Despite their widespread use, distributional models of word meaning have been criticised as ‘disembodied’ in that they are not *grounded* in perception and action (Perfetti, 1998; Barsalou, 1999; Glenberg and Kaschak, 2002). This lack of grounding contrasts with many experimental studies suggesting that meaning is acquired not only from exposure to the linguistic environment but also from our interaction with the physical world (Landau et al., 1998; Bornstein et al., 2004). This criticism has led to the emergence of new models aiming at inducing perceptually grounded semantic representations. Essentially, existing approaches learn meaning representations from multiple views corresponding to different modalities, i.e. linguistic and perceptual input. To approximate the perceptual modality, previous work has relied largely on semantic attributes collected from humans (e.g., is round, is sour), or on automatically extracted image features. Semantic attributes have a long-standing tradition in cognitive science and are thought to represent salient psychological aspects of word meaning including multisensory information. However, their elicitation from human subjects limits the scope of computational models to a small number of concepts for which attributes are available.

In this thesis, we present an approach which draws inspiration from the successful application of attribute classifiers in image classification, and represent images and the concepts depicted by them by automatically predicted visual attributes. To this end, we create a dataset comprising nearly 700K images and a taxonomy of 636 visual attributes and use it to train attribute classifiers. We show that their predictions can act as a substitute for human-produced attributes without any critical information

loss. In line with the attribute-based approximation of the visual modality, we represent the linguistic modality by textual attributes which we obtain with an off-the-shelf distributional model. Having first established this core contribution of a novel modelling framework for grounded meaning representations based on semantic attributes, we show that these can be integrated into existing approaches to perceptually grounded representations. We then introduce a model which is formulated as a stacked autoencoder (a variant of multilayer neural networks), which learns higher-level meaning representations by mapping words and images, represented by attributes, into a common embedding space. In contrast to most previous approaches to multimodal learning using different variants of deep networks and data sources, our model is defined at a finer level of granularity—it computes representations for individual words and is unique in its use of attributes as a means of representing the textual and visual modalities.

We evaluate the effectiveness of the representations learnt by our model by assessing its ability to account for human behaviour on three semantic tasks, namely word similarity, concept categorisation, and typicality of category members. With respect to the word similarity task, we focus on the model’s ability to capture similarity in both the meaning and appearance of the words’ referents. Since existing benchmark datasets on word similarity do not distinguish between these two dimensions and often contain abstract words, we create a new dataset in a large-scale experiment where participants are asked to give two ratings per word pair expressing their semantic and visual similarity, respectively. Experimental results show that our model learns meaningful representations which are more accurate than models based on individual modalities or different modality integration mechanisms. The presented model is furthermore able to predict textual attributes for new concepts given their visual attribute predictions only, which we demonstrate by comparing model output with human generated attributes. Finally, we show the model’s effectiveness in an image-based task on visual category learning, in which images are used as a stand-in for real-world objects.



# Lay Summary

Humans possess a rich knowledge of words and their meaning. Such knowledge includes, among other things, the perceivable physical properties (e.g., visual appearance) of the real-world objects to which the words refer and how these relate to each other. It enables us to recognise objects by means of our senses, to interact with them and to say something about them. An extensive amount of work in cognition research has been devoted to explaining the complex phenomena related to learning, mentally representing and processing aspects of this knowledge. Different classes of models of representations of words (in the form of, e.g., lists of continuous numbers) have been proposed. Typically, such models have been evaluated as to how well they can provide an account for human behaviour on specific tasks (e.g., categorisation, the grouping of different objects or words into categories). From a practical perspective, meaning representations are furthermore crucial for many natural language applications, such as information retrieval or document classification.

A major class of models automatically construct representations from text corpora. These models represent words by their relation to other words, based on the hypothesis that words which appear in similar linguistic contexts tend to have similar meanings. For example, the words *peel* and *cut* are often mentioned together with *apple*, *onion*, *potato*, etc. Despite their widespread use, this class of models of word meaning has been criticised for its exclusive reliance on text data, which contrasts with many experimental studies suggesting that humans learn the meaning of words not only from exposure to language but also from their interaction with the physical world (e.g., by means of their visual or olfactory senses). The criticism has led to the emergence of new models aiming at inducing *perceptually grounded* meaning representations. Essentially, existing approaches rely on multiple views corresponding to different modalities, i.e. linguistic and perceptual input (e.g., text and images), and combine these modalities into a joint representation. To approximate the perceptual modality, previous work has largely used semantic attributes collected from humans (e.g., is round, is sour), or abstract features automatically extracted from images. Using semantic attributes to represent word meaning is appealing—they have a long-standing tradition in cognitive science and are thought to represent salient psychological aspects of word meaning including multisensory information. However, their elicitation from human subjects is time-consuming and limits the scope of computational models to a small number of words for which attributes are available.

In this thesis, we present an approach to learning visually grounded meaning representations, in which we draw inspiration from computer vision research and represent images and the objects depicted by them by *automatically* obtained visual attributes (e.g., an image of an *apple* could evoke the attributes round, green, has a stalk, etc.). To this end, we create a dataset comprising images labelled with visual attributes and use it to train a system which, given a new image, predicts the absence or presence of attributes in the image. In line with the attribute-based approximation of the visual modality, we represent words in the linguistic modality by textual attributes (e.g., *fruit*, *harvest*, etc. for the word *apple*) which we obtain with an off-the-shelf text-processing system. We then introduce a novel model which learns meaning representations by simultaneously mapping words and images, represented by attributes, into a single representation. In contrast to previous approaches to multimodal learning, our model is unique in its use of semantic attributes as a means of representing the textual and visual modalities.

We present qualitative and quantitative results of our representation model in terms of its ability to simulate human behaviour on different semantic tasks, including word similarity (i.e., rating the similarity between two words on an ordinal scale) and categorisation. With respect to the word similarity task, we focus on the model's ability to capture similarity in both the meaning and appearance of the objects the words refer to. For this purpose, we create a new dataset in an experiment where participants are asked to give two ratings per word pair expressing their semantic and visual similarity. We furthermore demonstrate that the presented model is able to predict textual attributes for new objects given their automatically obtained visual attributes only. Finally, we show the model's effectiveness in an image-based task on category learning.

# Acknowledgements

First and foremost I am grateful to my supervisors Mirella Lapata and Vittorio Ferrari. Without them, the present thesis would not exist. Mirella's support, feedback and enthusiasm over the last few years have been invaluable for the thesis and my development as a researcher. From her I learned, inter alia, the importance of keeping the big picture and key questions in mind.

I furthermore thank my thesis examiners, Hinrich Schütze and Victor Lavrenko.

A great thanks goes also to my colleagues and friends from ILCC, especially to my fellow office mates, Siva, Lea, Hadi, and Yannis, and to Herman, Spandana, Bharat, Annie, Dominikus, Des, Zhengshuai, etc.

I would also like to thank my family and Michael.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

*(Carina Silberer)*

# Table of Contents

<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation and Central Claims . . . . .	1
1.2 Contributions . . . . .	3
1.3 Terminology and Notation . . . . .	4
1.4 Evaluation Methodology . . . . .	6
1.5 Thesis Structure . . . . .	8
1.6 Published Work . . . . .	10
<b>2 Background: Meaning Representations</b>	<b>11</b>
2.1 Attribute-based Models . . . . .	11
2.2 Distributional Lexical Semantics . . . . .	13
2.2.1 Vector Space Models . . . . .	13
2.2.2 Generative Latent Variable Models . . . . .	19
2.3 Distributed Lexical Semantics . . . . .	19
2.4 Discussion . . . . .	22
2.5 Grounded Models of Lexical Semantics . . . . .	23
2.5.1 Sources of perceptual information . . . . .	23
2.5.2 Integration mechanism . . . . .	25
2.6 Conclusions . . . . .	28
<b>3 Grounded Models Using Human Input</b>	<b>29</b>
3.1 Semantic Attribute Production Norms as a Proxy for Perceptual Information . . . . .	30
3.2 Image Labels as a Proxy for Perceptual Information . . . . .	31

3.3	Models . . . . .	33
3.3.1	Attribute-topic Model . . . . .	34
3.3.2	Global Similarity Model . . . . .	36
3.3.3	Canonical Correlation Analysis . . . . .	38
3.3.4	Discussion . . . . .	39
3.4	Experiments . . . . .	40
3.4.1	Experiment 1: Perceptual Information from Attribute Norms . . . . .	40
3.4.2	Experiment 2: Feature Engineering Attribute Norms . . . . .	47
3.4.3	Experiment 3: Visual Information from Image Labels . . . . .	49
3.5	Conclusions . . . . .	51
<b>4</b>	<b>Attribute-centric Representation</b>	<b>54</b>
4.1	Motivation for (Visual) Attributes . . . . .	55
4.2	Visual Attributes from Images . . . . .	56
4.2.1	Visual Attributes in Computer Vision . . . . .	56
4.2.2	Image Collections . . . . .	58
4.2.3	The Visual Attributes Dataset (VISA) . . . . .	60
4.2.4	Automatically Extracting Visual Attributes . . . . .	64
4.2.5	Deriving Visual Representations of Concepts . . . . .	68
4.3	Textual Attributes . . . . .	69
4.3.1	Textual Attributes from Strudel . . . . .	71
4.4	Experiment 4: Grounding Lexical Models with Attributes . . . . .	73
4.4.1	Data . . . . .	74
4.4.2	Evaluation Task . . . . .	75
4.4.3	Model Parameters . . . . .	75
4.4.4	Results & Discussion . . . . .	76
4.5	Conclusions . . . . .	79
<b>5</b>	<b>Visually Grounded Semantic Representations with Autoencoders</b>	<b>81</b>
5.1	Deep Learning in Artificial Neural Networks . . . . .	83
5.1.1	(Deep) Neural Networks . . . . .	83
5.1.2	Multimodal Deep Learning . . . . .	85
5.2	Autoencoders . . . . .	87
5.2.1	Basic Autoencoders . . . . .	87
5.2.2	Denoising Autoencoders . . . . .	88
5.2.3	Stacked Autoencoders . . . . .	89

5.3	Grounded Semantic Representations with Autoencoders . . . . .	90
5.3.1	Architecture . . . . .	90
5.3.2	Model Details . . . . .	92
5.3.3	Model Properties . . . . .	94
5.4	Conclusions . . . . .	95
<b>6</b>	<b>Experiments: Simulating Human Behaviour in Cognitive Tasks</b>	<b>96</b>
6.1	Experiment 5: Word Similarity . . . . .	97
6.1.1	Elicitation of Evaluation Dataset . . . . .	97
6.1.2	Experimental Setup . . . . .	99
6.1.3	Results and Discussion . . . . .	103
6.2	Experiment 6: Concept Categorisation . . . . .	105
6.2.1	Experimental Setup . . . . .	105
6.2.2	Results and Discussion . . . . .	108
6.3	Experiment 7: Typicality Ratings . . . . .	108
6.3.1	Experimental Setup . . . . .	109
6.3.2	Results and Discussion . . . . .	110
6.4	Conclusions . . . . .	111
<b>7</b>	<b>Image-related Tasks</b>	<b>113</b>
7.1	Experiment 8: Generation of Attributes . . . . .	113
7.1.1	Data . . . . .	114
7.1.2	Visual Attribute Generation . . . . .	115
7.1.3	Textual Attribute Generation . . . . .	116
7.1.4	Evaluation measures . . . . .	118
7.1.5	Results and Discussion . . . . .	118
7.2	Experiment 9: Visual Category Learning . . . . .	120
7.2.1	Visual Category Learning . . . . .	122
7.2.2	Method . . . . .	124
7.2.3	Experimental Setup . . . . .	127
7.2.4	Results . . . . .	129
7.3	Conclusions . . . . .	133
<b>8</b>	<b>Conclusions</b>	<b>135</b>
8.1	Main Findings . . . . .	135
8.2	Future Work . . . . .	136

<b>A</b>	<b>VisA Dataset</b>	<b>140</b>
A.1	Concepts and Synsets in VisA . . . . .	140
A.2	Annotation interface . . . . .	153
<b>B</b>	<b>Instructions to Participants in Word Similarity Study</b>	<b>154</b>
	<b>Bibliography</b>	<b>158</b>



# List of Figures

2.1	Example: Word-document matrix and vectors . . . . .	15
2.2	Example: Textual context snippets of words . . . . .	16
2.3	Example: Word-word matrix and vectors . . . . .	17
2.4	Example: Basic (artificial neural) network . . . . .	20
2.5	Illustration of different integration mechanisms . . . . .	26
3.1	Example: Attribute norms of McRae et al. (2005) . . . . .	31
3.2	Example: Images and labels in ESP and LabelMe . . . . .	34
3.3	Plate diagram for model of Andrews et al. (2009) . . . . .	35
3.4	Example: Representation with model of Andrews et al. (2009) . . . . .	36
3.5	Example: Representation with model of Johns and Jones (2012) . . . . .	37
3.6	Example: Representation with CCA model . . . . .	39
4.1	Example: WordNet/ImageNet subnetwork . . . . .	59
4.2	Example: Images from ImageNet with bounding box annotations . . . . .	60
4.3	Illustration: Attribute classes contained in VISA dataset . . . . .	61
4.4	Example: Attribute predictions for concepts encountered during training . . . . .	67
4.5	Example: Attribute predictions for concepts unseen during training . . . . .	68
4.6	Results: Precision-Recall curve for attribute classifiers . . . . .	69
4.7	Illustration: Construction of visual representations on the basis of visual attribute predictions . . . . .	70
4.8	Illustration: Construction of textual vector representations . . . . .	73
5.1	Illustration: Architecture of a basic autoencoder . . . . .	87
5.2	Illustration: Architecture of a denoising autoencoder . . . . .	89
5.3	Illustration: Architecture of SAE model . . . . .	91
7.1	Results: Comparison of predicted visual attributes and CSLB norms . . . . .	116
7.2	Results: Comparison of predicted textual attributes and CSLB norms . . . . .	117

7.3	Example: Variants in object reference . . . . .	121
7.4	Example: Task on visual category learning . . . . .	123
7.5	Illustration: Computation of category-based representation . . . . .	125
7.6	Illustration: Model used for task on visual category learning . . . . .	128
7.7	Results: Visual category learning. Effectiveness of models for different levels of categorisation . . . . .	133
A.1	Annotation interface for VisA. . . . .	153

# List of Tables

3.1	Statistics of ESP and LabelMe . . . . .	33
3.2	Results: Model comparison on Nelson et al. (1998) Models augmented with attribute norms . . . . .	43
3.3	Results: Model comparison on Nelson et al. (1998) Models augmented with inferred attributes . . . . .	44
3.4	Results: Model comparison on attribute inference . . . . .	46
3.5	Results: Model comparison on Finkelstein et al. (2002) Models augmented with inferred attributes . . . . .	46
3.6	Results: Model comparison on Nelson et al. (1998) Models augmented with visual vs. non-visual attributes . . . . .	48
3.7	Results: Model comparison on Nelson et al. (1998) Models augmented with image labels . . . . .	50
4.2	Example: Human-authored attributes for four concepts . . . . .	62
4.3	Example: Seven concepts and their eight most similar concepts on the basis of their visual, textual and bimodal representations . . . . .	71
4.4	Example: Extracted textual attributes for four concepts . . . . .	72
4.5	Results: Model comparison on Nelson et al. (1998) for seen concepts. Models augmented with automatically extracted visual and textual at- tributes . . . . .	76
4.6	Results: Model comparison on Nelson et al. (1998) for unseen con- cepts. Models augmented with automatically extracted visual and tex- tual attributes . . . . .	77
4.7	Results: Model comparison on Nelson et al. (1998). Models aug- mented with automatically extracted visual and textual attributes . . .	77
4.8	Results: Effectiveness of the McRae norms on Nelson et al. (1998) . .	78
5.1	Example: Attribute-based input to SAE model . . . . .	90

6.1	Example: Human-produced semantic and visual similarity ratings . . .	98
6.2	Results: Effectiveness of SAE model on word similarity task . . . . .	103
6.3	Example: Word pairs with highest semantic and visual similarity in SAE model . . . . .	104
6.4	Example: Clusters produced with SAE model . . . . .	106
6.5	Example: Clusters produced with visual SAE model . . . . .	107
6.6	Example: Clusters produced with textual SAE model . . . . .	107
6.7	Results: Model comparison on concept categorisation . . . . .	108
6.8	Results: Model comparison on typicality task . . . . .	110
6.9	Results: Comparison of SAE model and O'Connor et al. (2009) . . .	111
7.1	Example: Concepts and their visual attributes . . . . .	114
7.2	Example: Comparison of gold and generated attributes for <i>leek</i> . . . .	115
7.3	Example: Concepts and their textual attributes (inferred vs. extracted)	119
7.4	Results: Model comparison on visual category learning . . . . .	130
7.5	Results: Visual category learning. Distinction between known and un- known objects in the VISA dataset . . . . .	132
A.1	List of concepts, their WordNet synset IDs, and the corresponding synsets. . . . .	152

# Chapter 1

## Introduction

This thesis addresses the problem of grounding lexical meaning representations in the visual world. In this chapter we present the motivation for studying this problem, lay down the central claims of the thesis and give an overview of its structure.

### 1.1 Motivation and Central Claims

Humans generally possess a rich semantic knowledge of words and concepts. Such knowledge represents words' real-world referents and their perceivable physical properties (e.g., visual appearance), as well as how these interact and relate to each other. It is this knowledge that enables us to recognise objects and entities by means of our senses, to interact with them and to verbally convey information about them (McRae and Jones, 2013). An extensive amount of work in cognition research has been devoted to approaches and theories that explain the complex phenomena related to learning, representing and processing aspects of this knowledge. The ongoing debate over the properties of mental lexical representations, which underlie the understanding of linguistic phenomena, has given rise to different classes of models of meaning representations. From a practical perspective, meaning representations are crucial for many natural language applications (Turney and Pantel, 2010), which spurred research on models for automatic representation learning. Practical advantages of such models of meaning are, for example, that the same model can be used for different applications and can be adapted to specific problems (e.g., Landauer and Dumais, 1997; Collobert et al., 2011).

A well known class of such models automatically constructs representations from text corpora. They represent words by their relation to other words, based on the *dis-*

*tributonal hypothesis* (Harris, 1954) postulating that words which appear in similar linguistic contexts tend to have similar meanings. For example, the words *peel* and *cut* are collocates of *apple*, *onion*, *potato*, and *carrot*. Applications in which corpus-based models have been successfully used include document classification (Klementiev et al., 2012; Sebastiani, 2002), information retrieval (Manning et al., 2008), word sense discrimination (Schütze, 1998), frame-semantic role labelling (Roth and Lapata, 2015), and language modelling (Bengio et al., 2003). They have gained popularity in cognitive science being considerably successful at simulating human behaviour in various tasks (e.g., semantic priming, Lund and Burgess, 1996; Landauer and Dumais, 1997, or synonym selection, Bullinaria and Levy, 2012; Padó and Lapata, 2007).

There is a clear analogy between modelling semantics on the basis of text data and human acquisition of knowledge through exposure to linguistic input. However, many experimental studies suggest that word meaning is acquired not only from exposure to the linguistic environment but also from our interaction with the physical world (Landau et al., 1998; Bornstein et al., 2004). Beyond language acquisition, there is considerable evidence across both behavioural experiments and neuroimaging studies that the perceptual associates of words play an important role in language processing (for a review see Barsalou, 2008). It is for these reasons that, despite their widespread use, corpus-based models have been criticised as “disembodied” in that they are not *grounded* in perception and action (Perfetti, 1998; Barsalou, 1999; Glenberg and Kaschak, 2002).

In contrast, numerous theories and models in cognitive science are based on representations involving semantic attributes (McRae et al., 2005; Vinson and Vigliocco, 2008; Cree et al., 1999; Vigliocco et al., 2004) which represent perceived physical and functional properties associated with the referents of words. For example, *apples* are typically green or red, round, shiny, smooth, crunchy, tasty, and so on; *dogs* have four legs and bark, whereas *chairs* are used for sitting. However, these attributes are not obtained automatically but are either hand-coded or elicited from humans (e.g., the attribute norms from McRae et al., 2005), limiting the scope and applicability of computational models based on them.

The present thesis addresses the criticism on corpus-based models and sets forth an approach to ground meaning representations in the visual world by leveraging textual and visual information. In other words, the objective of the thesis is to derive and study *bimodal* or *visually grounded* representations of concepts. Our central claims are therefore:

**Integration Hypothesis:** The integration of visual and text-based information of concrete concepts yields meaning representations which more closely approximate the conceptual knowledge humans possess than purely text-based models. We test this hypothesis by assessing the ability of bimodal meaning representations to account for human behaviour in cognitive tasks, and compare their effectiveness with unimodal models.

**Attribute-based Representation:** The visual modality can be approximated by information which is rendered in natural language attributes and extracted from images. This underlies the assumption that we can use images as a stand-in for concrete concepts (objects). We examine this claim by assessing whether models leveraging such visual information (a) can account better for human behaviour than purely corpus-based models on a task which taps into the ability to judge the visual similarity of concrete concepts, and (b) are useful for inferring knowledge of new concrete concepts in cases where only images depicting those concepts are available.

**Joint Models:** The visual and textual modalities are interrelated and it is therefore beneficial to use *joint* integration methods which derive bimodal meaning representations by finding and exploiting intermodal associations. We test this hypothesis firstly by experimentally comparing different modality integration mechanisms. Secondly, if the hypothesis is true, it should be possible to infer some aspects of one modality from the other. We introduce a model which learns visually grounded representations by considering the two modalities in concert, and test the second claim by showing that this model can infer linguistic information when presented only with visual information.

## 1.2 Contributions

This thesis makes the following contributions to the problem of visually grounding lexical meaning representations:

**Representation Framework** We propose an attribute-centric approach to representing perceptual information for the purpose of learning visually grounded meaning representations. Specifically, we automatically predict the presence or absence of visual attributes (e.g., made of wood, is furry) in images, and use this as visual representation of concrete concepts. In doing so, we draw inspiration from computer vision research,

where the use of such natural language attributes to represent visual phenomena has experienced a growing interest. To the best of our knowledge, we are the first to use them in computational models of semantic representation.

**Bimodal Modelling of Word Meaning** We introduce a novel model for visual grounding which draws elements from connectionist, attribute-based, and distributional models of semantic memory. Our model is formulated as a neural network architecture that induces word representations by mapping linguistic and visual input into a common bimodal space. Both input modalities are rendered in attributes, where visual attribute information is obtained as outlined above, and linguistic attributes are extracted from texts using an off-the-shelf distributional approach. We show that the model, firstly, can account for human behaviour on tasks related to word similarity. Secondly, that it can infer explicit textual information when only given visual information as input. Thirdly, we show that the bimodal model yields representations useful for a visual categorisation task when presented with images only.

**Datasets** We have created two new datasets, both publicly available,<sup>1</sup> which we hope will be useful for further progress in the development and evaluation of visually grounded meaning representations. The first dataset (VISA) contains visual attribute annotations for approximately 500 concrete concepts. Specifically, these concepts (listed in Appendix A.1) are represented in the image database ImageNet (Deng et al., 2009) and the attribute production norms of McRae et al. (2005). Our second dataset consists of semantic and visual similarity ratings for 7,576 concept pairs. Each concept is covered by VISA and occurs in approximately 30 pairs. We obtained the similarity ratings using Amazon Mechanical Turk.

### 1.3 Terminology and Notation

**Terminology (Words).** We follow the standard literature and use the term *word* to denote any sequence of non-delimiting symbols. Two identical sequences of non-delimiting symbols are occurrences (tokens) of the same word (type) (Dale et al., 2000). The distinction between types and tokens is crucial in the context of counting words in a corpus. For example, the sentence *They moved out of the flat, the girl points out.* contains 8 types and 10 tokens (punctuation not counted).

---

<sup>1</sup>The datasets are available at [homepages.inf.ed.ac.uk/s1151656/resources.html](http://homepages.inf.ed.ac.uk/s1151656/resources.html).



**Terminology (Concepts and Categories).** Unless otherwise stated, we will use the term *concept* to denote the mental representation (knowledge) of objects belonging to basic-level classes, such as *dog*, *table*, *car*. We will use the term *category* to refer to superordinate-level classes of objects, such as ANIMAL, FURNITURE, VEHICLE. Note that the standard notion of these terms is less restrictive in that concepts as mental representations are not bound to a specific level of abstraction, and likewise, categories refer to equivalent classes of objects of any level of abstraction (e.g., Murphy and Medin, 1985; Rosch et al., 1976). Concepts and categories are linguistically expressed through words in *italics* and SMALL CAPITALS, respectively.

**Terminology (Attributes and Norms).** By the term *attributes* we refer to semantic properties or characteristics of concepts (or categories), expressed by words which people would use to describe their meaning. Our definition of *attributes* is essentially equivalent to what is commonly known as *semantic features*<sup>2</sup> in the literature of cognitive science, where they have been used in numerous theories and models of knowledge representation in human cognition (see, e.g., Yee et al., 2013; McRae et al., 2005; Rosch et al., 1976, and the references therein). For example, Rosch et al. (1976) created representations corresponding to attribute lists which they collected from human subjects (e.g., the category BIRD is represented by has feathers, has wings, has beak, lays eggs, flies, etc.). In a study involving a large group of participants, McRae et al. (2005) elicited attribute *production norms*, i.e. attributes for concepts which are found to be commonly used by people to describe them. See Chapter 3 (Section 3.1) for details on these norms. For further details on attributes and their use in models of knowledge representation, see Chapter 2 (Section 2.1).

We focus on two types of attributes and, conversely to the literature in cognitive sciences, distinguish them according to the form (*modality*) in which they can be accessed. Attributes referring to visually discernible properties are called *visual*. Examples are *furry*, *has legs*, *eats*. Attributes referring to properties that can be mined from text data are called *textual* or *linguistic*. Examples are *a mammal*, *dies*, *gives birth*. Note that our definition of textual attributes does not explicitly exclude properties which are visually perceivable. For example, the knowledge that *dogs* have ears may also be

---

<sup>2</sup>In order to avoid confusion, we will use the term *feature* only to refer to a measurable (possibly abstract) property of a general object, as used in machine learning and pattern recognition. For example, if the object is an image, a feature is derived from pixels and may denote, e.g., an edge or an interest point.

inferred from text data. However, existing studies suggest that information derived from text corpora prevalently capture encyclopaedic, functional and discourse-related properties of concepts (e.g., Baroni et al., 2010; Andrews et al., 2009).

**Mathematical Notation.** We will be using the following mathematical notation throughout the thesis:

- Matrices are denoted by capital bold-face letters, e.g.,  $\mathbf{W}$ .  $\mathbf{W}_j$  indicates the  $j$ th row, and  $\mathbf{W}_{.k}$  the  $k$ th column of  $\mathbf{W}$ . The component of matrix  $\mathbf{W}$  at row  $j$  and column  $i$  is denoted by  $W_{ji}$ .
- Vectors are denoted by lower case bold-face letters, e.g.,  $\mathbf{x}$ . A specific vector  $i$  is indicated by superscript in parenthesis, e.g.,  $\mathbf{x}^{(i)}$ , or by  $\mathbf{x}_i$ . The  $j$ th component of  $\mathbf{x}$  is denoted by  $x_j$ .
- $\mathbf{W}^T$  denotes the transpose of matrix  $\mathbf{W}$ . Analogously,  $\mathbf{x}^T$  denotes the transpose of vector  $\mathbf{x}$ .
- The cardinality of a set  $A$  is denoted  $|A|$ .

## 1.4 Evaluation Methodology

Throughout the thesis we will quantitatively evaluate the models against humans, mostly on cognitive tasks related to word similarity. Similarity is generally viewed as fundamental to cognition, it is found to play a major role in, for example, problem solving (Bassok, 1990; Novick, 1990; Kolodner, 1993), categorisation (Medin and Schaffer, 1978; Nosofsky, 1986; Rosch and Mervis, 1975), memory retrieval (Hintzman, 1986), decision making (Medin et al., 1995), and inductive reasoning (Osherson et al., 1990). Understanding the cognitive processes which underlie the assessment of similarity has been a central goal in cognitive psychology research (see, e.g., Goldstone and Son, 2005, for an overview of major psychological models of similarity). The evaluation of models of semantic representations by their ability to account for phenomena associated with similarity has therefore a long history. Semantic similarity (or the more general notion of relatedness) of pairs or groups of words is also crucial for many practical applications, including automatic thesauri creation (Grefenstette, 1994), information retrieval (Xu and Croft, 2000), word sense disambiguation (Yarowsky, 1992) and dis-

crimination (Pantel and Lin, 2002), metonymy resolution (Nissim and Markert, 2003), etc.

With respect to the previously mentioned corpus-based vector representations, the key assumption which ties them closely to similarity is that spatial closeness between vectors estimate similarity. It is therefore common in language research to quantitatively evaluate them by their ability to directly predict the semantic similarity between words, or to account for linguistic phenomena which are found to be dependent on similarity, such as categorisation. In Section 1.4.1 we will briefly explain our evaluation methodology which follows the common practice.

Note, however, that predicting similarity well does not imply that a model of representations gives a complete account for human lexical knowledge. It does hence not imply either that it is generally beneficial for all applications using word representations, since individual applications may require a lexical model to capture additional or other aspects of word meaning (e.g., ambiguity).

### 1.4.1 Correlation Analysis

We explain our general evaluation procedure by means of the explicit word similarity task, which has become a standard experimental methodology in both natural language processing and cognition research (Resnik, 1995; Agirre et al., 2009; Finkelstein et al., 2002, *inter alia*). The methodology regarding other cognitive tasks (e.g., word association or categorisation) is analogous (see Chapter 3, Section 3.4, and Chapter 6 for more details).

We evaluate the vector-based models against human-produced semantic similarity ratings of a set of word pairs. Typically, a dataset of human ratings (a.k.a. *benchmark*) is elicited by presenting human subjects with word pairs and asking them to rate their semantic similarity along an ordinal scale. These judgements are then averaged over the subjects to obtain a single score (Rubenstein and Goodenough, 1965; Miller and Charles, 1991; Finkelstein et al., 2002). This elicitation methodology has been validated to produce reliable mean ratings in virtue of high correlations between independent studies on the same word pairs. For example, Finkelstein et al. (2002) and Resnik (1995) elicited ratings for Miller and Charles's (1991) pairs and reported a correlation of .95 and .96 with Miller and Charles, respectively.

For each word pair of the dataset, we estimate the similarity of the words by comparing the corresponding vectors using a (geometrical) metric (e.g., the cosine similarity,

Chapter 2, Section 2.2.1). Applying correlation analysis, we can then compare the similarity estimates with the human ratings. Regarding the choice of the correlation coefficient, Spearman’s rank correlation coefficient ( $\rho$ ) and Pearson’s product moment correlation coefficient ( $r$ ) have been used in the literature, depending on the specifics of the task and the benchmark dataset at hand. If not otherwise stated, we will apply Spearman’s  $\rho$ , since it is less sensitive to extreme values.

Note that the correlation coefficient is generally a poor evaluation measure, since it can be sensitive to heavy-tailed distributions. An extreme example is the pairwise comparison of all words in the vocabulary, since every word is highly dissimilar to the vast majority of words in the vocabulary. However, the selection of the word pairs for creating an evaluation dataset is usually conducted in such a way that the pairs cover a range in semantic distances. Ideally, this process leads to a balanced set of pairs and avoids the problem of a heavy-tailed list of ranked pairs.

## 1.5 Thesis Structure

We begin by summarising existing work on models of meaning representation (Chapter 2). Next, we present a comparative study of three different perceptually grounded models (Chapter 3). Chapter 4 describes our approach to representing visual information through attributes. Chapter 5 introduces our model for visually grounding lexical meaning representations. Subsequently, we present a range of experiments (Chapters 6 and 7) which evaluate our attribute-centric approach and our model, and conclude the thesis in Chapter 8.

**Chapter 2** reviews three major classes of computational models of semantic representation—attribute-based, distributional, and distributed models—and discusses their main differences. Furthermore, it gives an overview of existing work on perceptually grounded models of lexical semantics, and identifies two important criteria by which the models can be characterised: the source of perceptual information and the mechanism used to integrate perceptual and textual data.

**Chapter 3** experimentally compares three perceptually grounded distributional models. We focus on different modality integration mechanisms, and provide all models with the same linguistic and perceptual input data, where the latter consists of human produced information (attribute norms or image labels, respectively). Our experimen-

tal results (on word association and similarity) show that all models benefit from the integration of perceptual data. We find that joint models, which perform integration by exploiting the interrelationships between the perceptual and linguistic data, obtain a closer fit with human judgements compared to a concatenation approach. We furthermore find that visual attributes from norming data are a beneficial means to represent visual information and to *visually* ground meaning compared with image labels.

**Chapter 4** elaborates on the idea of using attribute-centric representations of visual information and presents our approach to obtain these automatically from images. We first introduce our database (VISA) which comprises a taxonomy of visual attributes, visual attribute annotations of real-world concepts, and a large set of images depicting these concepts. We explain how we use VISA to train classifiers which predict visual attribute occurrences in images. We describe how we leverage the classifiers' predictions to automatically obtain visual attribute-based representations of concepts, and show that these can be effectively integrated with textual attribute information (Baroni et al., 2010) to yield promising results.

**Chapter 5** draws on the findings of Chapter 3 that the modalities are preferably integrated in a joint manner, and presents our novel model for visually grounded lexical meaning representations. It applies deep learning techniques in a neural network architecture for modality integration, using our attribute-centric representation as input. Specifically, our model is based on stacked denoising autoencoders (SAE) and derives bimodal meaning representations by jointly mapping the visual and linguistic modalities to a common hidden space.

**Chapter 6** assesses the effectiveness of our bimodal SAE model to simulate human behaviour in cognitive tasks related to concept similarity. To this end, we evaluate the SAE against human judgements on concept similarity, categorisation and typicality. Before presenting our experiments, we describe how we collected ratings for the first task using Amazon Mechanical Turk (AMT), where the annotators were asked to judge both semantic and *visual* similarity of concept pairs. The experimental results show that our model is better across all three tasks in accounting for human behaviour than comparison models (the model's unimodal variants and related bimodal models), and more effective in almost all cases than a purely text-based neural network model.

**Chapter 7** examines the benefit of our approach for two image-related tasks. We first demonstrate the ability of our attribute classifiers to generalise to unseen concepts. For this purpose, we apply the classifiers on images depicting concepts unknown to VISA, and compare their predictions against human generated normed attributes (Devereux et al., 2013). Similarly, we show the ability of our SAE model to infer textual attributes when only given visual input (the aforementioned attribute predictions), again comparing against normed attributes. Our second task evaluates the SAE on visual concept learning, where the task requires generalisation to a category from images depicting concrete concepts (Jia et al., 2013). Specifically, the goal is to decide for each of a series of concepts (e.g., *white fox*, *leatherback turtle*), whether it belongs to a given category (e.g., CARNIVORE). The category is hereby defined by a set of images depicting real-world concepts, and, likewise, a concept is represented by an image.

**Chapter 8** concludes the thesis with a summary of the main findings in light of our claims, and highlights avenues for further research.

## 1.6 Published Work

Some of the work presented in this thesis has been published previously. Most of the material in Chapter 3 is covered in Silberer and Lapata (2012). Chapter 4 is an elaboration of Silberer et al. (2013). The model described in Chapter 5 and its evaluation, Experiments 5 and 6 in Chapter 6, are presented in Silberer and Lapata (2014).

# Chapter 2

## Background: Meaning Representations

Humans possess a rich semantic knowledge of words and concepts. It captures the perceivable physical properties of their real-world referents, such as their visual appearance, their behaviour, and the relations that hold between them, including their interaction with each other. It is this knowledge that enables us to recognise objects and entities by means of our senses, to interact with them and to verbally convey information about them (McRae and Jones, 2013). Encoding this knowledge or some of its aspects is the goal of computational models of semantic representations. Existing models have placed emphasis on different aspects of meaning, depending ultimately on the task at hand.

The approach to semantic representation we present in this work, in particular in Chapters 4 and 5, exhibits characteristics from three different types of models which we review in this chapter: attribute-based (Section 2.1), distributional (Section 2.2), and distributed models (Section 2.3). In Section 2.5, we discuss recent approaches to modelling word meaning which aim to derive representations which are *grounded* in perception.

### 2.1 Attribute-based Models

A long-standing tradition in cognitive science is the assumption that meaning representations are based on attributes (e.g., Mervis and Rosch, 1981; Sloman et al., 1998). These are human-produced natural language properties, such as *is-a tree*, *has bark*, *is green*, *grows*, and typically encode knowledge of concepts with respect to their tax-

onomic relations to other concepts (hyponymy – is-a; meronymy – has-a), to their sensory properties (visual, acoustic, etc.) as well as their motoric characteristics as components of actions (behaviour). Attribute-based theories of lexical semantic representation<sup>1</sup> (Cree et al., 1999; Vigliocco et al., 2004; Jones et al., 2015, inter alia) use such attributes to computationally model phenomena of human cognition, e.g., categorisation and lexical priming.

Traditionally, attribute-based representations have been either directly hand-coded by the researchers, or induced in distributed models using the attributes as knowledge source. Goal of the latter is to investigate the mechanisms which underlie the learning of representations in the first place as well as to examine the interplay of the representations with other cognitive processes (e.g., Rogers and McClelland, 2004, but see Section 2.3 for details on distributed models). Classical examples of the former, hand-coded models, are Collins and Loftus (1975) and Smith et al. (1974). Collins and Loftus (1975) represent semantic knowledge in a network, where each concept (referred to by a content word of any part-of-speech) is represented by a single node, and nodes are connected via edges corresponding to the attributes that hold between them (e.g., *cherries*–is–*red*). The edges are labelled with weights indicating the importance of an attribute for a concept. Knowledge is accessed by spreading node activations through the network, starting with the nodes of the concepts in question and following the edges in decreasing order of their weights. The model of Smith et al. (1974) represents concepts as sets of attributes of two types: defining attributes common to all concepts of a (super-ordinate) category (e.g., *has wings*), and characteristic attributes essential for a particular concept (e.g., *flies*). Semantic processing is performed by computing the intersection of the sets of any two concepts. The model was shown to account for human behaviour on categorisation. This was assessed by comparing the reaction times of humans for concept-category pairs with model produced typicality ratings. An issue with models using experimenter-generated attributes as those mentioned above is that these were defined particularly for specific models and are thus prone to a lack of generality and psychological validity.

Modern attribute-based models use data collected in attribute norming studies, in which humans are presented with a series of words and asked to list relevant attributes of the things to which the words refer (Vinson and Vigliocco, 2008; Devereux et al., 2013; McRae et al., 2005; see also Section 3.1 for more details on the latter). Such

---

<sup>1</sup>In the context of semantic representations, attributes are often called features or properties in the literature. For the sake of consistency of the present work, we will adhere to the former term.



attribute norms<sup>2</sup> capture knowledge in addition to that used in the aforementioned methods, such as functional knowledge in the sense of the actions one can perform with objects, or go beyond objects and include events (Vinson and Vigliocco, 2008). Even though not undisputed, attribute norms are widely regarded as proxy for sensorimotor experience. They provide a cue to aspects of human meaning representations which have developed through interaction with the physical environment (McRae et al., 2005), and are used to verbally convey perceptual and sensorimotor information (e.g., *is yellow, smells bad, used by twisting*).

Researchers have used attribute norms to test theories and understand phenomena pertaining to the representation and processing of semantic knowledge, and to induce computational models of semantic representation (e.g., Grondin et al., 2009; Taylor et al., 2012), categorisation and its structure (Rosch and Mervis, 1975; Voorspoels et al., 2008; O'Connor et al., 2009), and category-specific disorders (Tyler and Moss, 2001; Rogers et al., 2004).

## 2.2 Distributional Lexical Semantics

In analogy with human acquisition of knowledge through exposure to linguistic input, distributional models of word meaning specify mechanisms for automatically constructing semantic representations from text corpora. They represent words by their relation to other words, based on the *distributional hypothesis* (Harris, 1954) postulating that words that appear in similar linguistic contexts tend to have related meanings. Distributional models thus clearly differ from attribute-based models presented in the previous section in that they capture information on how to *use* words.

### 2.2.1 Vector Space Models

A well known instance of distributional models are vector space models (VSMs). They are also termed semantic space models, as they typically represent words as points in a high-dimensional space. The components of the corresponding vectors encode the statistical distribution over some co-occurring contextual elements or features (e.g., Lund and Burgess, 1996; Padó and Lapata, 2007; Erk and Padó, 2008). Words (i.e. points) that are nearby in the space are assumed to be related in meaning, as they exhibit a

---

<sup>2</sup>They are often termed semantic feature production norms (e.g., McRae et al., 2005) or property norms (e.g., Devereux et al., 2013) in the literature.

similar contextual distribution. Analogously, points that are far away are assumed to be semantically dissimilar or unrelated.

VSMs have been successfully used in many natural language applications. Examples include information retrieval (Manning et al., 2008), document classification (Sebastiani, 2002), question answering (Tellex et al., 2003), information extraction (Paşca et al., 2006), semantic role labelling (Pennacchiotti et al., 2008), word sense discrimination (Schütze, 1998), word sense disambiguation (Padó and Lapata, 2007), thesaurus construction (Grefenstette, 1994), and many more (see Turney and Pantel, 2010). They have likewise gained popularity in cognitive science (e.g., LSA, Landauer and Dumais 1997; HAL, Lund and Burgess 1996), being considerably successful at simulating human behaviour in various tasks including semantic priming (Lund and Burgess, 1996; Landauer and Dumais, 1997; Padó and Lapata, 2007), deep dyslexia, text comprehension, synonym selection (Bullinaria and Levy, 2012; Landauer and Dumais, 1997; Padó and Lapata, 2007), word association, similarity judgements (Landauer and Dumais, 1997), and categorisation (Bullinaria and Levy, 2012) (see Griffiths et al., 2007b, and the references therein).

Typically, VSMs are constructed by analysing a text corpus and extracting the co-occurrence frequency of each target word with its contextual elements, such as context words or documents. Each target word is then represented as a vector whose components correspond to contextual elements and whose entries give their frequency of co-occurrence with the target word. These raw counts are subsequently turned into weights (e.g., using weighting schemes such as mutual information or *tf-idf*) that abstract from raw frequency counts and express the importance of the contextual elements for a target word. The vectors of all target words are generally stored as row vectors in a sparse matrix, representing the semantic space. An example is given in Figure 2.1 (on the left), where the contextual elements correspond to documents. The dimensionality of the space may further be reduced by means of an appropriate method, such as singular value decomposition (SVD, Golub and Reinsch, 1970).

Having constructed a VSM, we can mathematically compare the meaning of two words, for example by geometrically estimating their similarity, e.g., as the cosine of the angle (Deerwester et al., 1990) between their vectors. An example in a two-dimensional space is presented in Figure 2.1 (on the right), where the angle between the vectors of *cake* and *bread* is very small and the two words are therefore considered highly similar, as opposed to *cake* and *apple*.

Researchers have used various types of contextual elements, weighting schemes,

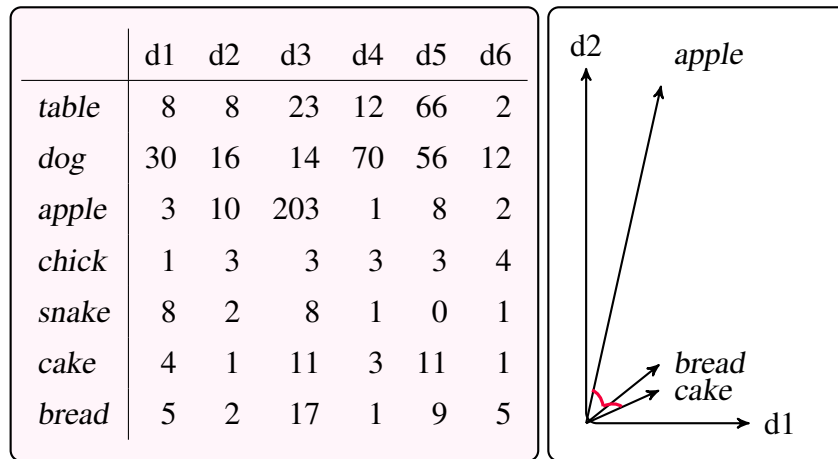


Figure 2.1: Left: Example of a word-document matrix extracted from a text corpus. Target words are represented by row vectors over documents they occur in (columns). Right: Representations of the words *cake*, *bread*, and *apple* in a two-dimensional semantic vector space. (Word counts extracted from British National Corpus (BNC).)

dimensionality reduction techniques, and comparison measures. We will give details to each of them below, as far as relevant for this thesis (see also Turney and Pantel, 2010, for a more detailed overview).

**VSM Types** Contextual elements can be text passages of an underlying corpus, e.g., paragraphs or whole documents (*word-document* matrices, e.g., Landauer and Dumais, 1997), context words (*word-word* matrices, e.g., Lund and Burgess, 1996; Bullinaria and Levy, 2012), or more sophisticated elements, such as syntactic dependencies (Padó and Lapata, 2007), selectional preferences (Erk and Padó, 2008), or semantic attributes (Baroni et al., 2010).

*Word-document matrices* are derived by counting the frequency with which each target word occurs in each document (or paragraph) of an underlying text corpus. An example for this matrix type is given in Figure 2.1 (left-hand side). *Word-word matrices*, in turn, represent words by other words (the context words) with which they co-occur in a text corpus. They are created by counting the frequency with which a context word occurs within a window of words surrounding the target word, aggregated over all occurrences of the target word in the text corpus. An example of context snippets is given in Figure 2.2, and Figure 2.3 (left-hand side) shows a word-word matrix.

Other approaches represent word meaning by means of *word-attribute matrices*.

... was sitting at the first *table* past the *kitchen*, finishing his lunch.  
 And without another word he left the *table*, went out of the *kitchen*,...  
 ... chef patissier of The Connaught, prepared the *apple* pastry *dessert*.  
 They finished with a *dessert* of nuts, honey and *apple*.  
 He went off into the *kitchen* and came back with an *apple* pie.  
 As the *chicks* *hatched*, they were moved into an out-building, ...  
 As soon as a domestic hen *chick* *hatches* it starts pecking at grains ...  
 Moreover, *snakes* never inject all their *venom* in a single strike, ...  
 ... the upas tree (Moraceae), or *venom* from toadskin or *snakes*.

Figure 2.2: Example context snippets for the target words *table*, *apple*, *chick*, and *snake*. Words highlighted in blue are examples of co-occurring words. Context snippets were extracted from the BNC.

They are similar to word-word matrices, with the difference that context words for a given target word are extracted in such a way that these can be interpreted as attributes of the latter. With reference to the previous section on attribute-based models (Section 2.1), these VSMs employ the distributional approach to word meaning as a means to *automatically* acquire attribute-based descriptions of words from text corpora. An example is the model by Baroni et al. (2010), whose approach is based on the co-occurrence counts of the target, its context words *and* the contextual elements linking them. We will describe the model in more detail in Chapter 4 (Section 4.3).

**Weighting Schemes** For word-document matrices, the *tf-idf* (term frequency–inverse document frequency) family of schemes for weighting co-occurring contextual elements is widely applied. We will use the following definition (see, e.g., Salton and Buckley, 1988, for alternatives):

$$\text{tf-idf}(w, d) = \text{tf}(w, d) \cdot \text{idf}(w) = \text{frequ}(w, d) \cdot \log \frac{N}{n}, \quad (2.1)$$

where  $\text{frequ}(w, d)$  denotes the frequency of occurrence of word  $w$  in document  $d$ ,  $N$  denotes the total number of documents in a corpus collection, and  $n$  is the number of documents containing  $w$ . The *idf*-factor leads to a higher weight for words that occur only in a few documents, and a lower weight for those occurring in many documents as these are supposedly not meaningful with respect to a specific document. Note that

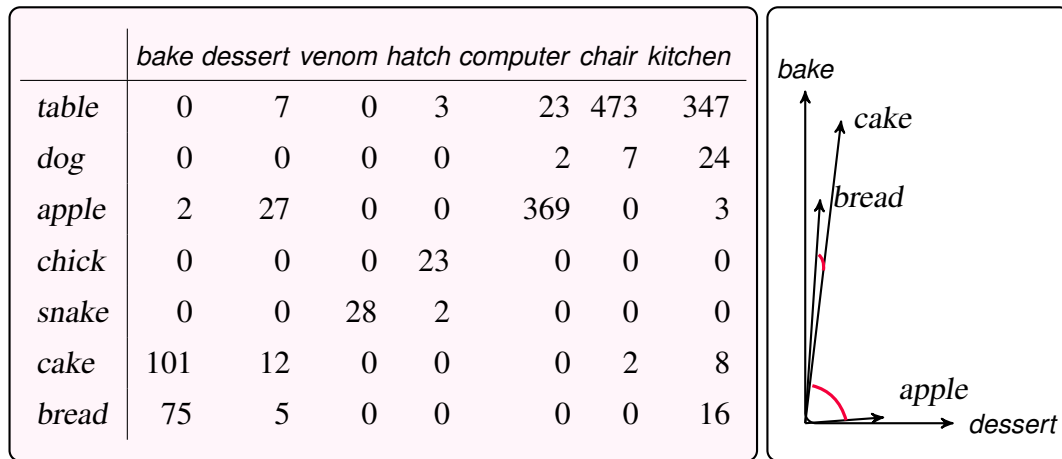


Figure 2.3: Left: Example of a word-word matrix extracted from a text corpus. Target words are represented by row vectors over co-occurring context words (columns). Right: Representations of the words *cake*, *bread*, and *apple* in a two-dimensional semantic vector space. (Word counts extracted from BNC.)

the idf-factor plays no role in that it is cancelled out when used in conjunction with the cosine similarity (see Equation (2.4)).

For word-word matrices, several weighting schemes have been used in the literature (see Bullinaria and Levy, 2007, for an overview). As an example, we give the *ratio* weighting below which proved to be a robust measure in Mitchell’s (2011, p. 45) experiments on similarity and synonymy tasks. It is defined as follows:

$$\text{ratio}(w_i|w_j) = \frac{P(w_i|w_j)}{P(w_i)} = \frac{\text{frequ}(w_i, w_j) \sum_{k=1}^K \text{frequ}(w_k)}{\text{frequ}(w_i) \text{frequ}(w_j)}, \quad (2.2)$$

where the probability  $P(w_i)$  of word  $w_i$  denotes the maximum likelihood estimate of its occurrence frequency in the dataset, the conditional probability  $P(w_i|w_j)$  of  $w_i$  given word  $w_j$  is estimated by the frequency with which the words co-occur, and  $K$  denotes the total number of considered context words (i.e. the columns of the word-word matrix).

**Dimensionality Reduction Techniques** Singular value decomposition (SVD, Golub and Reinsch, 1970) is a standard mathematical technique for reducing the dimensionality of semantic spaces, which has particularly proven useful for uncovering latent (i.e. hidden) semantic structure of text data (Deerwester et al., 1990; Landauer and Dumais, 1997).

Let  $\mathbf{M}$  be a real matrix with  $n$  rows and  $m$  columns, i.e.  $\mathbf{M} \in \mathbb{R}^{n \times m}$ . SVD computes a factorisation of  $\mathbf{M}$  into three component matrices, i.e.

$$\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T, \quad (2.3)$$

where  $\mathbf{U} \in \mathbb{R}^{n \times n}$  and  $\mathbf{V}^T \in \mathbb{R}^{m \times m}$  are orthonormal matrices<sup>3</sup> and  $\Sigma \in \mathbb{R}^{n \times m}$  is a diagonal matrix of rank  $r$  with the singular values of  $\mathbf{M}$  on its main diagonal in descending order.

Landauer and Dumais (1997) used SVD for their latent semantic analysis (LSA) model of word meaning, in which a semantic space of dimension  $r \leq \min(n, m)$ , represented by a word-document matrix  $\mathbf{M}$ , is transformed to a (lower-dimensional) latent space of dimension  $d < r$ . This is accomplished by setting all but the first  $d$  singular values in  $\Sigma$ , which are the latent components explaining best the co-occurrence of words, to zero and re-multiplying the matrices.  $\mathbf{M}_d$  is then an approximation of  $\mathbf{M}$ , where similar words and documents, respectively, are now geometrically close to each other even if they have never co-occurred in the original space. Note, that  $\mathbf{M}_d \in \mathbb{R}^{n \times m}$  still applies.

SVD has since been applied to all types of VSMs, including word-word matrices (Bullinaria and Levy, 2012) and vector spaces whose word representations were constructed on the basis of text *and* image data (Bruni et al., 2014).

**Similarity Measures** We can quantify the similarity between two words,  $w_u$  and  $w_v$ , by calculating the similarity between their vector representations,  $\mathbf{u}$  and  $\mathbf{v}$ , using a variety of similarity measures (see Weeds et al., 2004; Turney and Pantel, 2010, for overviews). Throughout this thesis, we will be using the cosine similarity measure, which returns the cosine of the angle between the vectors:

$$\text{cosine}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|} = \sum_{i=1}^n \frac{u_i v_i}{\sqrt{\sum_{i=1}^n u_i^2} \sqrt{\sum_{i=1}^n v_i^2}} \quad (2.4)$$

We give a two-dimensional example in Figures 2.1 and 2.3 (right-hand side), where the word *cake* is estimated to be more similar to *bread* than to *apple* due to their vectors being geometrically closer in space.

---

<sup>3</sup>For an orthonormal matrix  $\mathbf{Q}$  it holds that  $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}$ , where  $\mathbf{I}$  is the identity matrix. That is, the columns and rows of  $\mathbf{Q}$  are orthogonal unit vectors.

## 2.2.2 Generative Latent Variable Models

Topic models for word meaning represent words in a probabilistic way, namely by a set of topics which are modelled by probability distributions over words. Latent Dirichlet allocation (LDA, Blei et al., 2003) is a well-known example of such topic models. This hierarchical Bayesian model views each document as a finite random mixture over a set of latent, i.e. unobserved, topics, and each topic as an infinite mixture over an underlying set of topic probabilities (Blei et al., 2003). LDA, akin to SVD, can be used as a means for reducing the dimensionality of a word-document matrix. In contrast to SVD, however, it is a generative model, specifying the creation of a collection of documents (represented by the word-document matrix) by inferring the latent variables that are most likely responsible for the observed data, i.e. the words of the documents.

The generative process is broken down into probabilistic steps as follows: In order to generate a collection of documents, a set of topics is generated by sampling a multinomial probability distribution  $\phi$  from a Dirichlet distribution. For each document  $d$  to be created, a distribution over topics  $\theta$  is sampled from a Dirichlet distribution. Then, for each of the words in  $d$  to be generated, a topic is drawn from  $\theta$ , and a word is finally drawn from the multinomial distribution over words associated with the topic. Learning of the latent variables that best fit the word-document matrix, (i.e. which explain best the generation of the document collection), is performed through Bayesian inference.

## 2.3 Distributed Lexical Semantics

Distributed models of word meaning specify mechanisms for learning representations corresponding to vectors of distributed patterns of activation across neuron-like units. The activation of a unit is caused by the weighted activations of other units it is connected to. Unlike the components in distributional or attribute-based models, the individual units do not usually correspond to interpretable features or words. Instead, a meaningful feature may be encoded by a distribution of activities across several units (Rogers and McClelland, 2004, p. 77).

Distributed representations with artificial neural networks as their most common type of architecture lie at the core of connectionist models (Hinton, 1981, 1986; Rumelhart et al., 1986a). The basic architecture of an artificial neural network (henceforth network) contains one *layer* of *input* units and one layer of *output* (or *target*) units

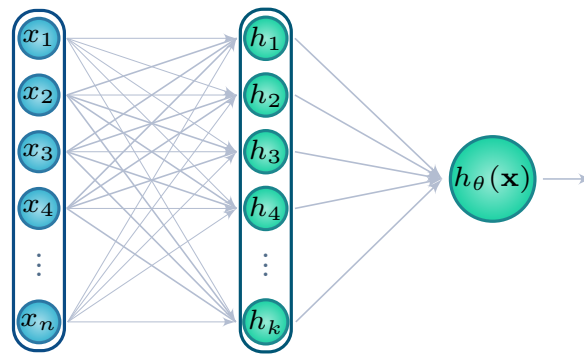


Figure 2.4: Example of a basic (artificial neural) network consisting of one input layer (with  $n$  units), one hidden layer (with  $k$  units), and one output layer (with 1 unit). The arrows represent the direction in which the activation of one unit is passed as input to another unit. Parameter  $\theta$  comprises all weight parameters to be learned.

which are often, but not necessarily, connected through internal units, referred to as *hidden* units (see Figure 2.4).

**Early Connectionist Models** Connectionist models have been extensively used to study and simulate phenomena of semantic knowledge as well as the consequences of its impairments and disorders by introducing damage to the architecture of the network (see Cree and Armstrong, 2012; Jones et al., 2015, *inter alia*, for a review). Many early network models were trained using hand-crafted features in the form of semantic attributes (e.g., `has_legs`) or labels (e.g., `cat`). These either explicitly correspond to the units of the semantic representations (e.g. Farah and McClelland, 1991), or they are given as input or target output to learn abstract distributed representations in the network’s hidden layer (e.g. Hinton and Shallice, 1991; Westermann and Mareschal, 2014). Other work (Tyler et al., 2000; McRae et al., 1997; Cree et al., 1999) does not rely on features specifically created for a network model, but instead makes use of semantic attributes empirically elicited in norming studies (see Section 3.1 for details on attribute norms). In the model by Rogers et al. (2004), the hidden semantic representation to be learned functions as connecting layer between both verbal units informed by attribute norms and visual units corresponding to visual properties of objects.

**(Deep) Neural Networks** As briefly discussed above, connectionist approaches have a long tradition in cognitive science (dating back to Hinton, 1981, 1986), but only recent achievements (LeCun et al., 1990; Hinton and Salakhutdinov, 2006; Hinton et al.,



2006; Ranzato et al., 2006) and advances in computer technology made learning in more complex network architectures feasible. This led to the emergence of a new research area in machine learning commonly referred to as *deep learning*, along with a surge/resurrection of interest in their application in natural language processing, computer vision and other disciplines (see, e.g., the NIPS 2013 and 2014 Workshops on Deep Learning<sup>4</sup> and Representation Learning<sup>5</sup>, the ICML 2013 Workshop on Deep Learning for Audio, Speech, and Language Processing<sup>6</sup>, and the ACL 2012 Tutorial on Deep Learning for NLP<sup>7</sup>.)

Since then, a variety of (deep and shallow) network architectures have been proposed that learn word representations corresponding to vectors of activation of network units, and are often referred to as *word embeddings*. A notable difference between these models and the early models outlined above is that learning is performed on the basis of unlabeled text corpora.

The model we propose in this thesis employs deep learning methods to learn visually grounded distributed word representations. We will give more details on deep learning along with the presentation of our model in Chapter 5.

Here, we will confine ourselves to briefly refer to models whose induced word representations have been leveraged in a wide range of NLP tasks (Collobert and Weston, 2008; Mnih and Hinton, 2009; Huang et al., 2012). These network models embed each word into a continuous space via an embedding matrix to be learned. Subsequent layers of the network use these embeddings by mapping them to a prediction of the target output. Typically, the embeddings are initialised randomly and then learned by backpropagating the derivative of the objective function (with respect to the network weights) through the network to the embedding matrix. Focusing on language modelling,<sup>8</sup> the network by Bengio et al. (2003) is trained to predict a probability distribution over the next word given preceding words, and, similarly, Mnih and Hinton's (2009) objective is to predict the embedding of the next word. Collobert and Weston (2008) introduced the use of a (pairwise) ranking criterion by which their network is trained to give a higher score to correct word sequences than to noisy ones, where noise was introduced by replacing the middle word of a sequence by a random word. Similar paradigms have been used by other neural network approaches to inducing word

---

<sup>4</sup><https://sites.google.com/site/deeplearningworkshopnips2013/>

<sup>5</sup><https://sites.google.com/site/deeplearningworkshopnips2014/>

<sup>6</sup><https://sites.google.com/site/deeplearningicml2013/>

<sup>7</sup><http://www.acl2012.org/program/tutorial2-2.asp>

<sup>8</sup>Neural network language models can be employed, e.g., during decoding for statistical machine translation (i.e. to ensure a fluent translation; Vaswani et al., 2013; Schwenk, 2007, 2010).

embeddings (e.g., Turian et al., 2010; Huang et al., 2012; Mikolov et al., 2013b), or to associating visual and linguistic data (e.g., Kiros et al., 2014b; Socher et al., 2013b, see Chapter 5). Collobert and Weston (2008) showed how thereby *pre-trained* word representations can be used to initialise a deep neural network architecture applicable to several NLP tasks, such as part-of-speech (PoS) tagging and semantic role labelling, and jointly optimised the network parameters (including those for the word embeddings) on the different tasks using labelled data (see also Collobert et al., 2011).

In later work, Mikolov et al. (2013a,b) present the *continuous skip-gram model* which has become one of the standard choices for NLP approaches leveraging word representations. We will give more details on this model in Chapter 6, where we compare it experimentally to our own model.

Word embeddings induced by the aforementioned methods have been used to initialise neural network models addressing, inter alia, sentiment analysis, semantic relation classification (Socher et al., 2012), parsing (Socher et al., 2013a), question answering (Iyyer et al., 2014), and image caption generation (Kiros et al., 2014b), or have served as word features in existing systems for, e.g., chunking, NER (Turian et al., 2010), dependency parsing (Bansal et al., 2014), or frame-semantic role labelling (Roth and Lapata, 2015). Moreover, they have been used directly as semantic representations for measuring relational similarity or answering analogy questions (e.g., Mikolov et al., 2013a,c; Levy and Goldberg, 2014).

These distributed models, however, usually cannot deal with out-of-vocabulary words. Some models adopt hybrid approaches of distributional and distributed methods and learn an encoding matrix which maps distributional word representations (Section 2.2) into a distributed space and which hence can be applied to encode new words (e.g., Bespalov et al., 2011). Similarly, our model presented in Chapter 5 learns distributed representations by leveraging attribute-based representations obtained with, inter alia, a distributional model.

## 2.4 Discussion

Connectionist models are in parts similar to semantic space models which arise from the application of mathematical techniques (e.g., singular value decomposition) and project word representations to a new space over latent variables. A key difference between the two types of models, however, is that connectionist representations are *learned* gradually by iteratively adjusting the weights that connect the units, whereas

representations of the latter type arise from a one-time computation (Rogers and McClelland, 2004, p. 77) applied onto count-based representations. Both, distributional models and modern network approaches applied to natural language processing, are based on the hypothesis that the meaning of words is determined by their relation to other words. Their induced word representations capture language use as a result of their induction from naturally occurring linguistic information, i.e. text corpora. Whether one type of the two corpus-based approaches is in general preferable over the other has been subject of recent studies (e.g., Lebet et al., 2013; Baroni et al., 2014; Lebet and Collobert, 2014; Levy et al., 2015). In contrast, attribute-based models capture taxonomic, sensorimotor and perceptual characteristics of concepts, by virtue of using human-produced information (hand-coded or empirically derived attributes) of concepts.

## 2.5 Grounded Models of Lexical Semantics

In the previous sections we discussed three major strands of models of lexical representations which focus on different aspects of word meaning. Despite their widespread use, corpus-based models have been criticised as “disembodied” in that they are not *grounded* in perception and action (Barsalou, 1999; Glenberg and Kaschak, 2002; Perfetti, 1998). This lack of grounding contrasts with many experimental studies suggesting that word meaning is acquired not only from exposure to the linguistic environment but also from our interaction with the physical world (Landau et al., 1998; Bornstein et al., 2004). Beyond language acquisition, there is considerable evidence across both behavioural experiments and neuroimaging studies that the perceptual associates of words play an important role in language processing (for a review see Barsalou, 2008).

In recent years, new types of models of word meaning have emerged that integrate both corpus-based (*textual*) and perceptual data in order to derive grounded representations. The models differ in terms of the source of perceptual information used as well as the methods that are applied for integrating different modalities. In this section we will give a review of these models.

### 2.5.1 Sources of perceptual information

Some models use attribute norms obtained in longitudinal elicitation studies (cf. Sections 2.1 and 3.1) as an approximation of the perceptual environment (Andrews et al.,

2009; Steyvers, 2010; Johns and Jones, 2012; Silberer and Lapata, 2012). Others focus on the visual modality as a major source of perceptual information and exploit image databases, such as ImageNet (Deng et al., 2009, see Section 4.2.2 for details) or ESP (von Ahn and Dabbish, 2004, see Section 3.2 for details). With a few exceptions that leverage human produced image labels (e.g., *red*, *dog*) as a proxy (Bruni et al., 2012a; Hill and Korhonen, 2014), most methods automatically extract visual information from images (Feng and Lapata, 2010; Silberer et al., 2013; Kiela and Bottou, 2014; Bruni et al., 2011, 2012a,b, 2014; Silberer and Lapata, 2014).

Feng and Lapata (2010) and Bruni et al. (2011, 2012a,b, 2014) use the bag-of-visual-words (BoVW) model to represent images as histograms over *visual words*. The BoVW model is an analogy to the bag-of-words approach in natural language processing, in which documents are represented by means of the words they contain.<sup>9</sup> A dictionary of visual words (codebook) is first derived by clustering feature descriptors extracted from a collection of images. An image can then be represented by the histogram of visual words, obtained by accumulating, for each cluster, the descriptors found in the image that are members of the particular cluster. With the purpose to introduce weak geometry, Bruni and colleagues partition an image uniformly into smaller regions and perform the BoVW approach on each image region instead of a whole image, applying spatial binning (Lazebnik et al., 2006). That is, an image is represented as the concatenation of histograms of visual words, where each histogram is derived from a spatial region of the image. Feng and Lapata (2010) and Bruni et al. (2011, 2012a,b, 2014) use SIFT (scale-invariant feature transform, Lowe, 2004) descriptors, which is an approach to extract descriptions of visual image features that are invariant to scaling and rotation and partially to changes in illumination and affine transformations. Bruni et al. (2012a) additionally use color space features as an alternative to SIFT descriptors. Kiela and Bottou (2014) follow the recent trend in image representation learning and learn distributed representations for the visual modality by means of convolutional neural networks, which directly operate on the pixel-level of images. Finally, combinations of different sources of perceptual information have also been suggested. Roller and Schulte im Walde (2013) combine visual information obtained with the BoVW model with data from norming studies, and Hill and Korhonen (2014) combine the latter with image labels.

With exception of the model by Feng and Lapata (2010), who extract all informa-

---

<sup>9</sup>Technically, the document representation as a bag of words is the transpose of a word-document matrix.

tion from a corpus of bimodal documents (i.e. BBC news articles and their associated images), data sources for the perceptual and textual modality in the aforementioned approaches are decoupled in that they are gathered independently and do not co-occur (i.e. linguistic data from text corpora and perceptual information from attribute norms or image databases).

In this thesis, we present another approach which draws inspiration from the successful application of attribute classifiers in object recognition, and represents images by visual attributes (e.g., `has_legs`, `is_round`; see Chapter 4). We will show that these (automatically predicted) attributes can act as substitutes for attribute norms in grounded models of semantic representation without any critical information loss (Chapters 4 and 6; Silberer et al., 2013; Silberer and Lapata, 2014).

## 2.5.2 Integration mechanism

Existing models can be broadly distinguished by the integration mechanism that is applied in order to obtain a perceptually grounded semantic space, and whether the process exploits the interrelationships between the modalities.

A shallow approach is to derive a new grounded space by concatenating the vectors (or matrices) corresponding to a word’s perceptual and linguistic representation (or to the two semantic spaces), respectively (Kiela and Bottou, 2014; Bruni et al., 2011, 2012a), without performing any further methods on the concatenation. For example, given two vectors  $\mathbf{v} = [a_1, a_2, \dots, a_n]$  and  $\mathbf{w} = [b_1, b_2, \dots, b_m]$ , their combined representation in the new space is  $[a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_m]$ . In the case of integration by concatenation, the components of the original unimodal spaces are directly adopted as the individual components that constitute the new grounded space. This mechanism is illustrated in Figure 2.5 (a). Bruni et al. (2011, 2012a) create an individual distributional space for the visual and linguistic modality, respectively, prior to their concatenation. Kiela and Bottou (2014) perform the same two-step approach, but use distributed representations for the modalities. Johns and Jones (2012) also concatenate distributional representations, but they focus on inferring missing perceptual information by leveraging the redundancies between the perceptual and linguistic modality prior to their concatenation. The approach is illustrated in Figure 2.5 (b). We discuss this model in more detail in Chapter 3 (Section 3.3.2).

Other approaches infer bimodal representations over latent variables responsible for the co-occurrence of words over featural dimensions. A model akin to Latent Se-

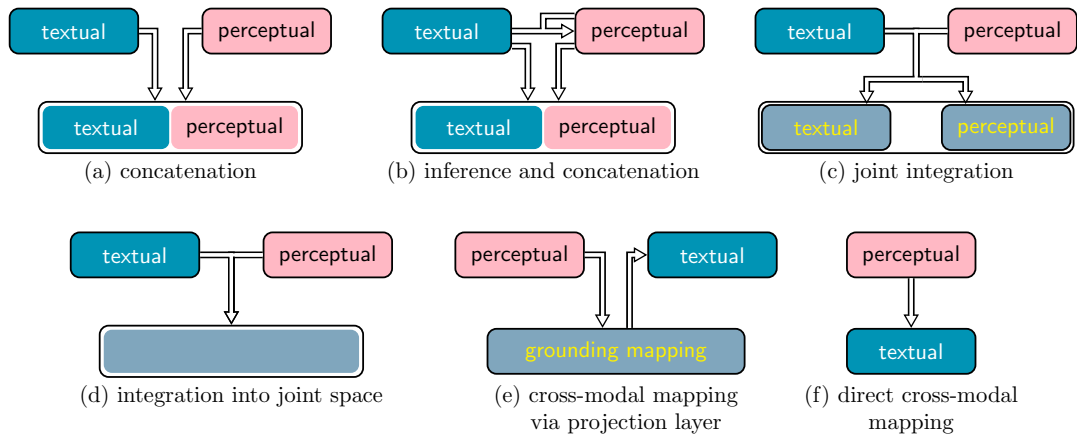


Figure 2.5: Illustration of different integration mechanisms.

mantic Analysis (Landauer and Dumais, 1997) is proposed by Bruni et al. (2012b, 2014) who concatenate the two matrices pertaining to a linguistic and visual VSM and subsequently project them onto a lower-dimensional space using SVD (Section 2.2.1). We will give a more detailed description of this model in Chapter 6. In the grounded space obtained by applying SVD, perceptual and linguistic components of the original matrices may now be collapsed to one latent variable. Yet, two matrices can be directly derived by means of the mappings  $\Sigma_d \mathbf{V}_L^T$  and  $\Sigma_d \mathbf{V}_V^T$  (see Section 2.2.1), consisting only of the columns corresponding to the linguistic and visual components, respectively. Namely by projecting the original matrix to a visually grounded linguistic space ( $\mathbf{U} \Sigma_d \mathbf{V}_L^T$ ) and a linguistically grounded visual space ( $\mathbf{U} \Sigma_d \mathbf{V}_V^T$ ), respectively. (See Figure 2.5 (c) for an illustration of this mechanism.)

Several models (Andrews et al., 2009; Steyvers, 2010; Feng and Lapata, 2010; Roller and Schulte im Walde, 2013) present an extension of latent Dirichlet allocation (LDA, Section 2.2.2). This type of perceptually grounded models treat both, words in documents and other perceptual units, as observed variables to learn topic distributions. The hereby inferred perceptually grounded representations of the words correspond to distributions over components (i.e. topics) which can not be attributed to a particular modality (Figure 2.5 (d)). We will give more details of this approach (Andrews et al., 2009) in Chapter 3 (Section 3.3.1). Hill and Korhonen (2014) propose a distributed approach and extend Mikolov et al.’s (2013a) skip-gram neural language model (described in Section 2.3), in a fashion analogous to Andrews et al. (2009).

In summary, despite differences in formulation, most existent models conceptualise the problem of perceptually grounding meaning representations as one of learning

from multiple views corresponding to different modalities. These models still represent words as vectors resulting from the combination of representations with different statistical properties that do not necessarily have a natural correspondence (e.g., text and images). These new grounded models of word meaning have been shown to account for human behaviour on a range of cognitive tasks, including lexical substitution (Andrews et al., 2009), word association (Andrews et al., 2009; Feng and Lapata, 2010; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013), semantic priming (Andrews et al., 2009; Johns and Jones, 2012), word similarity (Feng and Lapata, 2010; Silberer and Lapata, 2012, 2014; Bruni et al., 2014; Kiela and Bottou, 2014), compositionality (Roller and Schulte im Walde, 2013), and concept categorisation (Silberer and Lapata, 2014).

Other models that leverage linguistic and visual information from images have been developed with particular tasks in mind. We can distinguish between approaches that exploit visual information for linguistic tasks, such as measuring word relatedness (Leong and Mihalcea, 2011), retrieving word translations (Bergsma and Van Durme, 2011), or predicting selectional preferences (Bergsma and Goebel, 2011), and those that integrate linguistic information for image-oriented tasks, such as image or description retrieval (e.g., Gong et al., 2014; Weston et al., 2010; Socher et al., 2014; Kiros et al., 2014a,b), image annotation or caption generation (e.g., Barnard et al., 2003; Feng and Lapata, 2013; Kiros et al., 2014b; Mao et al., 2014), or object classification and zero-shot learning (e.g., Frome et al., 2013; Socher et al., 2013b; Lazaridou et al., 2014).

The models addressing linguistic tasks do not learn a joint representation. With respect to the image-related models, most researchers adopt a distributed approach and learn a joint representation by projecting both modalities to a bimodal space (Figure 2.5 (c), (d)), or by performing cross-modal learning, projecting the representations in one modality into the space of the other modality via a projection layer (Figure 2.5 (e), e.g., Socher et al., 2013b; Lazaridou et al., 2014). These models intersect with multimodal (deep) learning in networks, which we will discuss in more detail in Chapter 5. An alternative to the latter is to directly train a mapping from one space to the other (Socher et al., 2014). In this case, however, no bimodal representation is learned (Figure 2.5 (f)).

In Chapter 5 (Section 5.3), we present our approach to visually grounded models

of word meaning. It employs multimodal deep learning methods to map visual and textual information into a joint space, corresponding to the mechanism illustrated in Figure 2.5 (d). Our preference for this mechanism over the other methods is based on two assumptions: Firstly, the perceptual and linguistic modalities share correlated information and their unification in a joint space is cognitively more plausible than simply adding the modalities ((a) and (b)), since it is unlikely that humans have separate representations for different aspects of word meaning (Rogers et al., 2004). The joint space is derived by exploiting the interrelationships between the modalities, which is not the case with mechanism (a). Mechanism (e) (and (f)) relies upon a single input modality which contradicts the central assumption discussed above that humans learn word meaning through the exposure to both perceptual and linguistic experience, providing complementary information. In fact, mechanism (e) can be interpreted as learning meaning representations by trying to produce one modality (e.g., language) through the exposure to the respective other modality (e.g., perception), and prior to the experience of the former. Finally, mechanisms (d) and (c) both perform integration by jointly taking into account the two modalities. Whether one type is preferable over the other may therefore depend on the specific algorithms employed.

## 2.6 Conclusions

In this chapter, we reviewed three major strands of computational models of semantic representations, which derive lexical meaning representations either on the basis of text corpora or by means of human-produced attributes of concepts. We referred to the criticism on corpus-based approaches of not being grounded in the physical world as a result of their use of purely distributional statistics. We discussed recent work which deals with this issue and approaches the problem of perceptually grounding meaning representations by integrating perceptual and corpus-based information. In our discussion we focussed on the source of information which existing approaches use to approximate the perceptual modality, and the mechanism they employ for modality integration.



# Chapter 3

## Grounded Models Using Human Input

In the previous chapter we reviewed existent work on perceptually grounded models of lexical semantics and distinguished them according to the used type of perceptual information and the employed integration mechanism. The subject of this chapter is a closer examination of the latter. More precisely, we present a comparative study of three perceptually grounded distributional models and focus on the different mechanisms used for the integration of textual and perceptual data, addressing the following questions:

1. Does the integration of perceptual and textual information yield a better fit with behavioural data compared to a model that considers only one data source?
2. What is the best way to integrate the two information sources?
3. How accurately can we approximate perceptual information for words that do not have any?
4. What type of readily available perceptual information can we exploit? (e.g., visual, auditory, etc.)?

The first model, described in Section 3.3.1 and originally proposed by Andrews et al. (2009), is an extension of latent Dirichlet allocation (LDA, Blei et al., 2003, see Section 2.2.2). The integration mechanism of the model is an instance of type (d) illustrated in Figure 2.5 (page 26). The second model is based on Johns and Jones (2012) who represent the meaning of a word as the concatenation of its textual and its perceptual vector (Section 3.3.2). It is an instance of type (b) (Figure 2.5, page 26). Finally, we propose canonical correlation analysis (CCA, Hotelling, 1936; Haroon

et al., 2004) as our third model in Section 3.3.3. CCA is a data analysis and dimensionality reduction method for joint dimensionality reduction across two (or more) spaces that provide heterogeneous representations of the same objects. The assumption is that the representations in these two spaces contain some joint information that is reflected in correlations between them. This model is an instance of type (c) (Figure 2.5, page 26).

In our experiments (Section 3.4), we first compare the three models using attribute norms as a proxy for perceptual information (Sections 3.4.1 and 3.4.2). Subsequently, we approximate perceptual information with image labels (Section 3.4.3). We start with a description of these two data sources (Sections 3.1 and 3.2), followed by details on the three models (Section 3.3).

### 3.1 Semantic Attribute Production Norms as a Proxy for Perceptual Information

As discussed in Chapter 2 (Section 2.1), attribute norms can stand in as a proxy for sensorimotor experience, and are therefore useful for studying the integration of perceptual and textual information without being susceptible to the effects of noise, e.g., coming from processing of images and other modalities. Norms often cover a small fraction of the vocabulary of an adult speaker due to the effort involved in eliciting them, which makes their large-scale use difficult, but they provide a good starting point, serving as an upper bound of what can be achieved when integrating detailed perceptual information with text-based distributional models.

In this thesis, we rely on the widely used norming study of McRae et al. (2005, henceforth referred to as McRae norms). The norms contain attribute lists for concrete nouns referring to 541 animate and inanimate concepts. The authors unified synonymous attributes during recording and included in the list for a given concept only attributes that had been listed by at least five (out of thirty) participants. In total, the norms contain 2,526 unique attributes out of which 824 co-occur with at least two different nouns. Each attribute is assigned its production frequency, i.e. the number of participants who listed a specific attribute for a concept, and is furthermore categorised into one of nine knowledge types, such as visual or taxonomic information.

Table 3.1 presents examples of attributes participants listed for the nouns *apple*, *dog*, and *table* together with their knowledge types. The table shows probability dis-

Attributes	Concepts			Type
	<i>table</i>	<i>dog</i>	<i>apple</i>	
has_4_legs	.25	.36	0	visual-form/surface
used_for_eating	.44	0	0	function
furniture	.12	0	0	taxonomic
is_red	0	0	.51	visual-colour
is_crunchy	0	0	.21	tactile
is_round	.19	0	.16	visual-form/surface
beh_-barks	0	.42	0	sound
beh_-chases	0	.12	0	visual-motion
tastes_sour	0	0	.12	taste
is_domestic	0	.10	0	encyclopedic

Figure 3.1: Attribute norms for the nouns *table*, *dog*, and *apple* shown as distributions.

tributions over attributes given the words obtained by normalising the attribute production frequencies:

$$P(a_k|w) = \frac{\text{frequ}(a_k, w)}{\sum_{m=1}^A \text{frequ}(a_m, w)}, \quad (3.1)$$

where  $\text{frequ}(a_k, w)$  is the production frequency of attribute  $a_k$  for word  $w$  and  $A$  is the total number of attributes. Recently, another set of attribute norms has been released (Devereux et al., 2013), which is similar in fashion to the McRae norms and covers 639 nominal concepts.

## 3.2 Image Labels as a Proxy for Perceptual Information

The visual modality is a major source of perceptual information and, as the experimental results reported later in this chapter suggest (Section 3.4.2), constitute a strong complement to textual information for modelling word meaning.

Images represent a natural and direct source of visual information. Moreover, they are ubiquitous and therefore provide easily accessible information for a potentially unrestricted number of target concepts (in contrast to, e.g., attribute production norms). In recent years, many collections of images annotated with object labels have emerged

proving invaluable for computer vision research, especially as training and evaluation benchmarks for object detection and recognition. These object labels can be regarded as a human generated proxy for the visual modality, lying between information directly extracted from image data by means of computer vision techniques and attribute production norms.

Most datasets focus on a relatively small number of object classes and provide no (e.g., Caltech-101, Fei-Fei et al., 2007; Caltech-256, Griffin et al., 2007) or little information that goes beyond class labels, such as categories<sup>1</sup> (e.g., Cifar-100, Krizhevsky, 2009) or bounding boxes localising the objects present in an image (e.g., Pascal-VOC, Everingham et al., 2012). Some datasets have been created as part of online crowdsourcing efforts engaging a large number of humans and as a result cover thousands of different object classes with annotations ranging from object labels, image tags, bounding boxes, semantic attributes etc. (ESP, von Ahn and Dabbish, 2004; ImageNet, Deng et al., 2009; MIRFlickr, Huiskes and Lew, 2008; LabelMe, Russell et al., 2008).

In this chapter, we will make use of the ESP and LabelMe datasets, which contain five labels (out of 19K types) and four labels (out of 8K types) on average for each image, respectively, with 68K and 75K images in total (see also Table 3.1).<sup>2</sup> Figure 3.2 shows examples of images and their labels from the two datasets. ESP (von Ahn and Dabbish, 2004) was acquired through an online game, where online users were paired up and presented with the same image picked from a pool of images randomly collected from the web. Both players then typed in descriptive strings for the objects shown in the image, trying to guess what their partner was entering with respect to the objects. Once there was a matching string among the guesses, that is, they agreed on an appropriate label, the game continued with another image. The rationale behind this protocol was that matching guesses are typically good descriptive labels for the image. To increase the number of different labels for an image, labels that had been agreed upon by a certain number of pairs were listed as prohibited words for the next players being presented with the image.

LabelMe (Russell et al., 2008) is a database and a web-based annotation tool for sharing, annotating, and querying images. Image annotation consists of drawing a polygon around each object or parts of it (i.e. providing their outline and location) and

---

<sup>1</sup>Krizhevsky (2009) refer to categories as *superclasses*. For example, the object classes *beaver*, *dolphin* and *otter* are members of the superclass AQUATIC MAMMAL.

<sup>2</sup>The numbers refer to the LabelMe version downloaded in April 2012, and the ESP dataset obtained from the website of J. Langford (<http://www.hunch.net/~jl/>). Von Ahn has made available a larger version of ESP of 100K images at <http://www.cs.cmu.edu/~biglou/resources/> (last accessed in May 2015).

	ESP	LabelMe
number images	67,796	75,243
avg number labels/image	5.1	4.0
number labels	346,794	302,613
number labels (types)	19,354	7,873
number labels associated with $\geq x$ images		
x = 40	911	645
x = 80	529	463
x = 160	290	314
x = 320	161	206

Table 3.1: Statistics of the ESP and LabelMe image datasets.

assigning a name and, optionally, attributes to the object.

There is a notable difference between ESP and LabelMe with respect to the available image labels (cf. Table 3.1). Despite containing a larger number of images, LabelMe has considerably less label types than ESP (8K and 19K types, respectively). Furthermore, LabelMe contains primarily images that do not focus on a certain object. An example are street scenes where houses, streets, or cars are illustrated. Since the annotators in LabelMe were asked to draw a polygon around a certain object and provide a descriptive label for that area, the labels of an image in LabelMe often correspond to objects *co-occurring* in the image. ESP, in contrast, consists exclusively of images and their labels, with the images often focussing on a particular object. Hence, labels in ESP tend to name objects and *parts* and *attributes*. The statistics of the dataset labels reflect this observation: the ten most frequent labels in LabelMe are *person, walking, tree, window, building, sky, car, road, table, door*. In ESP, the most frequent words are *white, black, blue, man, red, green, woman, logo, yellow, tree*.

### 3.3 Models

This section presents the three models which we will experimentally compare in the subsequent section. Recall that the comparative study performed in this chapter addresses the question of what is the best way to integrate linguistic and perceptual information. We thus chose models which represent the different core integration

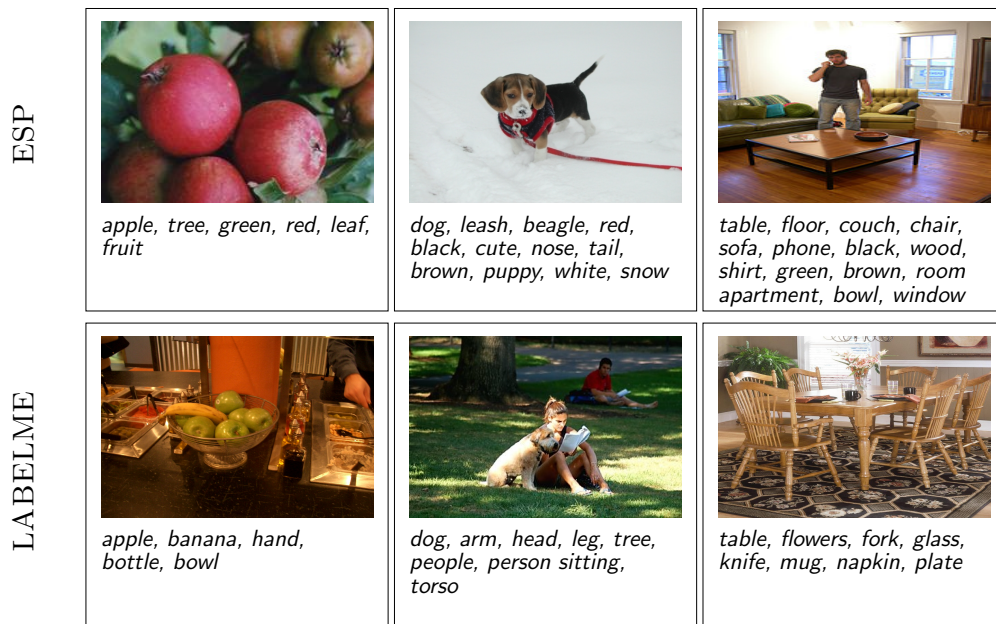


Figure 3.2: Examples of images and labels in ESP and LabelMe.

mechanisms identified in the previous chapter and illustrated in Figure 2.5 (Page 26): concatenation approaches ((a),(b), Section 3.3.2) and the two joint mechanisms ((c), Section 3.3.3, and (d), Section 3.3.1).

### 3.3.1 Attribute-topic Model

Andrews et al. (2009) present an extension of LDA (Blei et al., 2003) where words in documents as well as their associated attributes are treated as observed variables that are explained by a generative process. The underlying training data consists of a corpus  $\mathcal{D}$  where each document is represented by words and their frequency of occurrence within the document. In addition, those words of a document for which also attribute information is available (e.g., they are included in attribute norms) are paired with one of their attributes, where an attribute is sampled according to the attribute distribution given that word (see, e.g., Equation (3.1), Page 31).

For example, suppose a document  $d_j$  consists of the sentence *Mix in the apple, celery, raisins, and apple juice*. Suppose further that for all content words except of *mix* and *juice* attribute information is available. Then, a representation for  $d_j$  is *mix:1, apple;is\_red:2, celery;has\_leaves:1, raisin;is\_edible:1, juice:1*.

The plate diagram in Figure 3.3 illustrates the graphical model in detail. Each document  $d_j$  in  $\mathcal{D}$  is generated by a mixture of components  $\{x_1, \dots, x_c, \dots, x_C\} \in \mathcal{C}$ ;

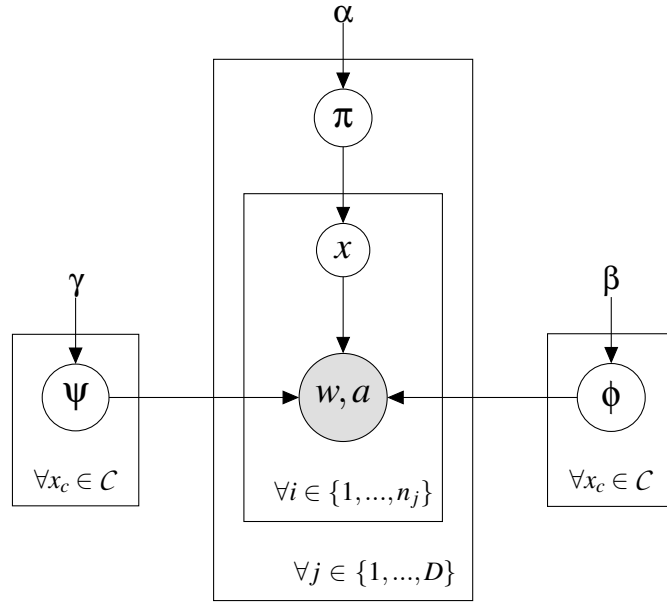


Figure 3.3: Attribute-topic model. The components  $x_{ji}$  of a document  $d_j$  are sampled from  $\pi_j$ . For each  $x_c = x_{ji}$ , a word  $w_{ji}$  is drawn from distribution  $\phi_c$  and an attribute  $a_{ji}$  is drawn from distribution  $\psi_c$ .

a component  $x_c$  comprises a latent discourse topic coupled with an attribute cluster originating from an external source of perceptual information (e.g., attribute norms). A discourse topic belonging to  $x_c$ , in turn, is a distribution  $\phi_c \in \phi = \{\phi_1, \dots, \phi_C\}$  over words, and an attribute cluster is a distribution over attributes,  $\psi_c \in \psi = \{\psi_1, \dots, \psi_C\}$ .

In order to create document  $d_j$ , a distribution  $\pi_j$  over components is sampled from a Dirichlet distribution parametrised by  $\alpha$ . To generate each word  $w_{ji} \in \{w_{j1}, \dots, w_{jn_j}\}$ , a component  $x_c = x_{ji}$  is drawn from  $\pi_j$ ;  $w_{ji}$  is then drawn from the corresponding distribution  $\phi_c$ . If there is attribute information available for  $w_{ji}$ , it is coupled with an attribute  $a_{ji}$  which is correspondingly drawn from  $\psi_c$ . A symmetric Dirichlet prior with hyperparameters  $\beta$  and  $\gamma$  is placed on  $\phi$  and  $\psi$ , respectively. The probability of the corpus  $\mathcal{D}$  is defined as:

$$P((w \cup a)_{1:D} | \phi, \psi, \alpha) = \prod_{j=1}^D \int d\pi_j \prod_{i=1}^{n_j} P(\pi_j | \alpha) \sum_{c=1}^C P(w_{ji} | x_{ji} = x_c, \phi) P(a_{ji} | x_{ji} = x_c, \psi) P(x_{ji} = x_c | \pi_j) \quad (3.2)$$

where  $D$  is the number of documents and  $C$  the predefined number of components. Computing the posterior distribution  $P(\phi, \psi, \alpha, \beta, \gamma | (w \cup a)_{1:D})$  of the hidden variables

$$apple \begin{bmatrix} x_1 & x_2 & x_{12} & \dots & x_{28} & x_{75} & x_{107} & x_{119} & x_{125} & x_{148} & x_{182} & \dots & x_{266} & x_{326} & x_{349} & x_{350} \\ 3e-5 & 3e-5 & 0 & \dots & 5e-4 & 9e-4 & .09 & .002 & 7.6e-5 & 2e-4 & .003 & \dots & 0 & 0 & 3e-6 & 0 \end{bmatrix}$$

Figure 3.4: Example of the representation of the meaning of *apple* with the model of Andrews et al. (2009) .

given the data is generally intractable:

$$P(\phi, \psi, \alpha, \beta, \gamma | (w \cup a)_{1:D}) \propto P((w \cup a)_{1:D} | \phi, \psi, \alpha) P(\phi | \beta) P(\psi | \gamma) P(\alpha) P(\beta) P(\gamma) \quad (3.3)$$

Equation (3.3) may be approximated using the Gibbs sampling procedure described in Andrews et al. (2009).

Inducing attribute-topic components from a document collection  $\mathcal{D}$  with the extended LDA model just described gives two sets of parameters: word probabilities given components  $P_W(w_i | X = x_c)$  for  $w_i, i = 1, \dots, N$ , and attribute probabilities given components  $P_A(a_k | X = x_c)$  for  $a_k, k = 1, \dots, A$ . For example, most of the probability mass of a component  $x$  would be reserved for the words *apple*, *fruit*, *lemon*, *orange*, *tree* and the attributes *is\_red*, *tastes\_sweet*, *is\_round* and so on.

Word meaning in this model is represented by the distribution  $P_{X|W}$  over the learned components (see Figure 3.4 for an example). Assuming a uniform distribution over components  $x_c$  in  $\mathcal{D}$ ,  $P_{X|W}$  can be approximated as:

$$P_{X=x_c|W=w_i} = \frac{P(w_i|x_c)P(x_c)}{P(w_i)} \approx \frac{P(w_i|x_c)}{\sum_{l=1}^C P(w_i|x_l)} \quad (3.4)$$

where  $C$  is the total number of components. The model can be also used to infer attributes for words for which no attribute information is available. The probability distribution  $P_{A|W}$  over attributes given a word  $w_i$  is simply inferred by summing over all components  $x_c$  for each attribute  $a_k$ :

$$P_A(a_k | W = w_i) = \sum_{c=1}^C P(a_k | x_c) P(x_c | w_i) \quad (3.5)$$

### 3.3.2 Global Similarity Model

Johns and Jones (2012) propose an approach for generating perceptual representations for words by means of global lexical similarity. Their model does not place so much



$$\begin{array}{l}
\text{apple} \begin{bmatrix} \dots & d_{16} & \dots & d_{322} & \dots & d_{2469} & d_{2470} & \dots & d_D \\ \dots & 1 & \dots & 1 & \dots & 0 & 1 & \dots & 0 \end{bmatrix} \begin{bmatrix} a\_fruit & has\_fangs & is\_crunchy & \dots & is\_yellow & is\_red & is\_green & is\_round \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \end{bmatrix} \\
\text{apple} \begin{bmatrix} \dots & d_{16} & \dots & d_{322} & \dots & d_{2469} & d_{2470} & \dots & d_D \\ \dots & 1 & \dots & 1 & \dots & 0 & 1 & \dots & 0 \end{bmatrix} \begin{bmatrix} a\_fruit & has\_fangs & is\_crunchy & \dots & is\_yellow & is\_red & is\_green & is\_round \\ .006 & 1.8e-5 & 8e-4 & \dots & .004 & .004 & .006 & .02 \end{bmatrix}
\end{array}$$

Figure 3.5: Example of the representation for *apple* before (first row) and after (second row) applying the perceptual inference method of John & Jones (2012).

emphasis on the integration of perceptual and linguistic information, rather its main focus is on inducing perceptual representations for words with no perceptual correlates. Their idea is to assume that lexically similar words also share perceptual attributes and hence it should be possible to transfer perceptual information onto words that have none from their linguistically similar neighbours.

Let  $\mathbf{T} \in \{1, 0\}^{N \times D}$  denote a binary word-document matrix, where each cell records the presence or absence of a word in a document. Let  $\mathbf{P} \in [0, 1]^{N \times A}$  denote a perceptual matrix, representing a probability distribution over attributes for each word (see, e.g., Equation (3.1), Page 3.1). A word's meaning is represented by the concatenation of its textual and perceptual vectors (see Figure 3.5, second row). If a word is lacking attribute information, its perceptual vector will be all zeros. Johns and Jones (2012) propose a two-step estimation process for words without perceptual vectors. Initially, a perceptual vector is constructed based on the word's weighted similarity to other words that have non-zero perceptual vectors:

$$\mathbf{p}_{inf} = \sum_{i=1}^N \mathbf{t}_i * \text{sim}(\mathbf{t}_i, \mathbf{p})^\lambda \quad (3.6)$$

where  $\mathbf{p}$  is the representation of a word with a textual vector but an empty perceptual vector,  $\mathbf{t}_i$  are composite representations consisting of textual and perceptual vectors,  $\text{sim}$  is a measure of distributional similarity such as the cosine similarity,  $\lambda$  a weighting parameter, and  $\mathbf{p}_{inf}$  the resulting inferred representation of the word. The process is repeated a second time, so as to incorporate the inferred perceptual vector in the computation of the inferred vectors of all other words. An example of this inference procedure is illustrated in Figure 3.5.

### 3.3.3 Canonical Correlation Analysis

Our third model uses canonical correlation analysis (CCA, Hotelling, 1936; Haroon et al., 2004) to learn a joint semantic representation from the textual and perceptual views. Given two random variables  $\mathbf{x}$  and  $\mathbf{y}$  (or two sets of vectors), CCA can be seen as determining two sets of basis vectors in such a way, that the correlation between the projections of the variables onto these bases is mutually maximised (Borga, 2001). In effect, the representation-specific details pertaining to the two views of the same phenomenon are discarded and the underlying hidden factors responsible for the correlation are revealed.

In our case the linguistic view is represented by a word-document matrix,  $\mathbf{T} \in \mathbb{R}^{N \times D}$ , containing information about the occurrence of each word in each document. The perceptual view is captured by a perceptual matrix,  $\mathbf{P} \in [0, 1]^{N \times A}$ , representing words as a probability distribution over attributes (see, e.g., Equation 3.1).

CCA is concerned with describing linear dependencies between two sets of variables of relatively low dimensionality. Since the correlation between the linguistic and perceptual views may exist in some nonlinear relationship, we use a kernelised version of CCA, kernel canonical correlation analysis (kCCA, Haroon et al., 2004), which first projects the data into a higher-dimensional feature space and then performs CCA in this new feature space. The two kernel matrices are  $\mathbf{K}_T = \mathbf{T}\mathbf{T}'$  and  $\mathbf{K}_P = \mathbf{P}\mathbf{P}'$ .

After applying kCCA we obtain two matrices projected onto  $L$  basis vectors,  $\mathbf{C}_t \in \mathbb{R}^{N \times L}$ , resulting from the projection of the textual matrix  $\mathbf{T}$  onto the new basis and  $\mathbf{C}_p \in \mathbb{R}^{N \times L}$ , resulting from the projection of the corresponding perceptual attribute matrix.

The meaning of a word can thus be represented by its projected textual vector in  $\mathbf{C}_T$ , its projected perceptual vector in  $\mathbf{C}_P$  or their concatenation. Figure 3.6 shows an example of the textual and perceptual vectors for the word *apple* which were used as input for kCCA (first row) and their new representation after the projection onto new basis vectors (second row).

The kCCA model as sketched above will only obtain full representations for words with perceptual attributes available. One solution would be to apply the method from Johns and Jones (2012) to infer the perceptual vectors and then perform kCCA on the inferred vectors. Another approach which we assess experimentally (see Section 3.4.1) is to create a perceptual vector for a word that has none from its  $k$ -most (textually) similar neighbours, simply by taking the average of their perceptual vectors. This

$$\begin{array}{c}
 \text{apple} \quad \left[ \begin{array}{cccccc} \dots & d_{16} & \dots & d_{322} & \dots & d_{2470} & \dots & d_D \\ \dots & .006 & \dots & .003 & \dots & .1e-6 & \dots & 0 \end{array} \right] \left[ \begin{array}{cccccccc} a\_fruit & has\_fangs & is\_crunchy & \dots & is\_yellow & is\_red & is\_green & is\_round \\ .13 & 0 & .06 & \dots & .04 & .14 & .09 & .04 \end{array} \right] \\
 \text{apple} \quad \left[ \begin{array}{cccccc} k_1 & k_2 & k_3 & \dots & k_{409} & k_{410} \\ -.003 & -.01 & .002 & \dots & -.002 & -.01 \end{array} \right] \left[ \begin{array}{cccccc} k_1 & k_2 & k_3 & \dots & k_{409} & k_{410} \\ .008 & -.03 & -.008 & \dots & -.02 & -.07 \end{array} \right]
 \end{array}$$

Figure 3.6: Example representation for *apple* before (first row) and after (second row) applying CCA.

inference procedure can be applied to the original vectors or the projected vectors in  $\mathbf{C}_T$  and  $\mathbf{C}_P$ , respectively, once kCCA has taken place.

### 3.3.4 Discussion

Johns and Jones (2012) primarily present a model of perceptual inference, where textual data is used to infer perceptual information for words for which no attribute information is available. There is no means in this model to obtain a joint representation resulting from the mutual influence of the perceptual and textual views. As shown in the example in Figure 3.5, the textual vector on the left-hand side does not undergo any transformation whatsoever.

The generative model put forward by Andrews et al. (2009) learns meaning representations by simultaneously considering documents and attributes. Rather than simply adding perceptual information to textual data, it integrates both modalities jointly in a *single* representation which is desirable, at least from a cognitive perspective (see Section 2.5.2). Similarly to Johns and Jones (2012), Andrews et al.’s attribute-topic model can also infer perceptual representations for words that have none. The inference is performed automatically in an implicit manner during component induction.

In kCCA, textual and perceptual data represent two different views of the same objects and the model operates on these views *directly* without combining or manipulating any of them a priori. Instead, the combination of the two modalities is realised via correlating the linear relationships between them. A drawback of the model lies in the need of additional methods for inferring perceptual representations for words that have none.

## 3.4 Experiments

### 3.4.1 Experiment 1: Perceptual Information from Attribute Norms

The first experiment addresses our hypothesis that representations obtained by integrating textual and perceptual information yield a better fit with behavioural data than unimodal models. Goal of the experiment is furthermore to shed light on the question of which is the best integration mechanism. We therefore experimentally compare the bimodal models presented above on two tasks related to word similarity and association, respectively, and also contrast them to their unimodal variants. We furthermore evaluate the models also in terms of their ability to infer absent perceptual information.

**Data** All our simulations used a lemmatised version of the British National Corpus (BNC) as a source of textual information. The attribute norms of McRae et al. (2005, McRae norms) were used as a proxy for perceptual information. We encoded perceptual word vectors as a probability distribution over attributes computed according to Equation (3.1) (Page 31). The BNC comprises 4,049 texts totalling approximately 100 million words. The McRae norms consist of 541 words and 2,526 attributes; we used the 824 attributes which occur with at least two different words.

**Evaluation Tasks** Our evaluation experiments compared the models discussed above on three tasks. Two of them have been previously used to evaluate semantic representation models, namely word association and word similarity. In order to simulate word association, we used the human norms collected by Nelson et al. (1998).<sup>3</sup> These were established by presenting a large number of participants with a cue word (e.g., *rice*) and asking them to name an associate word in response (e.g., *Chinese, wedding, food, white*). For each cue word, the norms provide a set of associates and the frequencies with which they were named. We can thus compute the probability distribution over associates for each cue. Analogously, we can estimate the degree of similarity between a cue and its associates using our models (see the following section for details on the similarity measures we employed). Word association norms can be considered a reflection of the links between words as manifested in semantic memory, with the links most likely to be driven by semantic relations (McRae and Jones, 2013). The norms contain 63,619 unique normed cue-associate pairs in total. Of these, 25,968 pairs were

---

<sup>3</sup>Available at <http://w3.usf.edu/FreeAssociation/> (last accessed in April 2015).

covered by all models and 520 appeared in the McRae norms. Using correlation analysis, we examined the degree of linear relationship between the human cue-associate probabilities and the automatically derived similarity values. We follow previous work (Griffiths et al., 2007b; Nelson et al., 1998) in reporting correlation coefficients using Pearson’s  $r$ .<sup>4</sup> A rank correlation measure may be less informative on this dataset, since many cue-associate pairs share the same probability.

Our word similarity experiments used the WordSimilarity-353 test collection (Finkelstein et al., 2002)<sup>5</sup> which consists of similarity judgements for word pairs. For each pair, a judgement (on a scale of 0 to 10) was elicited from 13 or 16 human subjects (e.g., *tiger-cat* are very similar, whereas *delay-racism* are not). The average rating for each pair represents an estimate of the perceived semantic similarity of the two words. The task varies slightly from word association. Here, participants are asked to rate perceived similarity rather than to generate the first word that came to mind in response to a cue word. The collection contains ratings for 353 word pairs. Of these, 76 pairs appeared in our corpus and 3 in the McRae norms. We created grounded representations by inferring missing perceptual information with the models presented in Section 3.3 and, again, evaluated how well model produced similarities correlate with human ratings. We report Pearson’s  $r$  for comparison reasons and Spearman’s rank correlation coefficient ( $\rho$ ). The latter is the commonly used measure for this dataset (Agirre et al., 2009).

Our third task directly assessed the ability of the models to infer perceptual vectors for words that have none. To do this, we conducted 10-fold cross-validation on the McRae norms. We treated the perceptual vectors in each test fold as unseen, and used the data in the corresponding training fold together with the models presented in Section 3.3 to infer them. Then, for each word, we examined how close the inferred vector was to the actual one, via correlation analysis.

**Model Parameters** The attribute-topic model has a few parameters that must be instantiated. These include,  $C$ , the number of predefined components and the priors  $\alpha$ ,  $\beta$ , and  $\gamma$ . We adopted all parameter settings determined by Andrews et al. (2009). Specifically, we set the components  $C$  to 350 and placed a vague inverse gamma prior on  $\alpha$ ,

---

<sup>4</sup>Griffiths et al. (2007b) furthermore report how many times the word with the highest score under the model was the first associate in the human norms. This evaluation metric assumes that there are many associates for a given cue which unfortunately is not the case in our study which is restricted to the concepts represented in the McRae norms.

<sup>5</sup>Available at <http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/> (last accessed in May 2015).

$\beta$ , and  $\gamma$ .<sup>6</sup> To allow for comparison to a unimodal variant, we trained a vanilla LDA model on the BNC only and, following Andrews et al. (2009), set parameter  $C$  to 250.

For estimating word similarity within the attribute-topic model, we adopt Griffiths et al.'s (2007b) definition. The underlying idea is that word association can be expressed as a conditional distribution. If we have seen word  $w_1$ , then we can determine the probability that  $w_2$  will be also generated by computing  $P(w_2|w_1)$ . Assuming that both  $w_1$  and  $w_2$  came from a single component,  $P(w_2|w_1)$  can be estimated as:

$$P(w_2|w_1) = \sum_{c=1}^C P(w_2|x_c)P(x_c|w_1) \quad (3.7a)$$

$$P(x_c|w_1) \propto P(w_1|x_c)P(x_c), \quad (3.7b)$$

where  $P(x_c)$  is uniform, a single component  $x_c$  is sampled from the distribution  $P(x_c|w_1)$ , and an overall estimate is obtained by averaging over all  $C$  components.

Johns and Jones' (2012) model uses binary textual vectors to represent word meaning. If the word is present in a given document, that vector element is coded as one; if it is absent, it is coded as zero. We built a binary word-document matrix from the BNC over 14,000 lemmas. The value of the similarity weighting parameter  $\lambda$  was set to the same values reported by Johns and Jones ( $\lambda_1 = 3$  for Step 1 and  $\lambda_2 = 13$  for Step 2).

For the kCCA model, we represented the textual view with a word-document matrix. Matrix cells were set to their tf-idf values (cf. Equation (2.1), Page 16).<sup>7</sup> The textual and perceptual matrices were projected onto 410 vectors. As mentioned in Section 3.3.3, kCCA does not naturally lend itself to inferring perceptual vectors, yet a perceptual vector for a word can be created from its  $k$ -nearest neighbours. We inferred a perceptual vector by averaging over the perceptual vectors of the word's  $k$  most similar words; textual similarity between two words was measured using the cosine similarity of the two vectors representing them (cf. Equation (2.4)). To find the optimal value for  $k$ , we used one third of the cues of Nelson's (1998) norms as development set. The highest correlation was achieved with  $k = 2$  when the perceptual vectors were created prior to kCCA and  $k = 8$  when they were inferred on the projected textual and perceptual matrices.

**Results** In order to assess whether integrated perceptual and textual information account more for human behaviour than only one type of data source we measure the

<sup>6</sup>That is  $P(\bullet) = \exp(-\frac{1}{\bullet})\bullet^{-2}$ .

<sup>7</sup>Experiments with a binarised version of the word-document matrix consistently performed worse.

Models	Pearson's $r$		
	T	P	T+P
Attribute-topic	.12	.22	.35
Global similarity	.11	.22	.23
kCCA	.14	.29	.32
Upper Bound	.91		

Table 3.2: Performance of attribute-topic, global similarity, and kCCA models on a subset of the Nelson norms when taking into account the textual and perceptual modalities on their own (columns T and P, respectively) and in combination (T+P). All correlation coefficients are statistically significant ( $p < 0.01$ ).

performance of the models on the word association task when textual and perceptual information are both available. The results in Table 3.2 are thus computed on the subset of Nelson's norms (520 cue-associate pairs) that also appeared in the McRae norms and for which a perceptual vector was present. The table shows different instantiations of the three models depending on the type of modality taken into account: textual (column T), perceptual (P) or both (T+P).

As can be seen, Andrews et al.'s (2009) attribute-topic model provides a better fit with the association data when both modalities are taken into account (column T+P). A vanilla LDA model constructed solely on the BNC (column T) or the McRae norms (column P) yield substantially lower correlations. We observe a similar pattern with Johns and Jones' (2012) global similarity model. Concatenation of perceptual and textual vectors yields the best fit with the norming data (T+P), relying on perceptual information alone comes close (P), whereas textual information on its own seems to have a weaker effect (T). Note, that in this evaluation setting, the global similarity models does not infer any perceptual representations, since perceptual vectors are provided directly by the McRae norms for all words. The kCCA model takes perceptual and textual information as input in order to find a projection onto basis vectors that are maximally correlated. Although by definition the kCCA model must operate on the two views, we can nevertheless isolate the contribution of each modality by considering the vectors resulting from the projection of the textual matrix (T), the perceptual matrix (P) or their concatenation (T+P). We obtain best results with the latter representation; again we observe that the perceptual information is more dominant.

Models	Pearson's $r$
Attribute-topic	.15
Global similarity	.03
Global similarity $\ll$ kCCA	.12
$k$ -NN $\ll$ kCCA	.11
kCCA $\ll$ $k$ -NN	.12
Upper Bound	.96

Table 3.3: Performance of the attribute-topic, global similarity and kCCA models on the Nelson norms (entire dataset). All correlation coefficients are statistically significant ( $p < 0.01$ ).

Overall we find that the attribute-topic model and kCCA perform best. In fact the correlations achieved by the two models do not differ significantly, using a  $t$ -test (Cohen and Cohen, 1983). The performance of the global similarity model is significantly worse than the attribute-topic model and kCCA ( $p < 0.01$ ). Recall that the attribute-topic model (T+P) represents words as distributions over components, whereas the global similarity model simply concatenates the textual and perceptual vectors. The same input is also given to kCCA which in turn attempts to interpret the data by inferring common relationships between the two views. In sum, we can conclude that the higher correlation with human judgements indicates that integrating textual and perceptual modalities jointly is preferable to concatenation.

However, note that all models in Table 3.2 fall short of the human upper bound which we measured by calculating the *reliability* of Nelson et al.'s (1998) norms. Reliability estimates the likelihood of a similarly-composed group of participants presented with the same task under the same circumstances producing identical results. We split the collected cue-associate pairs randomly into two halves and computed the correlation between them; this correlation was averaged across 200 random splits. These correlations were adjusted by applying the Spearman-Brown prediction formula (Voorspoels et al., 2008).

The results in Table 3.2 are computed on a small fraction of Nelson et al.'s norms. One might even argue that the comparison is slightly unfair as the global similarity model is more geared towards inferring perceptual vectors rather than integrating the two modalities in the best possible way. To gain a better understanding of the models'



behaviour and to allow comparisons on a larger dataset and more equal footing, we also report results on the entire dataset (20,556 cue-associate pairs).<sup>8</sup> This entails that the models will infer perceptual vectors for the words that are not attested in the McRae norms. Recall from Section 3.3.3 that kCCA does not have a dedicated inference mechanism. We thus experimented with three options (a) interfacing the inference method of Johns and Jones (2012) with kCCA (global similarity  $\ll$  kCCA) (b) creating a perceptual vector from the words'  $k$ -nearest neighbours before ( $k$ -NN  $\ll$  kCCA) or (c) after kCCA takes place (kCCA  $\ll$   $k$ -NN).

Our results are summarised in Table 3.3. The upper bound was estimated in the same way as for the smaller dataset. Despite being statistically significant ( $p < 0.01$ ), the correlation coefficients are lower. This is hardly surprising as perceptual information is approximate and in several cases likely to be wrong. Interestingly, we observe similar modelling trends, irrespective of whether the models are performing perceptual inference or not. The attribute-topic model achieves the best fit with the data, followed by kCCA. The inference method here does not seem to have much of an impact: kCCA  $\ll$   $k$ -NN does as well as global similarity  $\ll$  kCCA. This is perhaps expected as the inference procedure adopted by Johns and Jones (2012) is a generalisation of our  $k$ -nearest neighbour approach. The global similarity model performs worst; we conjecture that this is due to the way semantic information is integrated rather than the inference method itself. KCCA works with similar input, yet achieves better correlations with the human data, due to its ability to represent the commonalities shared by the two modalities.

Taken together, the results in Tables 3.2 and 3.3 answer the question of what is the best way to integrate the modalities: models that capture latent information shared between the two modalities create more accurate semantic representations compared to simply treating the two as independent data sources.

In order to isolate the influence of the inference method from the resulting semantic representation we evaluated the inferred perceptual vectors on their own by computing their correlation with the original attribute distributions in the McRae norms. The correlation coefficients are reported in Table 3.4 and were computed by averaging the coefficients obtained for individual words. Here, the global similarity model achieves the highest correlation, and for a good reason. It is the only model with an emphasis on inference, the other two models do not have such a dedicated mechanism. KCCA has

---

<sup>8</sup>This excludes the data used as development set for tuning the  $k$ -nearest neighbours for kCCA.

Models	Pearson's $r$
Attribute-topic	.17
Global similarity	.25
Global similarity $\ll$ kCCA	.21
$k$ -NN $\ll$ kCCA	.19
kCCA $\ll$ $k$ -NN	.13

Table 3.4: Mean correlation coefficients between original and inferred attribute vectors in McRae et al.'s norms.

Models	Pearson's $r$	Spearman's $\rho$
Attribute-topic	.35	.43
Global similarity	.08	.09
Global similarity $\ll$ kCCA	.38	.38
$k$ -NN $\ll$ kCCA	.39	.39
kCCA $\ll$ $k$ -NN	.28	.26
Upper Bound	.98	

Table 3.5: Model performance on predicting word similarity. All correlation coefficients are statistically significant ( $p < 0.01$ ), except for the global similarity model.

in fact none, whereas in the attribute-topic model the inference of missing perceptual information is a by-product of the generative process. The results in Table 3.4 give an indicative answer to our question on how well we can approximate perceptual information for words that do not have any: the perceptual vectors are not reconstructed very accurately (the highest correlation coefficient is  $r = .25$ ), hence, better inference mechanisms are required for perceptual information to have a positive impact on semantic representation.

In Table 3.5 we examine the models' effectiveness on semantic similarity rather than association using the WordSimilarity-353 dataset (Finkelstein et al., 2002). The models were evaluated on 76 word pairs that appeared in the BNC. We inferred the perceptual vectors for 51 words. We computed the upper bound using the reliability method described earlier. Again, the joint models achieve better results than the simple concatenation model. The attribute-topic and kCCA models perform comparably, with

the global similarity model lagging substantially behind.

### 3.4.2 Experiment 2: Feature Engineering Attribute Norms

In the previous experiments we have used the McRae norms without any extensive feature engineering other than applying a frequency cut-off. However, these norms do not exclusively encode perceptual but also linguistic knowledge, such as taxonomic information (e.g., *a\_fruit*), or specify the function of a concept (e.g. *eaten\_in\_pies*; see Section 3.1).

Subject of this experiment is to unravel the contribution of attributes capturing purely visual information. For that purpose, we distinguish between models obtained by only integrating visual attributes contained in the McRae norms and models ignoring those by integrating all but the visual attributes of the norms (*non-visual models*). This provides a first insight into whether modelling word meaning based on textual and *visual* information represents a valuable approximation to grounded semantic models.

**Data** Analogously to the previous experiment, we used the BNC as a source of textual information, and the McRae norms as a proxy for perceptual information. As described in Section 3.1, the McRae norms contain a classification of each attribute into one of nine knowledge types. We use the attributes classified as visual (i.e. *visual-motion*, *visual-form/surface*, *visual-colour*) for the models integrating visual information and all other attributes, including other sensory attributes (e.g., *sound*, *smell*) and non-perceptual attributes (e.g., *taxonomic*), for the models integrating non-visual information. There are 676 visual attributes in the norms; we used the 295 attributes which are listed with at least two distinct concepts.

**Evaluation Task** We compare the models on the word association task employed in the previous experiment (Section 3.4.1) using Nelson et al.'s (1998) norms. Recall that our goal in this experiment is to specifically assess the contribution of visual information. For this reason, we do not perform inference of perceptual or, more precisely, visual vectors for words that have none. Consequently, we do not assess the performance of the models on a word similarity task as in the previous experiment, since this required the inference of vectors for 51 words.

**Model Parameters** We adopted all parameter settings from the previous experiment except for the global similarity model, which we now augmented with the same matri-

Models	Attributes	Pearson's $r$		
		T	P	T+P
Attribute-topic	visual	.12	.23	.25
Attribute-topic	non-visual	.12	.17	.30
Attribute-topic	all	.12	.22	<b>.35</b>
Global similarity	visual	.14	.23	.24
Global similarity	non-visual	.14	.17	.19
Global similarity	all	.14	.22	.24
kCCA	visual	.19	.31	<b>.37</b>
kCCA	non-visual	.17	.23	.28
kCCA	all	.14	.29	.32
Upper Bound	—	.91		

Table 3.6: Performance of the models on a subset of the Nelson norms when taking into account the textual and perceptual modalities on their own (columns T and P, respectively) and in combination (T+P). All correlation coefficients are statistically significant ( $p < 0.01$ ). The results differ according to which attribute class of the McRae norms was used.

ces as given to the kCCA model. More precisely, we now used the better performing tf-idf-weighted word-document matrix for the textual view instead of a binary matrix. Note that the global similarity model was obtained without performing inference (see Section 3.3.2).

**Results** Table 3.6 reports the results on the 493 pairs (101 cues) for which word representations were computed. Column T+P gives the effectiveness of the models in the bimodal setting, i.e. when integrating textual and either visual, non-visual or all McRae attributes. Overall, the two joint approaches, kCCA and the attribute-topic model, yield a better fit to human data irrespective of the attribute set (visual, non-visual, all). kCCA using visual attributes (row visual) achieved the highest correlation coefficient across all settings. Moreover, models augmented with visual input (columns T+P and P) perform comparably or even better in most cases than their variants using non-visual or all attributes (rows non-visual and all). Only the attribute-topic model seems to suffer from the lower number of attributes in the visual setting (column T+P), performing

comparably to the much simpler global similarity model, and being more effective when using all attributes.

In conclusion, the results in Table 3.6 suggest that visual attributes alone are the most salient and valuable source of perceptual information provided by attribute norms. The results moreover confirm the conclusion made in the previous experiment: joint models integrate the two modalities more effectively than a concatenation approach.

### 3.4.3 Experiment 3: Visual Information from Image Labels

As demonstrated in the previous experiments, attribute norms are a useful first approximation of perceptual and especially visual data. However, the effort involved in eliciting them limits the scope of any computational model based on normed data. We now shift to image datasets as a source of visual information and exploit the natural language descriptions the datasets provide for each image.

**Data** Analogously to the previous experiments, we used a lemmatised version of the BNC as a source of textual information. For the visual information, we used the image datasets ESP (von Ahn and Dabbish, 2004) and LabelMe (Russell et al., 2008) described in Section 3.2.

**Evaluation Tasks** Our evaluation experiments compared the models discussed in Section 3.3 on the word association task using the human norms collected by Nelson et al. (1998). The models covered 2,482 pairs of the Nelson norms.

**Model Parameters** The models outlined in Section 3.3 were built on either the ESP dataset, or the LabelMe dataset, or on a combination of both by treating them as one corpus (ESP+LabelMe). For integrating the image labels, we computed a word-word matrix representing the weighted co-occurrence frequency of the labels occurring in the respective dataset. We applied ratio weighting (Equation (2.2), Page 2.2). The dimensions (co-occurring labels) were determined experimentally on the held out Nelson development set from Experiment 1 (see Section 3.4.1). For the attribute-topic model, all labels with a frequency of at least 5 were considered. For kCCA, an optimal frequency threshold was found at 320 for the ESP dataset, 40 for LabelMe, and 100 for ESP+LabelMe. For the kCCA model we set an additional parameter, namely the number of basis vectors considered for the projection of the input data onto the new dimensions. We experimentally determined this number of dimensions on the Nelson

Models	Dataset	Pearson's $r$		
		T	V	T+V
Attribute-topic	ESP	.124	.133	.342
Attribute-topic	LabelMe	.124	.111	.330
Attribute-topic	ESP+LabelMe	.124	.112	.331
Global similarity	ESP	.172	.178	.179
Global similarity	LabelMe	.172	.130	.131
Global similarity	ESP+LabelMe	.172	.239	.248
kCCA	ESP	.230	.140	.191
kCCA	LabelMe	.220	.096	.181
kCCA	ESP+LabelMe	.196	.241	.254

Table 3.7: Performance of the attribute-topic, global similarity, and kCCA models on a subset of the Nelson norms (2,482 pairs) when taking into account the textual and perceptual modalities on their own (columns T and V, respectively) and in combination (T+V).

development set. For ESP, we employed the 150 basis vectors with highest correlation coefficients resulting from the application of kCCA, and for ESP+LabelMe the dimensions were set to 100. For LabelMe, all basis vectors were chosen.

As in Experiment 2 (Section 3.4.2), the global similarity model was obtained without performing inference, and the vectors of the two modalities are the same as we used as input for the kCCA model.

**Results** Table 3.7 shows the results on the Nelson pairs which were covered by the models obtained on the different databases. All models yield a lower correlation coefficient when using LabelMe compared to their variants based on ESP. The reason for this might be the different types of information provided by ESP (object labels and their parts and attributes) and LabelMe (labels of co-occurring objects) as pointed out in Section 3.2. The attribute-topic model based on both modalities (column T+V) yielded overall the best fit to human data independent of the database. The global similarity model and kCCA perform comparably when using the combination of ESP and LabelMe (ESP+LabelMe). Global similarity does not benefit from the information provided by either ESP or LabelMe; its textual representations (T) achieve compara-

ble or higher correlation coefficients than the visual modality (V) or their concatenation (T+V). Interestingly, the kCCA projections of the textual input onto the bimodal space (column T), despite using input identical to global similarity, outperform global similarity, which indicates that kCCA does benefit from visual information.

In summary, we confirm our conclusions from the previous experiments with limitations: only the joint models, derived on the basis of textual and visual information (image labels), yield a better fit with behavioural data than just a single modality. Also, the true benefit of using image labels as an approximation of the visual modality is less clear than it was the case with visual attributes from the McRae norms (see Section 3.4.2) and depends on the choice of the dataset. Finally, in order to leverage image labels, we had to determine an optimal threshold of label frequencies for each dataset. We find that the main advantage of image labels over norms is that the former have a higher word coverage (in the form of image labels). Since they can therefore provide an approximation of visual information for a larger number of words, additional inference methods such as those performed in Experiment 1 (Section 3.4.1) are not necessary to the same extent for our purpose of modelling word meaning representations. Moreover, we demonstrated that image datasets can be extended, e.g. by merging different datasets, in order to obtain better coverage. An extension of attribute norms is less straightforward.

### 3.5 Conclusions

In this chapter, we compared three different models of semantic representation which compute word meaning on the basis of textual and perceptual information. The models differ in terms of the mechanisms by which they integrate the two modalities. In the attribute-topic model (Andrews et al., 2009), the textual and perceptual views are integrated via a set of latent components that are inferred from the *joint* distribution of textual words and perceptual words (attributes or image labels). The model based on canonical correlation analysis (Hardoon et al., 2004) integrates the two views by deriving *consensus* representations based on the correlation between the linguistic and perceptual modalities. Johns and Jones' (2012) similarity-based model simply concatenates the two representations. In addition, it uses the linguistic representations of words to infer perceptual information when the latter is absent.

Experiments on word association and similarity show that all models benefit from the integration of perceptual data. We find that joint models (i.e. of types (c) and (d)

in Figure 2.5, Page 26) are superior as they obtain a closer fit with human judgements compared to the simple concatenation of the two views (Figure 2.5 (a),(b)) .

We also examined how these models, when augmented with attribute norms as source of perceptual information, perform on the perceptual inference task which has implications for the wider applicability of grounded semantic representation models. Johns and Jones' (2012) inference mechanism goes some way towards reconstructing the information contained in the attribute norms, however, further work is needed to achieve representations accurate enough to be useful in semantic tasks.

On the basis of these results, we formulate the following desiderata for models of perceptually grounded meaning representations (in descending order of importance):

- (1) The models should integrate different modalities by means of a joint mechanism (Figure 2.5 (c),(d)).
- (2) The models should offer the flexibility to map new words into the shared space, i.e. it should be able to derive meaning representations for out-of-vocabulary words which were not part of the underlying training data.
- (3) It would be desirable if the models had the flexibility to map just one modality into the shared space, and were possibly capable of inferring information about the missing modality.

In Chapter 5 we present our model which has been designed with these requirements in mind.

We furthermore examined in isolation the contribution of the visual modality as one type of perceptual information, which we approximated either with the visual attributes of McRae et al.'s (2005) norms or with human-generated image labels. Our experimental results on word association suggest that visual information plays a strong role in grounding meaning representations. Unfortunately, the applicability of both types of data sources, attribute norms and image labels, is limited in scope since their creation requires human efforts. This shortcoming is even more so the case with attribute norms due to their elicitation through laborious experimental studies. Despite the higher coverage of image labels, which gives them in this respect an advantage over norming data, inference methods are needed in both cases if we want to transfer perceptual information to words which have none. We hypothesise that the carefully coded visual attributes are more appropriate than image labels for transferring visual



knowledge to new concepts, since the latter are more susceptible to noise and may also provide non-visual information (e.g., training is used as a label in ESP co-occurring with *bike*).

A natural avenue is the development of semantic representation models that exploit *automatically* induced perceptual information from data that is both naturally occurring and easily accessible, such as images, whilst retaining the features of attribute norms. The latter include interpretability (i.e. they are transcribed in language), cognitive plausibility (i.e. they describe visual phenomena similarly to how humans describe them, Section 2.1), and potential to generalise to new concepts.

# Chapter 4

## Attribute-centric Representation

In the previous chapter we concluded that visual attributes are a valuable source of information for perceptually grounded meaning representations. In the remainder of the thesis we will therefore adopt an attribute-centric approach to meaning representations and focus on the visual modality as a major source of perceptual information. Instead of relying on attribute norms (or image labels), we will use computer vision techniques to automatically obtain visual attribute representations from images. Consequently, this alleviates issues of the former information sources, namely their dependence on human input whatsoever or on inference methods for concepts for which visual information is absent. This automatic, image-based approach furthermore benefits from the fact that image data is ubiquitous and easily accessible.

Our choice of an attribute-centric approach is also motivated by theoretical arguments from cognitive science and computer vision research, as we will outline in Section 4.1. In line with the attribute-centric representation for the visual modality (presented in Section 4.2), we represent the textual modality by means of textual attributes which we automatically extract from text data using an existent distributional method (Baroni et al., 2010, Section 4.3). Analogously to the extraction of visual attributes from images, this approach is scalable to a large number of arbitrary concepts with little effort in contrast to the labour-intensive human-based elicitation of norms.

In experiments on human word association data (Section 4.4), we demonstrate the benefit of these attribute-based representations when using them as input to the models presented in the previous chapter. Specifically, the experiments address the following three questions:

- (1) Do *automatically* predicted visual attributes improve the effectiveness of distributional models?

- (2) Are there performance differences among different models, i.e. are some models better suited to the integration of this type of visual information?
- (3) How do computational models fare against human-produced norming data?

For the sake of clarity it may be helpful to recall the definition of the term *attribute* which we have already introduced in Chapter 1 (Section 1.3). We use the term *attribute* to refer to a semantic property of a concept in natural language. Attributes referring to visually discernible properties are called *visual*. Examples are *furry*, *has legs*, *eats*. Attributes referring to properties that can be mined from text data are called *textual* or *linguistic*. Examples are *a mammal*, *dies*, *gives birth*. By the term *feature* we refer to a measurable property of an object in general as used in machine learning and pattern recognition. If the object is an image, a feature is derived from pixels and can denote, e.g., an edge or an interest point.

## 4.1 Motivation for (Visual) Attributes

From a cognitive perspective, the use of attributes for meaning representations is endorsed by its long-standing tradition in cognitive science, as discussed in Chapters 2 and 3 (Sections 2.1 and 3.1). In brief, attributes are the medium humans naturally use to verbally convey perceptual, taxonomic, sensorimotor, and functional knowledge of concepts in natural language.

From a computer vision perspective, attributes are advantageous for several reasons. In order to describe visual phenomena (e.g., objects, scenes, faces, actions) in natural language, computer vision algorithms traditionally assign each instance a categorical label (e.g., *apple*, *sunrise*, *Sean Connery*, *drinking*). Attributes, on the other hand, offer a means to obtain semantically more fine-grained descriptions. They can transcend category and task boundaries and thus provide a generic description of visual data and, consequently, their depictions (e.g., both *apples* and *balls* are round, *forks* and *rakes* have a handle and have tines). In addition to facilitating inter-class connections by means of shared attributes, intra-class variations can also be captured, hence offering a means to discriminate between instances of the same category (e.g., *birds* can have long beaks or short beaks). Moreover, attributes allow to generalise to new instances for which there are no training examples available. We can thus say something about depicted entities without knowing their object class. This makes attributes efficient, since they obviate the training of a classifier for each category.

From a modelling perspective, attributes occupy the middle ground between non-linguistic (low- or mid-level) image features and linguistic words. More precisely, attributes constitute a medium that is both, machine detectable and human understandable. They crucially represent image properties, however by being words themselves, they can be easily integrated in any text-based model thus eschewing known difficulties with rendering images into word-like units.

## 4.2 Visual Attributes from Images

For the reasons given in the previous section, the use of attributes for computer vision tasks, such as image classification, has experienced a growing interest in recent years. We discuss related work on supervised methods for the prediction of visual attributes from images (Section 4.2.1) along with existing data sets created for this purpose (Section 4.2.2). Since our work differs from previous work in focus and scope, we require a new dataset, and describe its creation in Section 4.2.3. We subsequently explain how we use the dataset to train attribute classifiers (Section 4.2.4), which we then apply to derive visual attribute-based representations for concepts depicted in images (Section 4.2.5).

### 4.2.1 Visual Attributes in Computer Vision

The field of computer vision deals with the automation of visual processing, aiming for computers to understand image data at the same level as humans. In practice, computer vision systems extract information from images useful for solving tasks including, *inter alia*, image classification (i.e. the determination of the classes of the objects present in an image), object detection (i.e. the determination of the class and the location of objects present in an image), or scene classification. Methods addressing such tasks require the images to be represented by means of extracted *features*, where the choice of specific feature types and their representation (*feature descriptors*) is a critical part.

Popular approaches to vision tasks have used local descriptors, such as SIFT (scale-invariant feature transform, Lowe, 2004) or HOG (Histogram of Oriented Gradients, Dalal and Triggs, 2005), which represent images by means of low-level features. Recently, new types of methods have been proposed aiming at more sufficient descriptions of visual content in form of intermediate or mid-level features. One class of such methods derives part-based image representations (e.g., Felzenszwalb et al., 2010).

Another class of approaches learns hierarchical mid-level feature representations. Examples of the latter include bag-of-visual-words (Sivic and Zisserman, 2003) or spatial pyramids (Lazebnik et al., 2006), which build upon low-level feature descriptors by combining them to a more global image representation. More recently, a new class of feature learning methods has emerged which directly operates on the pixel-level using convolutional neural networks (CNNs). CNNs learn hierarchical representations in an unsupervised way using a deep network architecture (see Chapter 5 for more details on the latter). Another class of methods emphasises the need of compact, semantically meaningful intermediate-level representations which are interpretable by both machines and humans, and thus promotes the use of visual attributes (a.k.a. semantic features in the literature).

Initial work on visual attributes for image data (Ferrari and Zisserman, 2007) focussed on simple colour and texture attributes (e.g., blue, stripes) and showed that these can be learned in a weakly supervised setting from images returned by a search engine when using the attribute as a query. Farhadi et al. (2009) were among the first to use visual attributes in an object recognition task. Using an inventory of 64 attribute labels, they developed a dataset of approximately 12,000 instances representing 20 objects from the PASCAL Visual Object Classes Challenge 2008 (Everingham, M. and Van Gool, L. and Williams, C. K. I. and Winn, J. and Zisserman, A., 2008). Visual semantic attributes (e.g., hairy, four-legged) were used to identify familiar objects and to describe unfamiliar objects when new images and bounding box annotations were provided. Lampert et al. (2009) showed that attribute-based representations can be used to classify objects when there are no training examples of the target classes available (a task referred to as *zero-shot learning*), provided their attributes are known. Their dataset contained over 30,000 animal images and used 85 attributes (e.g. brown, stripes, furry, paws) from the norming study of Osherson et al. (1991). Similar work was done by Parikh and Grauman (2011), who use relative attributes indicating their degree of presence in an image compared to other images (e.g. more smiling than). The use of attributes for zero-shot learning was also explored in the context of scene classification (Patterson et al., 2014) and action recognition (Liu et al., 2011).

Russakovsky and Fei-Fei (2010) learned classifiers for 20 visual attributes on ImageNet (Deng et al., 2009) with the goal of making visual inter-category connections across a broad range of classes on the basis of shared attributes (e.g., striped animals and striped fabric). The ability of attributes to capture intra-category variations has in

turn been leveraged in approaches for face verification<sup>1</sup> (Kumar et al., 2011), domain-specific image retrieval (Kumar et al., 2011; Patterson et al., 2014; Rastegari et al., 2013), and fine-grained object recognition (Duan et al., 2012). The use of visual attributes extracted from images in models of semantic representations is novel to our knowledge.

## 4.2.2 Image Collections

A key prerequisite for learning attribute classifiers for images is the availability of training data comprising a large number of images along with attribute annotations. Existing image databases of objects and their attributes focus on a small number of categories (Farhadi et al., 2009), or on a specific category, such as animals (Animals with Attributes, Lampert et al., 2009), birds (Caltech-UCSD Birds-200-2011, Wah et al., 2011), faces (FaceTracer, Kumar et al., 2008), or clothing items (Chen et al., 2012). Some databases provide attribute annotations for scenes (Laffont et al., 2014; Patterson et al., 2014). Other, general-purpose image collections cover a broad range of object categories, but provide no (ESP, von Ahn and Dabbish, 2004; MIR Flickr, Huiskes and Lew, 2008)<sup>2</sup> or little (ImageNet, Russakovsky and Fei-Fei, 2010; Deng et al., 2009; LabelMe, Russell et al., 2008) attribute information.

Since our goal is to develop models that are applicable to many words from different categories, we created a new dataset. It shares many features with previous work (Lampert et al., 2009; Farhadi et al., 2009), but differs in focus and scope, covering a larger number of object classes and attributes. We chose to create the dataset on top of ImageNet due to its high coverage of different objects, its use of the hierarchical structure of WordNet (Fellbaum, 1998) to organise the objects, and the high quality of its images (i.e. cleanly labelled and high resolution). We describe our dataset in Section 4.2.3, but first give a brief overview on WordNet and ImageNet.

### The WordNet and ImageNet Databases

WordNet (Fellbaum, 1998) is an English lexical database which groups synonymous content words (i.e. words denoting the same concept) into sets (*synsets*). The synsets

---

<sup>1</sup>The task of face verification is to decide whether two faces are of the same individual (Kumar et al., 2011).

<sup>2</sup>ESP and MIRFlickr contain image tags which could potentially be used to automatically gather attribute annotations. See, for instance, Sharma and Jurie (2011); Rohrbach et al. (2010) and also Chapter 3 (Section 3.2).

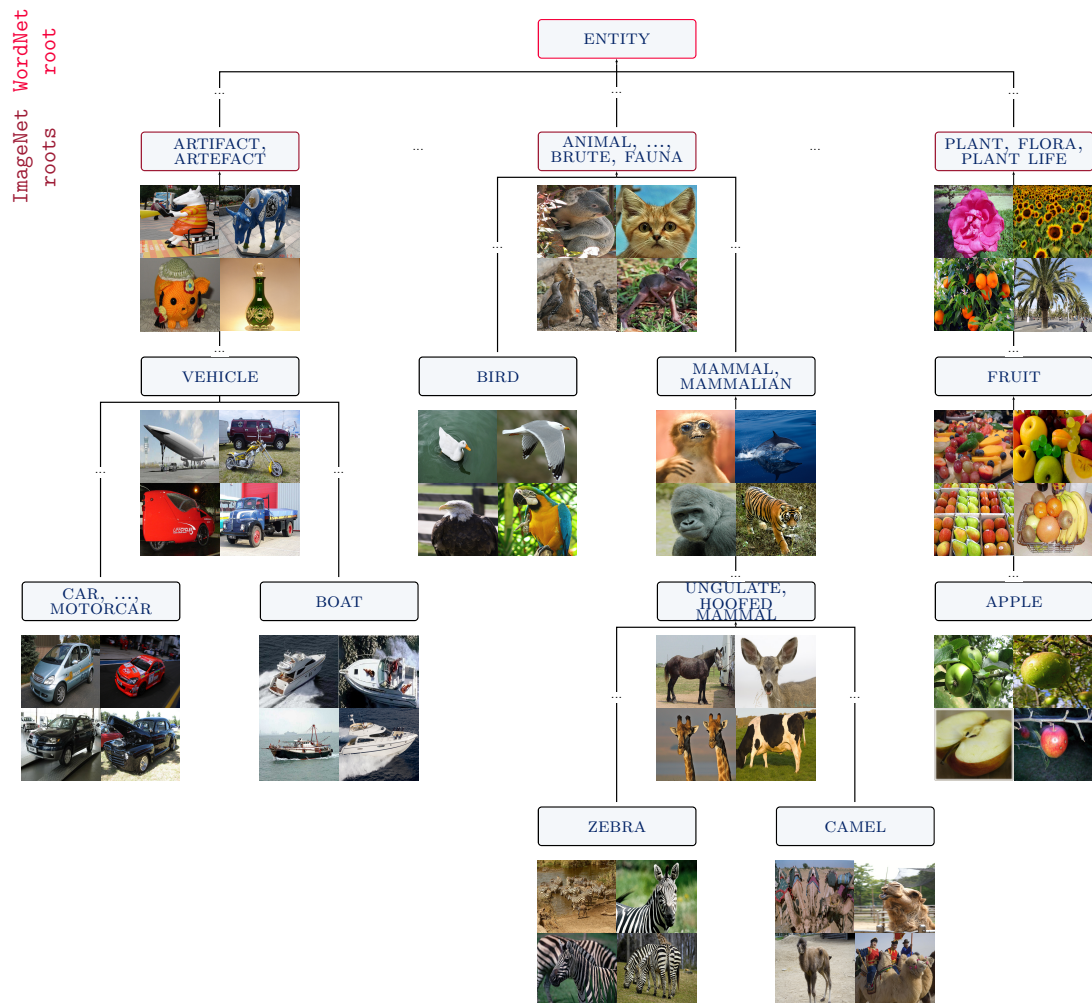


Figure 4.1: Extract of the WordNet/ImageNet hierarchy. Dots represent synsets that we have omitted from the hierarchy (for the sake of brevity).

denoted by words of each individual part-of-speech (PoS) are interlinked by means of semantic relations, such as hypernymy (e.g., synset {*motor vehicle, automotive vehicle*} is a hypernym of {*car, auto, automobile, machine, motorcar*}), or meronymy (e.g., {*car wheel*} is a part of {*motor vehicle, automotive vehicle*}).

ImageNet<sup>3</sup> (Deng et al., 2009) is an ontology of images organised according to the nominal hierarchy of WordNet (Fellbaum, 1998). Figure 4.1 shows an example of the WordNet/ImageNet subnetwork where synsets of nominal concepts are interlinked by hypernymy relations, and are populated with images of the corresponding objects. ImageNet has more than 14 million images assigned to more than 21K WordNet

<sup>3</sup>ImageNet is available at <http://www.image-net.org>.



Figure 4.2: Images from ImageNet for *dog* (synset  $\{dog, domestic\ dog, Canis\ familiaris\}$ , n02084071) and *screwdriver* ( $\{screwdriver\}$ , n04154565) with bounding box annotations (green rectangles).

synsets of all levels of categorisation, i.e. superordinate-level synsets (e.g.,  $\{vehicle\}$ ,  $\{animal, \dots, fauna\}$ , or  $\{plant\}$ ; Figure 4.1), basic-level synsets (e.g.,  $\{boat\}$ ,  $\{zebra\}$ ,  $\{apple\}$ ), and subordinate-level synsets (e.g.,  $\{mountain\ zebra\}$ ,  $\{Granny\ Smith\}$ ,  $\{ferryboat\}$ ). The database provides additional information for a subset of images, including SIFT features (Lowe, 2004), attribute annotations, and bounding box information (see Figure 4.2 for an example of the latter). Currently, each of 3,000 synsets have 150 images on average with bounding boxes, and 400 synsets have images with attribute annotations from a set of 25 attributes. In our experiments we will use the bounding box information when available.

Deng et al. (2009) harvest images for ImageNet for an individual synset by querying several image search engines. These candidate images are then verified by human annotators using Amazon Mechanical Turk (AMT). The annotators are presented with a set of candidate images and the definition of the target synset, and are asked to decide for each image whether it depicts an object of the synset. Every candidate image is labelled by several annotators, where the number of annotators depends on a synset-dependent confidence score threshold that has to be reached.

### 4.2.3 The Visual Attributes Dataset (VISA)

**Concepts and Images** We created the dataset for the nominal concepts contained in the attribute production norms of McRae et al. (2005, henceforth McRae norms), as they cover a wide range of concrete concepts including animate and inanimate things (e.g., animals, clothing, vehicles) and are widely established in cognitive science re-



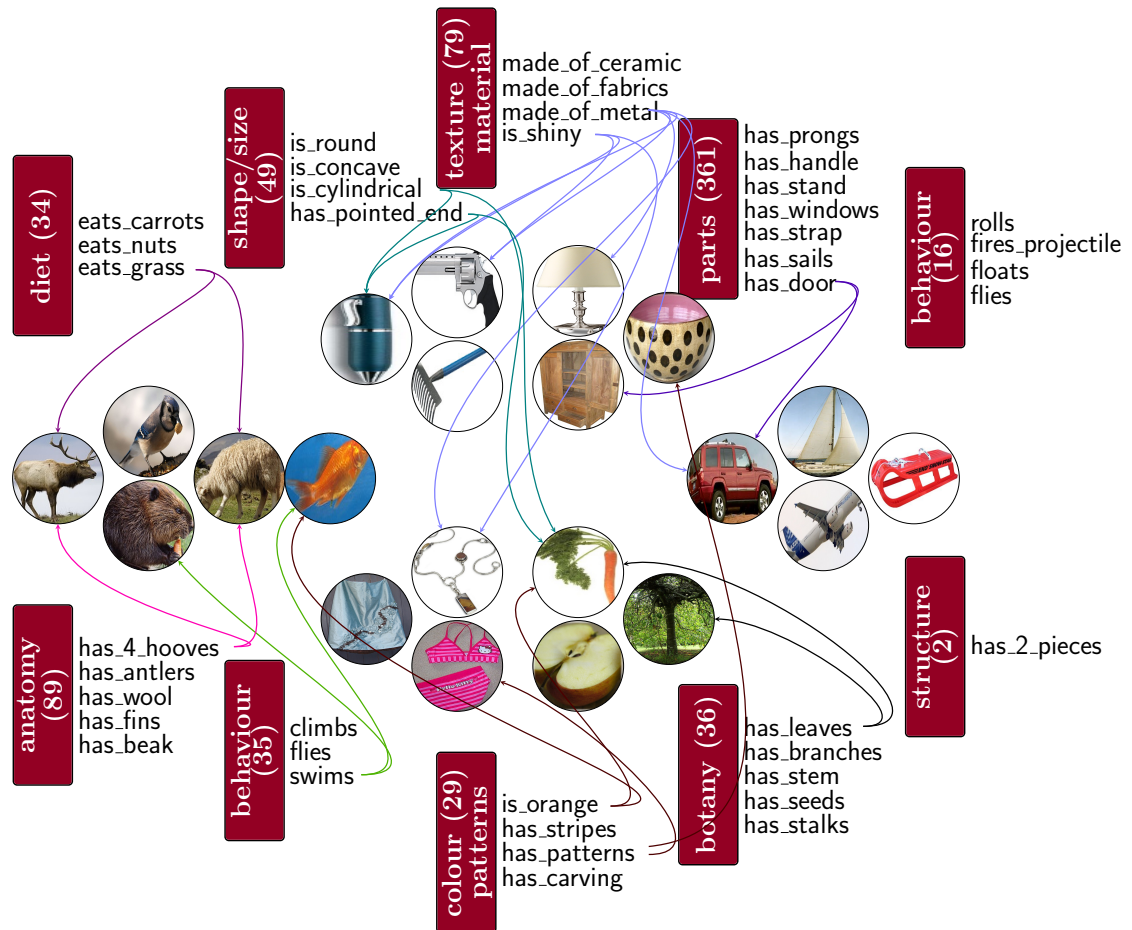


Figure 4.3: Attribute categories and examples of attribute instances and images. Parentheses denote the number of attributes per category.

search (see the description of the norms in Section 3.1).

Images for the concepts in the McRae norms were harvested from ImageNet (Deng et al., 2009, Section 4.2.2). The McRae norms contain 541 concepts out of which 516 appear in ImageNet<sup>4</sup> and are represented by nearly 700K images overall. The average number of images per concept is 1,310 with the most popular being *closet* (2,149 images) and the least popular *prune* (5 images). See Appendix A (Section A.1) for a list of all concepts and the synsets to which we mapped the former in order to gather the corresponding images from ImageNet.

**Attribute Annotation** Our aim was to develop a set of visual attributes that are both discriminating and cognitively plausible in the sense that humans would generally use

<sup>4</sup>Some words had to be modified in order to match the correct synset, e.g., *tank\_(container)* was found as *storage\_tank*.





	anatomy	has_mouth, has_head, has_nose, has_tail, has_claws has_jaws, has_neck, has_snout, has_feet, has_tongue
	behaviour	eats, walks, climbs, swims, runs
	colour_patterns	is_black, is_brown, is_white
	diet	drinks_water, eats_anything
	shape_size	is_tall, is_large
	botany	has_skin, has_seeds, has_stem, has_leaves, has_pulp
	colour_patterns	purple, white, green, has_green_top
	shape_size	is_oval, is_long
	texture_material	is_shiny
	behaviour	rolls
	colour_patterns	different_colors, is_black, is_red, is_grey, is_blue, is_white
	parts	has_4_wheels, has_steering_wheel, has_seat<ne>, has_windows has_engine<ne>, has_mirror, has_number_plate, has_bonnet has_trunk, has_windshield_wiper, has_roof, has_bumper, has_handle has_belts, has_light, has_windshield, has_door, has_brakes<ne>
	texture_material	made_of_metal
	colour_patterns	is_black, is_brown
	parts	has_rest_(musical), has_4_strings, has_bridge_(musical) has_board, has_scroll, has_tail_piece, has_curved_body has_pegs
	shape_size	is_hollow
	texture_material	is_shiny, has_f_holes, made_of_wood

Table 4.2: Human-authored attributes for *bear*, *eggplant*, *car*, and *violin*. <ne> stands for <no\_evidence>.

them to describe a concrete concept. As a starting point, we thus used the visual attributes from the McRae norms. Attributes capturing other primary sensory information (e.g., smell, sound), functional or motor properties, or encyclopaedic information were not taken into account. For example, *is\_purple* is a valid visual attribute for an *eggplant*, whereas *a\_vegetable* is not, since it cannot be visualised. Collating all the visual attributes in the norms resulted in a total of 676. Similar to Lampert et al. (2009) in their creation of the *Animals with Attributes* dataset (see Section 4.2.1), we conducted the annotation on a *per-concept* rather than a *per-image* basis (as for example Farhadi et al., 2009). However, our methodology is slightly different from Lampert et al. (2009) in that we did not simply transfer the attributes from the norms to the concepts in question but modified and extended them during the annotation process explained below,

using a small fraction of the image data as development set (see Section 4.2.4.1 for details on the development set).

For each concept (e.g., *bear* or *eggplant*), we inspected the images in the development set and chose all visual attributes contained in the McRae norms that applied. If an attribute was generally true for the concept, but the images did not provide enough evidence, the attribute was nevertheless chosen and labelled with `<no_evidence>`. For example, a *plum* has `_a_pit`, but most images in ImageNet show plums where only the outer part of the fruit is visible. We added new attributes which were supported by the image data but missing from the initial set as given by the norms. For example, `has_lights` and `has_bumper` are attributes of *cars* but are not included in the norms. In general we were conservative in adding new attributes as our aim was to preserve the cognitive plausibility of the original attribute norms. For this reason, we added entirely new attributes only when we considered them to be on the same level of granularity as the attributes of the McRae norms.

Appendix A (Section A.2) shows an example of the annotation interface that was used for this annotation procedure.

There are several reasons for choosing the described annotation scheme instead of transferring the McRae attributes directly. Firstly, it makes sense to select attributes corroborated by the images. Secondly, by looking at the actual images, we could eliminate errors in the McRae norms. For example, eight study participants erroneously thought that a *catfish* has `_scales`. Thirdly, during the annotation process, we normalised synonymous attributes (e.g., `has_pit` and `has_stone`) and attributes that exhibited negligible variations in meaning (e.g., `has_stem` and `has_stalk`). Finally, our aim was to collect an exhaustive list of visual attributes for each concept which is consistent across all members of a category. This is unfortunately not the case in the McRae norms. Participants were asked to list up to 14 different properties that describe a concept. As a result, the attributes of a concept denote the set of properties humans consider most salient. For example, both, *lemons* and *oranges* have `_pulp`. But the norms provide this attribute only for the second concept.

Annotation proceeded on a category-by-category basis, e.g., first all food-related concepts were annotated, then animals, vehicles, and so on. Two annotators (one of them is the author of this thesis) developed the set of attributes for each category. One annotator first labelled concepts with their attributes as described above, and the other annotator reviewed the annotations, making changes if needed. Finally, annotations were revised and compared per category in order to ensure consistency across all con-

cepts of that category. Attributes were grouped in ten general classes (e.g., anatomy, parts) shown in Figure 4.3.

Overall, we discarded or modified 262 visual attributes of the McRae norms, and added 294 attributes. On average, each concept was annotated with 15 attributes; approximately 11.5 of these were not part of the set of attributes created by the participants of the McRae norms for that concept even though they figured in the attribute sets of other concepts. Furthermore, on average two McRae attributes per concept were discarded. Examples of concepts and their attributes from our database<sup>5</sup> are shown in Table 4.2.

## 4.2.4 Automatically Extracting Visual Attributes

### 4.2.4.1 Data

For each concept in the VISA dataset, we partitioned the corresponding images into a training, development, and test set. For most concepts the development set contained a maximum of 100 images and the test set a maximum of 200 images. Concepts with less than 800 images in total were split into  $1/8$  test and development set each, and  $3/4$  training set. Image assignments to the splits were done randomly in general. However, we wanted the test set to be composed of as many images with bounding box annotations as possible (recall that ImageNet does not provide bounding boxes for each image, see Section 4.2.2). We therefore first assigned images with bounding boxes to the splits, starting with the test set, before assigning the remaining images. To learn a classifier for a particular attribute, we used all images in the training data, totalling to approximately 550K images. Images of concepts annotated with the attribute were used as positive examples, and the rest as negative examples.

### 4.2.4.2 Training Attribute Classifiers

In order to extract visual attributes from images, we follow previous work (Farhadi et al., 2009; Lampert et al., 2009) and learned one classifier for each attribute that had been assigned to at least two concepts in our dataset. We furthermore only considered attribute annotations that were corroborated by the images, that is, we ignored those labelled with `<no_evidence>`. This amounts to 414 classifiers in total.

We used an L2-regularised L2-loss linear support vector machine (SVM, Fan et al.,

---

<sup>5</sup>Available at <http://homepages.inf.ed.ac.uk/s1151656/resources.html>.

2008)<sup>6</sup> to learn the attribute predictions, and adopted the training procedure of Farhadi et al. (2009). We optimised cost parameter  $C$  of each SVM on the training data, randomly partitioning it into a split of 70% for training, and 30% for validation. The final SVM for an attribute was trained on the entire training data, i.e. on all positive and negative examples.

**Features** We used the four different feature types proposed by Farhadi et al. (2009)<sup>7</sup>, namely colour, texture, visual words, and edges. For each feature type, an image (or the image region defined by a bounding box) was represented using a bag-of-words approach.

**Background: Bag-of-visual-words (BoVW) Approach** The bag-of-words approach in computer vision (Sivic and Zisserman, 2003), analogous to the bag-of-words model described in Chapter 2 (Section 2.2), represents an image as a distribution (histogram) over words. In contrast to text processing, the set of words (i.e. the vocabulary or *codebook*) has to be learned automatically from an image collection in form of a set of quantised feature descriptors. This is typically conducted as follows: First, features in the images are detected and subsequently represented as compact numerical vectors (*feature descriptors*), describing local patches of pixels around the feature location. Then, a codebook of size  $k$  is generated by quantising the descriptors to  $k$  words, e.g., by clustering the feature descriptors using  $k$ -means clustering (Sivic and Zisserman, 2003), where each cluster center represents one word.

In order to transform an image to a bag-of-words representation using a given codebook, each feature descriptor of the image is assigned to the nearest word (i.e. cluster center), and the image is then represented as  $k$ -dimensional histogram of word counts resulting from the counted number of descriptors assigned to each word.

Texture descriptors (Varma and Zisserman, 2005) were computed for each pixel and quantised to the nearest 256  $k$ -means centers. Edges were detected using a standard Canny detector and their orientations were quantised into eight bins. Colour descriptors were computed in the LAB (CIE 1976 L\*a\*b\*) colour space. They were

<sup>6</sup>For a given set of instance-label pairs  $(\mathbf{x}_i, y_i), i = 1, \dots, l, \mathbf{x}_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$ , the SVM solves the optimisation problem  $\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \zeta(\mathbf{w}; \mathbf{x}_i, y_i)$ , where the L2-loss function is  $\zeta(\mathbf{w}; \mathbf{x}_i, y_i) = \max(1 - y_i \mathbf{w}^T \mathbf{x}_i, 0)^2$ .

<sup>7</sup>The code by Farhadi et al. (2009) is available at <http://vision.cs.uiuc.edu/attributes/> (last accessed in May 2015).

sampled for each pixel and quantised to the nearest 128  $k$ -means centers.

Visual words were constructed with a HOG (Histogram of Oriented Gradients, Dalal and Triggs, 2005) spatial pyramid. In the spatial pyramid approach (Lazebnik et al., 2006), descriptors are computed on different resolution levels (*scales*) of the image. This scale space is represented in octaves, where an increase of the scale by one octave means roughly halving the resolution of the image, and each octave is a set covering a fixed number of intermediate scale steps. In our case, 2 scales per octave were used. HOG descriptors themselves were computed using  $8 \times 8$  blocks and a 4 pixel step size. This means that an image is decomposed into spatial regions (*cells*), and for each cell a histogram of gradients is created over the pixels it contains, by first computing a gradient vector for each pixel and then quantising the vectors into 9 bins (orientations). Several adjacent cells are grouped into *blocks*, and a HOG descriptor is obtained for each block as a result of concatenating the corresponding histograms and subsequent normalisation. The algorithm follows a sliding window approach, in which a block is created by grouping adjacent cells every *pixel-step-size* pixels. In order to obtain the final visual words, the HOG descriptors were quantised into 1000  $k$ -means centers.

For each of the four feature types, individual histograms were computed for the whole image or a bounding box (if available). With the purpose to represent shapes and locations, six additional histograms were generated for each feature type. These were obtained by dividing the image (or region) into a grid of three vertical and two horizontal blocks, and computing a histogram for each block in the grid separately.

The resulting seven histograms per feature type were individually normalised with the  $l^2$ -norm and then stacked together resulting in the feature vector for an image.

#### 4.2.4.3 Evaluation

Figures 4.4 and 4.5 show classifier predictions for eight test images from concepts seen by the classifiers during training, and eight images from new, i.e. unseen, concepts not part of the VISA dataset, respectively. Most attributes predicted for the images from seen concepts (Figure 4.4) do indeed describe the depicted objects. However, not all of them are actually present in the image (e.g., *has\_tongue* was wrongly predicted for the image showing *rats*). Such mistakes indicate that some classifiers confused features relevant for recognising a particular attribute from features pertaining to correlated attributes (e.g., *has\_snout*). The attribute predictions for images from unseen concepts (Figure 4.5) demonstrate that the classifiers are indeed able to generalise to concepts

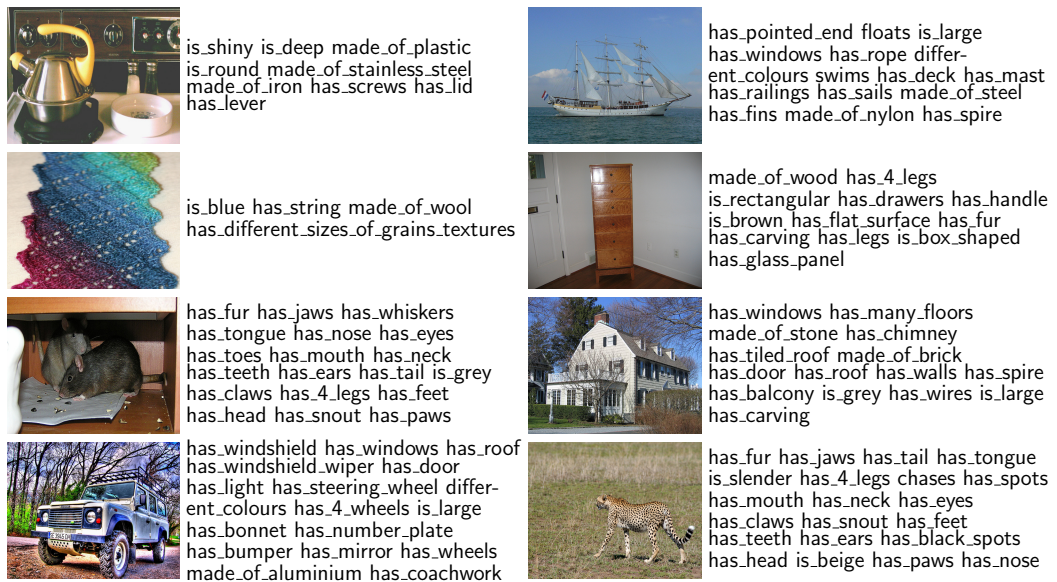


Figure 4.4: Attribute predictions for concepts encountered during training. (From top left to bottom right: *kettle*, *scarf*, *rat*, *jeep*, *yacht*, *bureau*, *house*, *cheetah*.)

they have not encountered during training. However, now, errors are not primarily due to attribute correlations (e.g., *has\_buds* for *ailanthus*), but some predictions are plainly wrong in that they lack a relation to the object class (e.g., *has\_windows* for *basket* or *made\_of\_wood* for *espresso\_maker*).

We also quantitatively evaluated the attribute classifiers by measuring the interpolated average precision (AP, Salton and McGill, 1986) on the test set. Since the reference annotations contained in VISA are concept-based, we perform the evaluation on the basis of concept-level predictions as the centroid of all attribute predictions for the images belonging to the same concept (see Section 4.2.5 for details on how we compute the concept-level predictions); specifically, we plot precision against recall based on a threshold.<sup>8</sup> Recall is the proportion of correct attribute predictions whose prediction score exceed the threshold to the true attribute assignments given by VISA. Precision is the fraction of correct attribute predictions to all predictions exceeding the threshold. The AP is then the mean of the maximum precision at eleven recall levels  $[0, 0.1, \dots, 1]$ . The precision/recall curve is shown in Figure 4.6; the attribute classifiers achieved a mean AP of 0.52.

<sup>8</sup>Threshold values ranged from 0 to 0.9 with 0.1 stepsize.



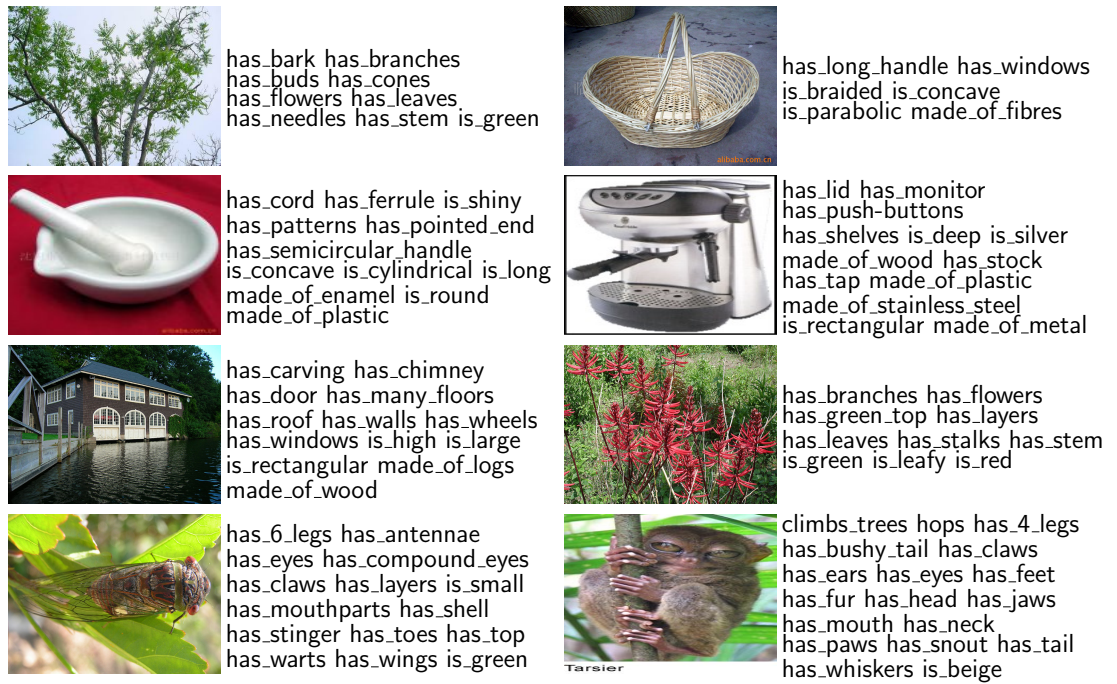


Figure 4.5: Attribute predictions for concepts not encountered during training. (From top left to bottom right: *ailanthus*, *mortar*, *boathouse*, *cicada*, *shopping basket*, *espresso maker*, *coraltree*, *titi monkey*.)

#### 4.2.5 Deriving Visual Representations of Concepts

Note that the classifiers predict attributes on an image-by-image basis; in order to describe a concept  $w$  by its visual attributes taking into account multiple images representing  $w$ , we need to aggregate their attributes into a single representation. We use a vector-based representation where each attribute corresponds to a dimension of an underlying semantic space and concepts are represented as points in this attribute space. Just as in text-based semantic spaces (see Chapter 2, Section 2.2.1), we can thus quantify similarity between two concepts by measuring the geometric distance of their vectors. Since we encode visual attributes, however, the underlying semantic space is perceptual, and so is the similarity we can measure.

We construct visual vector representations as follows. For each image  $x_w \in I_w$  of concept  $w$ , we output an  $A$ -dimensional vector containing prediction scores  $\text{score}_a(x_w)$  for attributes  $a = 1, \dots, A$ .<sup>9</sup> We transform these attribute vectors into a single vector  $\mathbf{p}_w \in \mathbb{R}^{1 \times A}$ , by computing the centroid of all vectors for concept  $w$ . That is, we

<sup>9</sup>For simplicity, we use the symbol  $w$  to denote both, the concept and its index. Analogously, symbol  $a$  denotes the attribute and its index.



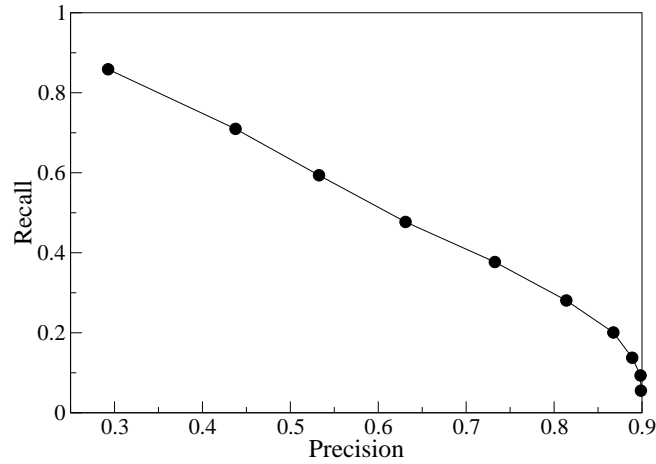


Figure 4.6: Attribute classifier performance for different thresholds  $\delta$  (test set).

average the scores for the various attributes:

$$\mathbf{p}_w = \left( \frac{1}{|I_w|} \sum_{x_w \in I_w} \text{score}_a(x_w) \right)_{a=1, \dots, A} \quad (4.1)$$

The construction process is illustrated in Figure 4.7 by the example concepts *chick* and *balloon*. In Table 4.3 (second column) we give the eight nearest neighbours for seven example concepts (first column) from our dataset. Nearest neighbours for a concept were found by measuring the cosine similarity (Equation (2.4), Page 18) between the visual attribute vectors  $\mathbf{p}$  of that concept and all other concepts in our dataset and choosing the ten concepts with the highest similarity. The examples show that the visual attribute representation is able to capture semantic similarity, attesting words of the same semantic category (e.g., vehicles or animals) as closest nearest neighbours. For comparison, the table also shows the eight nearest neighbours when the example concepts are represented by their textual attribute vectors (Table 4.3, third column) whose creation is discussed in the following section, and by their bimodal vector representations as learned with our bimodal stacked autoencoder (SAE) model (Table 4.3, last column), which will be described in Chapter 5 (Section 5.3).

### 4.3 Textual Attributes

Leveraging text corpora for the automatic extraction of attributes that describe concepts in a comparable way as attribute norms do has been the objective of several recently proposed approaches.

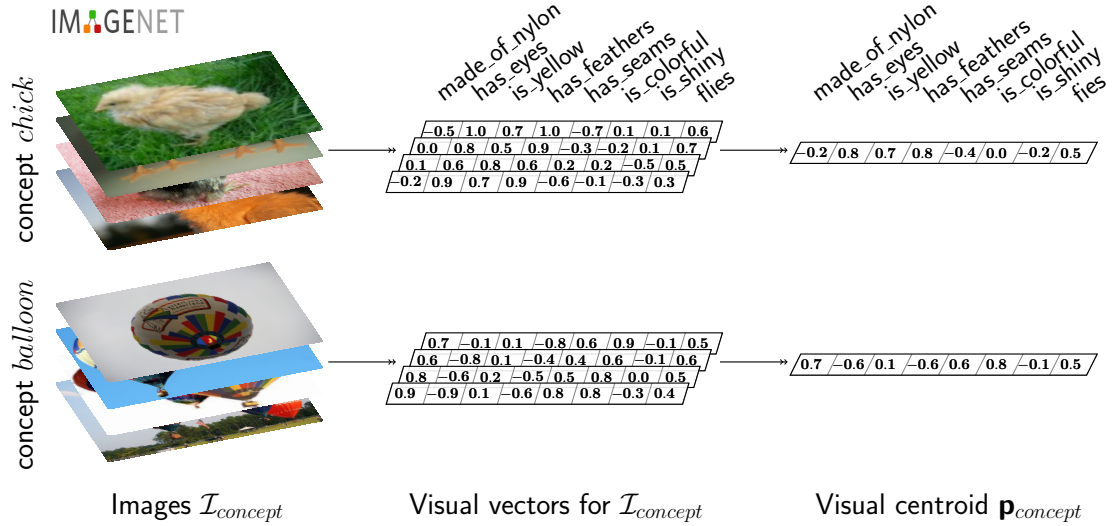


Figure 4.7: Construction of the visual representation for the concepts *chick* and *balloon*. Attribute classifiers predict attributes for example images depicting *chicks* and *balloons*. These prediction scores are then converted into vectors (first arrow). To compute a single visual attribute vector for the concepts, all vectors are aggregated into  $\mathbf{p}_{chick}$  and  $\mathbf{p}_{balloon}$ , respectively, according to Eq. (4.1) (second arrow).

Barbu (2008) used shallow methods in a bootstrapping approach on the basis of the McRae norms, where attributes were extracted by means of a combination of a pattern-based method and co-occurrence association measures. Devereux et al. (2009) extracted concept–relation–attribute triples from dependency-parsed corpora (e.g., *chick–be–bird*, *screwdriver–has–blade*), with the relation corresponding to the verb that occurs within the dependency path between the concept and the attribute. Their method relied on external semantic knowledge, WordNet (Fellbaum, 1998) and the McRae norms, from which they derived conditional probabilities of attribute classes given concept categories (e.g.,  $P(\text{body part}|\text{REPTILES})$ ) in order to re-rank and filter extracted candidate triples. Kelly et al.’s (2010) extension of Devereux et al.’s (2009) system additionally relied on manually generated extraction rules, leading to an improved precision (see also Kelly et al., 2014). Kelly et al. (2012) and Kelly et al. (2013) applied semi-supervised and minimally supervised methods, respectively, for concept–relation–attribute extraction.

A fully unsupervised template-based approach was proposed by Baroni et al. (Strudel, 2010) which extracts weighted concept–attribute pairs (e.g., *chick–bird:n*, *chick–brood:v*) from a text corpus. We opt for using Strudel to obtain textual attributes for concepts

Concept	Nearest Neighbours		
	Visual	Textual	Bimodal (SAE)
<i>ambulance</i>	<i>van truck taxi bus limousine jeep car train</i>	<i>helicopter trolley van taxi train truck scooter tricycle</i>	<i>taxi van truck bus train trolley limousine scooter</i>
<i>bison</i>	<i>ox bull pony elephant bear cow camel calf</i>	<i>elk buffalo deer caribou bear otter pig pony</i>	<i>buffalo bear elephant caribou deer sheep pig elk</i>
<i>brush</i>	<i>paintbrush pencil ladle hammer screwdriver pin bow_(weapon) hook</i>	<i>comb paintbrush vest scissors doll coat bag pencil</i>	<i>comb paintbrush pen- cil scissors razor pen screwdriver skis</i>
<i>dress</i>	<i>robe blouse camisole nightgown vest hose_(leggings) cloak pants</i>	<i>gown shirt skirt blouse jacket robe pants jeans</i>	<i>gown blouse robe skirt nightgown pants jeans vest</i>
<i>hut</i>	<i>shed shack barn cabin house church chapel cathedral</i>	<i>shack cottage bungalow cabin tent house build- ing barn</i>	<i>shack cabin house cottage bungalow barn apartment tent</i>
<i>microwave</i>	<i>oven shelves stove cabi- net freezer radio bureau bin_(waste)</i>	<i>stove oven freezer radio pot colander pan squid</i>	<i>radio stove oven freezer stereo fridge telephone dishwasher</i>
<i>scarf</i>	<i>gloves shawl socks sweater veil pajamas doll cap_(hat)</i>	<i>shawl sweater cloak veil gown robe vest coat</i>	<i>shawl sweater pajamas skirt socks veil cape cloak</i>

Table 4.3: Seven example concepts (column 1) and their eight most similar concepts computed on the basis of visual and textual attribute-based representations (columns 2 and 3, respectively) and bimodal representations learned by the SAE model (column 4) in order of decreasing cosine similarity.

due to its knowledge-lean approach—it merely expects PoS-tagged input—and the fact that it has a bias towards non-perceptual attributes such as actions, functions or situations (Baroni et al., 2010).

### 4.3.1 Textual Attributes from Strudel

In Baroni et al.’s (2010) system *Strudel*<sup>10</sup>, weighted word–attribute pairs (e.g., *chick–bird:n* (60.1), *chick–brood:v* (67.5), *chick–precocial:j* (45.8)) are extracted from a PoS-tagged and optionally lemmatised corpus in a fully automatic and unsupervised way. The weight between a word and an attribute expresses their strength of association. The attributes are not known a priori, but are directly acquired from the corpus. *Strudel* is

<sup>10</sup>The software is available at <http://clic.cimec.unitn.it/strudel/> (last accessed in May 2015).

Attribute	llr ( <i>chick</i> )	llr ( <i>snake</i> )	llr ( <i>cake</i> )	llr ( <i>bread</i> )
<i>bake:v</i>	–	1.5	540.0	541.3
<i>eat:v</i>	23.4	253.4	244.3	539.8
<i>hatch:v</i>	422.5	9.9	–	–
<i>nest:n</i>	413.0	26.5	–	–
<i>venom:n</i>	–	405.1	–	–
<i>flour:n</i>	–	–	244.3	403.3
<i>bite:v</i>	3.7	278.2	7.8	1.9
<i>sandwich:n</i>	–	–	8.1	265.7
<i>feed:v</i>	216.5	115.6	5.3	7.7
<i>kill:v</i>	51.08	172.4	-1.0	-2.7
<i>toast:v</i>	–	–	14.3	142.0
<i>dessert:n</i>	–	–	133.8	21.7
<i>prey:n</i>	–	128.5	–	–
<i>sugar:n</i>	–	–	114.0	15.2
<i>hiss:v</i>	–	93.3	–	–
<i>reptile:n</i>	–	90.1	–	–
<i>bird:n</i>	60.1	35.4	–	0.7
<i>egg:n</i>	59.6	44.9	40.1	31.3
<i>precocial:j</i>	45.8	–	–	–

Table 4.4: Extracted attributes and their weights (llr for log-likelihood ratio) for the concepts *chick*, *snake*, *cake*, and *bread*. Sorted in descending order of llr scores.

an instance of distributional models of meaning (Chapter 2, Section 2.2), but unlike the majority of these models, it induces meaning representations that describe a concept by means of its attributes instead of a bag of co-occurring words or text passages (see Chapter 2, Section 2.2.1).

Given a list of nominal target concepts, Strudel first scans a PoS-tagged corpus to identify, for each target concept, potential attributes which can be nouns, adjectives, or verbs co-occurring with the concept, using a set of 15 general rules. These rules act as filter for plausible attributes and are imposed on the tokens connecting the potential attribute–target concept pairs (e.g., *if adjective follows concept, connecting tokens must contain be; the connecting tokens can contain maximally one noun*). The connecting tokens are extracted along with the pairs in a generalised form, where all content words, aside from a few exceptions, have been replaced by their PoS tag. The distinct generalised connectors are subsequently used to obtain a score for each attribute–concept pair that denotes the importance of the potential attributes for the concept and can thus be used to rank the candidate pairs. The ranking approach is motivated by the assumption that true semantic attributes co-occur with the target in various constella-

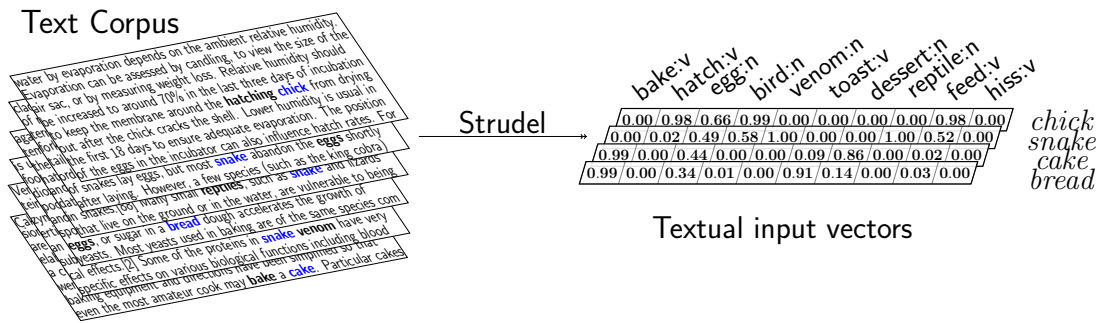


Figure 4.8: Construction of textual semantic space with Strudel for four example concepts (*chick*, *snake*, *cake*, and *bread*).

tions, whereas simply co-occurring words do not necessarily. To this end, the score is computed by measuring a log-likelihood ratio statistic (Dunning, 1993) based on the number of connector types (instead of tokens), with the results that potential attributes which are connected with a concept by various patterns receive a higher score than those connected (possibly frequently) by a single pattern. Strudel furthermore uses the generalised connectors to compute type sketches for concept-attribute pairs. As we will not make use of them we refer the interested reader to Baroni et al. (2010) for more details. Table 4.4 gives examples of attributes extracted for the concepts *chick*, *snake*, *cake*, *bread*, weighted by their log-likelihood scores (llr).

We convert the output from Strudel into textual meaning representations by constructing a textual semantic space over the extracted attributes, subject to some additional attribute filtering, and representing each target word as vector of that space, with the corresponding log-likelihood ratio scores as its entries. Figure 4.8 illustrates this with an example.

In accordance with the terminology for the visual modality, we will henceforth refer to the Strudel attributes as textual attributes.

## 4.4 Experiment 4: Grounding Lexical Models with Attributes

We evaluate the effectiveness of our attribute classifiers presented in Section 4.2.4 for visually grounding meaning representations in a preliminary experiment. Specifically, we integrate their predictions with textual attribute representations using the models described in Chapter 3 (Section 3.3), namely the attribute-topic model (Section 3.3.1),

the model based on kernel canonical correlation analysis (kCCA, Hardoon et al., 2004) (Section 3.3.3), and the global similarity model (Section 3.3.2). Recall, however, that the latter places emphasis on the inference of missing perceptual information from the linguistic modality, which is not necessary here since we acquire visual information through the attribute classifiers. We will therefore just retain the model’s integration mechanism consisting of the concatenation of the vectors representing the two modalities, and call it *concatenation model* instead (cf. Chapter 2, Figure 2.5, Page 26).

#### 4.4.1 Data

We represent the visual modality by the attribute vectors computed as shown in Equation (4.1) (Page 56) using the concepts and their images contained in the VISA dataset. The linguistic environment is approximated by textual attribute vectors derived from Strudel as explained in the previous section. We learned the underlying word–attribute pairs from a lemmatised and PoS-tagged (2009) dump of the English Wikipedia.<sup>11</sup>

Recall that the attribute-topic model requires a distribution over visual attributes from which it can sample attribute observations for a given word (cf. Chapter 3, Section 3.3.1). We thus normalise the attribute vector  $\mathbf{p}_w$  of each word  $w$  by dividing through the sum of its values:

$$\hat{\mathbf{p}}_w = \frac{\tilde{\mathbf{p}}_w}{\sum_{i=1}^A \tilde{p}_{wi}}, \quad (4.2)$$

where  $A$  denotes the number of attributes and  $\tilde{\mathbf{p}}_w$  is derived from  $\mathbf{p}_w$  by setting all scores less than a threshold  $\delta$  to 0:

$$\tilde{p}_{wi} = \begin{cases} p_{wi} & \text{if } p_{wi} \geq \delta \\ 0 & \text{if } p_{wi} < \delta \end{cases} \quad (4.3)$$

Training data for the attribute-topic model is a corpus  $\mathcal{D}$  of *textual* attributes. Each attribute is represented as a bag-of-concepts, i.e. words demonstrating the property expressed by the attribute (e.g., *vegetable:n* is a property of *eggplant*, *spinach*, *carrot*), coupled with their attribute weight. For some of these concepts, our classifiers predict visual attributes. In this case, the corresponding concept  $w$  is paired with one of its visual attributes, sampled from the distribution given by  $\hat{\mathbf{p}}_w$ . An example would be the representation of *anchor:n*  $\in \mathcal{D}$  through *boat,has\_windows:71*, *rope:75*, *yacht,has\_sails:71*, *chain:38*.

<sup>11</sup>The corpus can be downloaded at <http://wacky.sslmit.unibo.it/doku.php?id=corpora> (last accessed in May 2015).

The kCCA model (Section 3.3.3) receives as input a visual matrix,  $\mathbf{V} \in [0, 1]^{N \times A}$ , where each row corresponds to a normalised visual attribute vector (see Equation (4.2)) representing one of  $N$  target words, and a textual matrix,  $\mathbf{T} \in \mathbb{R}^{N \times D}$ , containing normalised textual attribute vectors as rows. Normalisation in the case of the textual modality is performed analogously to the visual modality, as given in Equation (4.2).  $D$  and  $A$  denote the number of textual and visual attributes, respectively.

The concatenation of matrices  $\mathbf{T}$  and  $\mathbf{V}$  directly yields the concatenation model.

#### 4.4.2 Evaluation Task

As in Chapter 3 (Section 3.4.1), we evaluate the three models on the word association norms collected by Nelson et al. (1998). The norms contain 63,619 unique cue-associate pairs. Of these, 435 pairs are covered by the VISA dataset and our models. We also experiment with 1,716 pairs that are *not* part of the McRae norms but belong to categories covered by our attribute taxonomy (e.g., ANIMALS, VEHICLES), and are present in our text corpus and ImageNet. Using correlation analysis, we examine the degree of linear relationship between the human cue-associate probabilities and the automatically derived similarity values.

#### 4.4.3 Model Parameters

In order to integrate the visual attributes with the models described in Section 3.3 we must select the appropriate threshold value  $\delta$  (see Equation (4.3)). We optimised this value on the development set of VISA (Section 4.2.4.1) and obtained best results with  $\delta = 0$ . We also experimented with thresholding the attribute prediction scores for individual images and with excluding attributes with low precision. In both cases, we obtained best results when using all attributes.

We could apply CCA to the vectors representing each image separately and then compute a weighted centroid on the projected vectors. We refrained from doing this as it involves additional parameters and assumes input different from the other models.

With regard to the textual attributes, we obtained a 9,394-dimensional semantic space after discarding word-attribute pairs with a log-likelihood ratio score less than 19.<sup>12</sup> We also discarded attributes co-occurring with less than two different words. For the attribute-topic model, the number of predefined components  $C$  was set to 10. We adopted the similarity measures used in the experiments of Chapter 3 (Section 3.4): In

<sup>12</sup>Baroni et al. (2010) use a similar threshold of 19.51.

	Nelson	Concat	kCCA	TopicAttr	TextAttr
Concat	0.24				
kCCA	0.30	0.72			
TopicAttr	0.26	0.55	0.28		
TextAttr	0.21	0.80	0.83	0.34	
VisAttr	0.23	0.65	0.52	0.40	0.39

Table 4.5: Correlation matrix for seen Nelson et al. (1998) cue-associate pairs and five models. All correlation coefficients are statistically significant ( $p < 0.01$ ,  $N = 435$ ). Correlation was measured using Spearman’s  $\rho$ .

the kCCA and concatenation model, we measured the similarity between two words using the cosine similarity (Section 2.2.1); in the attribute-topic model, similarity was measured as defined by Griffiths et al. (2007b) (see Chapter 3, Equation (3.7), Page 42), where the underlying idea is that word association can be expressed as a conditional distribution.

#### 4.4.4 Results & Discussion

Our results are broken down into seen (Table 4.5) and unseen (Table 4.6) concepts. The former are known to the attribute classifiers and form part of VISA, whereas the latter are unknown and are not included in the McRae norms (or VISA). We report the correlation coefficients (Spearman’s  $\rho$ ) we obtain when human-derived cue-associate probabilities (Nelson et al., 1998) are compared against the simple concatenation model (Concat), kCCA, and Andrews et al.’s (2009) attribute-topic model (TopicAttr). Table 4.7 displays the corresponding coefficients obtained with Pearson’s  $r$ . We also report the performance of a model that is based solely on the output of our attribute classifiers, i.e. without any textual input (VisAttr), and conversely the performance of a distributional model that uses textual attributes only without any visual input (TextAttr). The results in Tables 4.5 and 4.6 are displayed as a correlation matrix so that inter-model correlations can be observed.

As can be seen in Tables 4.5 and 4.7 (second column), two modalities are in most cases better than one when evaluating model performance on seen data. Differences in correlation coefficients between models with two versus one modality are all statistically significant ( $p < 0.01$  using a  $t$ -test), with the exception of Concat when compared



	Nelson	Concat	kCCA	TopicAttr	TextAttr
Concat	0.11				
kCCA	0.15	0.66			
TopicAttr	0.17	0.69	0.48		
TextAttr	0.11	0.65	0.25	0.39	
VisAttr	0.13	0.57	0.87	0.57	0.34

Table 4.6: Correlation matrix for unseen Nelson et al. (1998) cue-associate pairs and five models. All correlation coefficients are statistically significant ( $p < 0.01$ ,  $N = 1,716$ ). Correlation was measured using Spearman's  $\rho$ .

Models	Seen	Unseen
Concat	0.18	0.12
kCCA	0.24	0.17
TopicAttr	0.19	0.32
TextAttr	0.14	0.10
VisAttr	0.17	0.13

Table 4.7: Model effectiveness on seen and unseen Nelson et al. (1998) cue-associate pairs. Correlation was measured using Pearson's  $r$ . All correlation coefficients are statistically significant ( $p < 0.01$ ,  $N = 435$  seen and  $N = 1,716$  unseen).

against VisAttr. It is also interesting to note that TopicAttr is the least correlated model when compared against other bimodal models or single modalities (Table 4.5). This indicates that the latent space obtained by this model is most distinct from its constituent parts (i.e. visual and textual attributes). Perhaps unsurprisingly, Concat, kCCA, VisAttr, and TextAttr are also highly intercorrelated.

On unseen pairs (see Tables 4.6 and 4.7), Concat fares worse than kCCA and TopicAttr. kCCA and TopicAttr are significantly better than TextAttr and VisAttr ( $p < 0.01$ ). This indicates that our attribute classifiers generalise well beyond the concepts found in our database and can produce useful visual information even on unseen images. Compared to Concat and kCCA, TopicAttr obtains a better fit with the human association norms on the unseen data. The reason for this might be that TopicAttr benefits more from the size of the unseen data which is larger than the seen data. TextAttr

Models	Seen	
	$\rho$	$r$
All Attributes	0.28	0.25
Text Attributes	0.20	0.19
Visual Attributes	0.25	0.23

Table 4.8: Model effectiveness on seen Nelson et al. (1998) cue-associate pairs; models are based on gold human generated attributes (McRae et al., 2005). All correlation coefficients are statistically significant ( $p < 0.01$ ,  $N = 435$ ).

performs worse on the unseen data, which could be a result of our attribute selection in which we discarded attributes co-occurring with less than two different words contained in VISA.

To assess how computational models fare against human-produced norming data, we obtained distributional models from the McRae norms and measured their effectiveness on predicting Nelson et al.’s (1998) word-associate similarities. Each concept was represented as a vector with dimensions corresponding to attributes generated by participants of the norming study. Vector components were set to the frequency with which participants generated the corresponding attribute when presented with the concept. We measured the similarity between two words using the cosine similarity. Table 4.8 presents results for different model variants which we created by manipulating the number and type of attributes involved. The first model uses the full set of attributes present in the norms (All Attributes). The second model (Text Attributes) uses all attributes but those classified as visual (e.g., functional, encyclopaedic, auditory), representing the correspondence to our textual attributes. The third model (Visual Attributes) considers solely visual attributes.

We observe a similar trend as with our computational models. Taking visual attributes into account increases the fit with Nelson’s (1998) association norms, whereas visual and non-visual attributes on their own perform worse. Interestingly, kCCA’s performance is comparable to the All Attributes model (see Tables 4.5 and 4.7, second column), despite using automatic attributes (both textual and visual).

In summary, our results demonstrate that the integration of visual attributes to a distributional (corpus-based) model improves its effectiveness across the board. On the word association task, kCCA and the attribute-topic model give a better fit to human data when compared against simple concatenation and models based on a single

modality. KCCA consistently outperforms the attribute-topic model on seen data (its effectiveness is in fact comparable to the model that uses the human-generated McRae norms), whereas the attribute-topic model generalises better on unseen data (see Tables 4.5, 4.7, 4.8, and 4.6).

## 4.5 Conclusions

In this chapter, we presented the VISA dataset and described how we used it to learn classifiers which predict the absence or presence of visual attributes in images. We explained how these classifier predictions can ground the meaning of words in terms of the visual attributes of their real-world referents. We showed the effectiveness of the classifiers for learning visually grounded representations by integrating their predictions with textual attribute information, and evaluating the obtained representations on a word association task (Section 4.4). We used the same integration models as in the experiments conducted in Chapter 3 (Section 3.4). The results confirmed the superiority of the two joint models—the kCCA (Section 3.3.3) and the attribute-topic (Section 3.3.2) model—over the concatenation approach (see Section 4.4.4).

However, both joint models have shortcomings with respect to our desiderata for models of perceptually grounded meaning representations (see Chapter 3, Section 3.5). Recall that the attribute-topic model induces attribute-topic components from a corpus collection of words and their attributes in a generative process (see Section 3.3.1). It is not possible to embed out-of-vocabulary words (i.e. words which have not occurred in the corpus collection) in the bimodal space constructed over the inferred components. Likewise, its ability to infer missing perceptual information is limited to known words. On the other hand, our experimental results showed that the attribute-topic model is more effective on words unseen to the classifiers than kCCA. The kCCA model in turn is able to project new words not part of its training data into the space, and can even operate on just one modality by projecting a new word’s input vector onto the corresponding basis (see Section 3.3.3). However, it cannot infer information on the missing modality.

In the following chapter, we will introduce a new model which aims to combine the merits of the methods discussed above. More precisely, the model has the following properties, among others: (1) Like TopicAttr and kCCA, it uses a joint mechanism to integrate the modalities (Figure 2.5 (d), Page 26). (2) Like kCCA, it can yield meaning representations for new words and, (3), deal with missing modalities in the sense

that it allows to map concepts into the bimodal space for which only one modality is given. Moreover, due to our use of natural language attributes, it would be particularly desirable to be able to infer information rendered in one modality (e.g., textual attributes), given information of the respective other modality (e.g., image representations of new concepts). A model potentially owning this property could be used in applications going beyond those benefiting from meaning representations, e.g., image classification or description. Whereas the architecture of our model has the potential for this kind of inference, it figures difficult for both the kCCA and the attribute-topic model (cf., Chapter 3, Table 3.4, Page 46).

Among the models discussed so far (Concat, kCCA, TopicAttr) we will keep Concat (as an instance of a concatenation approach) and kCCA (as an instance of a simple joint approach) for comparison to our new model in the following chapters. We drop TopicAttr due to its limitation to concepts it has encountered during training.

# Chapter 5

## Visually Grounded Semantic Representations with Autoencoders

In Chapter 3, we experimentally compared different models of semantic representations which compute word meaning by integrating linguistic and visual information, and discussed their strengths and shortcomings. The models were augmented with visual information obtained from human input. We concluded that chapter by (i) listing desiderata with respect to the properties models of visually grounded meaning representations should fulfil, and (ii) pointing out the necessity of automatically obtaining visual information. In the previous chapter, we described our approach to automatically deriving visual and textual attribute-centric representations. In this chapter, we will present a novel model for visually grounded meaning representations which was designed with the above desiderata in mind. It applies deep learning techniques in a neural network architecture for modality integration, using our attribute-centric representations as input.

When presented with a mass of signals (e.g., sensory data such as visual or auditory signals), an essential key for making use of this input is to be able to capture the critical structure of its patterns. What humans achieve efficiently and robustly, deep learning tries to accomplish by inducing distributed representations (or features) of input data that are organised into multiple levels of non-linearity. Artificial neural networks, biologically inspired by our knowledge of the human brain, are the major kind of architectures used for deep learning (e.g., Arel et al., 2010; Hinton, 2007). Such connectionist approaches have had a long tradition in cognitive science, but only recent achievements (LeCun et al., 1990; Hinton and Salakhutdinov, 2006; Hinton et al., 2006; Ranzato et al., 2006) and advances in computer technology have rendered struc-

ture learning in *deep* architectures feasible which has subsequently led to the emergence of a new research area in machine learning. Since then, deep neural networks (NNs) have proven overwhelmingly successful in various tasks, as will be outlined in Section 5.1.1 (see also, e.g., Bengio et al., 2013, and the references therein).

There are several reasons why deep NNs are so powerful: Representation learning of the input data is accomplished in an unsupervised fashion. That is, features are not manually engineered for nor guided by a specific task at hand. Instead, regularities in large amounts of unlabelled data drive the learning process, whose goal is to unravel the factors underlying the structure of the data, and discard noise or other kinds of irrelevant information. Another crucial aspect of deep architectures is that they are arranged in multiple layers, where each layer evokes a representation composed by non-linear transformations of the patterns it receives as input from the previous layer. This potentially gives rise to learning a hierarchy of feature abstractions, corresponding to basic-level features on the lower layers to more abstract features on the higher levels.

Learning representations independently of a particular task also facilitates their use for a whole range of different tasks (e.g., Collobert et al., 2011). Once a good representation of the input, i.e. one with high expressive power, is learned (*unsupervised pre-training*), it can be refined (*fine-tuned*) for solving a specific task by means of an appropriate supervised criterion and labelled data.

Furthermore, (deep) NNs allow the composition of multiple types of information, e.g., different sensory sources, as has been demonstrated by various researchers. We give an overview of relevant work in this area in Section 5.1.2 (see also Chapter 2, Section 2.5). Finally, NNs can be used as a means for *non-linear* dimensionality reduction by decreasing the number of units in at least one of the internal layers. Autoencoders, presented in Section 5.2, were traditionally used for this purpose.

We will use deep learning techniques in a stacked autoencoder architecture to project linguistic and visual information onto a unified representation that fuses the two modalities together (cf. Figure 2.5 (d), Page 26). We introduce the details of our model in Section 5.3.

## 5.1 Deep Learning in Artificial Neural Networks

An artificial neural network (henceforth just *network*) consists of layers of units, where units of adjacent layers are fully<sup>1</sup> connected by weights. The basic network has a layer of *input* units and a layer of *output* (or *target*) units. *Deep* networks contain multiple *hidden* (a.k.a. *latent*) layers in between the input and output layer (see Figure 2.4, Page 20, for an illustration of a basic neural network with one hidden layer).

### 5.1.1 (Deep) Neural Networks

The application of deep network architectures is gaining increasing popularity for diverse tasks of natural language processing and computer vision, examples of which we will briefly give in this section. Please refer to, e.g., Deng (2014) or Bengio et al. (2013) for a more comprehensive overview.

As discussed in Section 2.3, neural networks have been employed on unlabelled text corpora to induce distributed lexical representations which capture word co-occurrence statistics. A great deal of work has initialised deep networks with these representations which used them to produce sentence representations in order to address NLP tasks, such as PoS-tagging, named entity recognition, semantic role labelling, chunking, (Collobert and Weston, 2008; Collobert et al., 2011), parsing, paraphrase detection, and sentiment analysis (Socher, 2014). The architecture of Collobert and Weston (2008) performs convolution on word windows of the sentence, whereas Socher (2014) presents deep learning architectures based on recursive neural networks (Goller and Küchler, 1996; Pollack, 1990). The latter operate on hierarchically structured data, where the same network model is recursively applied at every node of the structure. Socher's (2014) models can thus produce sentence representations by combining word representations according to the structure of the sentences (e.g., parse trees). Glorot et al. (2011), addressing domain-adaptation for sentiment classification, employ stacked denoising autoencoders (Vincent et al., 2010; see Section 5.2) to learn feature extractions from unlabelled review texts and use these to train sentiment classifiers on one source domain. In contrast to the former approaches, they do not make use of pre-trained word embeddings, but simply encode the input texts directly by binary vectors over n-grams.

All previously mentioned architectures are instances of *feed-forward* NNs which

---

<sup>1</sup>An exception are convolutional layers, in which each unit is connected to a *small subset* of contiguous units of the previous layer, where the weights are usually replicated over the previous layer.

differ from *recurrent* neural networks (RNNs) in that the latter have cyclic connections which allow them to model sequential data of arbitrary length. RNN-based language models have been successfully applied to speech and phoneme recognition (Mikolov et al., 2010 and Hochreiter and Schmidhuber, 1997; Graves et al., 2013, respectively), and to language generation conditioned on different inputs, e.g., in an end-to-end approach for machine translation (Sutskever et al., 2014), for Chinese poetry generation (Zhang and Lapata, 2014), and image description generation (see Section 5.1.2 for details).

In the area of computer vision, deep convolutional neural networks (CNNs) have proven highly effective for object classification and recognition. They have led to a major breakthrough on the ImageNet large-scale visual recognition challenge (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Russakovsky et al., 2014), they have been used by the top systems of the challenge on facial expression recognition (ICML 2013 Workshop on Representation Learning<sup>2</sup>, Goodfellow et al., 2013), and have achieved performance superior to the state-of-the-art on handwriting recognition (Cireşan et al., 2012).

CNNs and the feature representations they induce are widely used in computer vision systems and in approaches leveraging image data including multimodal models (Section 5.1.2, e.g., Frome et al., 2013; Mao et al., 2014; Kiros et al., 2014a,b; Kiela and Bottou, 2014). Recently, they have been also employed for large-vocabulary continuous speech recognition (Sainath et al., 2013). Models based on (deep) autoencoders haven been successful on knowledge-free handwriting recognition (Rifai et al., 2011) and were shown to be able to learn selective features for face and human body detectors from unlabelled data (Le et al., 2013). A hybrid approach was proposed by Kahou et al. (2013) for emotion recognition from videos (Dhall et al., 2013). They leveraged different features which were extracted with modality-specific deep networks (e.g., deep CNNs (Krizhevsky et al., 2012) for facial expressions, deep belief nets (DBNs) for audio features, and autoencoders for local motion features).

The approaches mentioned above exemplify most of the beneficial properties of (deep) network architectures listed in the beginning of this chapter (e.g., unsupervised/hierarchical feature learning and applicability to different tasks or domains). In contrast to our work, the networks operate on a single modality, i.e. language (text or speech) or vision. In the following section we outline related work on exploiting deep networks for modality composition.

---

<sup>2</sup><http://deeplearning.net/icml2013-workshop-competition/>



### 5.1.2 Multimodal Deep Learning

The use of stacked autoencoders to extract a shared lexical meaning representation is new to our knowledge, although, as we explain below related to a large body of work on deep learning in network architectures.

Srivastava and Salakhutdinov (2012) and Ngiam et al. (2011) were the first to address the problem of multimodal representation learning in neural network architectures which project different modalities onto a bimodal space (Figure 2.5 (d)). These networks are typically composed of different levels: the lower levels pertain to modality-specific layers or (sub-)networks (which are possibly pre-trained with a different type of architecture) whereas higher modality-unifying layers operate on top of them. Work which focusses on integrating words and images has used a variety of architectures including deep Boltzmann machines (DBMs; Srivastava and Salakhutdinov, 2012, 2014), restricted Boltzmann machines (RBMs; Sohn et al., 2014), and autoencoders (Feng et al., 2013). Similar methods were employed to combine other modalities such as speech and video or images using RBMs and DBNs (Ngiam et al., 2011; Huang and Kingsbury, 2013; Kim et al., 2013), deep autoencoders (Ngiam et al., 2011), or DBMs (Srivastava and Salakhutdinov, 2014).

Although our model is conceptually similar to these studies (especially those applying stacked autoencoders), it differs in at least two aspects. Firstly, many former models learn bimodal representations with the aim to reason about one modality given the respective other modality (e.g., Ngiam et al., 2011; Huang and Kingsbury, 2013; Sohn et al., 2014). In contrast, our goal is to learn bimodal representations in which complimentary and redundant information from different modalities is unified in an optimal way. Secondly, most approaches deal with a particular end task (e.g., image classification or speech recognition, but see Srivastava and Salakhutdinov, 2014, for an exception). They join the modalities in a *task-independent* unified representation by means of an unsupervised criterion (e.g., with an RBM or autoencoder), and fine-tune the network parameters with an appropriate supervised criterion on top of the joint representations (e.g., Huang and Kingsbury, 2013), or use the latter as features for training a conventional classifier (e.g., Ngiam et al., 2011; Sohn et al., 2014). In contrast, we do not address a specific task and fine-tune our autoencoder using a semi-supervised criterion. Specifically, we use a combined objective comprising the reconstruction of the input representation and the classification of the input object. The latter, supervised criterion is used as a means to drive the learning process, as we will explain in more

detail in Section 5.3.

Furthermore, our model is defined at a finer level of granularity than most previous work—it computes representations for *individual* words—and leverages information from decoupled data sources, i.e. image collections and text corpora. Former work on multimodal representation learning builds upon images and their accompanied tags (Srivastava and Salakhutdinov, 2014; Sohn et al., 2014; Feng et al., 2013), or sentential descriptions of the image content for the purpose of image and description retrieval (Kiros et al., 2014b; Mao et al., 2014; Socher et al., 2014; Karpathy and Fei-Fei, 2015).

Previous work on image description and retrieval differs from our model in that it directly derives *task-specific* bimodal representations. Existing models often initialise their networks with unimodal task-independent feature representations pre-trained in deep networks (e.g., CNNs trained on ImageNet; Krizhevsky et al., 2012), but integrate the modalities using a training criterion and an architecture suited to a particular task. For example, Mao et al. (2014) and Karpathy and Fei-Fei (2015) combine the modalities in an RNN which is trained to predict the next word given an image and previous words for the purpose of generating image descriptions or retrieving images given sentences or vice versa. Retrieval is also addressed by Kiros et al. (2014b) who project image representations onto the hidden states of an RNN representing a sentence, and optimise the projection matrix and the RNN parameters using the pairwise ranking loss as cost function (Weston et al., 2010; see also the winning system from Feng et al. 2013 of the multimodal learning challenge of an ICML 2013 workshop<sup>3</sup>). In Socher et al.’s (2014) approach to the retrieval task, the same training objective was used to optimise the parameters which project image embeddings onto sentence embeddings, hence performing cross-modal learning<sup>4</sup> (cf. Figure 2.5 (f), Page 26).

Cross-modal learning was also performed on the word- (or object-)level for the purpose of zero-shot learning (see Chapter 4, Section 4.2.1), where images of unseen classes were mapped onto the textual space so as to classify them by applying a  $k$ -nearest neighbours approach within this space (Socher et al., 2013b; Figure 2.5 (e)). Similarly, Frome et al. (2013) learn image-to-word embeddings and use a nearest neighbour approach to perform large-scale object recognition and zero-shot learning (in their deep network, the pre-trained image subnetwork was fine-tuned during the

---

<sup>3</sup>In this challenge, systems were presented with an image and two candidate sets of word tags and had to choose the correct set (Goodfellow et al., 2013).

<sup>4</sup>Socher et al.’s (2014) model formulation allows for the joint optimisation of the image and sentence embeddings, but they do not report results for this.

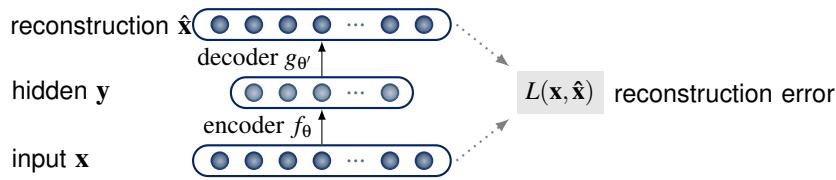


Figure 5.1: Basic autoencoder whose hidden layer  $\mathbf{y}$  yields a latent representation of some input  $\mathbf{x}$ .

cross-modal learning phase). Similar to our approach, the two latter cross-modal learning methods are instances of multimodal learning in deep networks using disjoint data sources, but unlike our work, their goal is not to learn joint bimodal meaning representations, but to tackle an image-based task by exploiting the textual modality.

## 5.2 Autoencoders

In this section, we review autoencoders with emphasis on aspects relevant to our model, which uses this type of neural network architecture to learn higher-level lexical meaning representations, as we describe in the subsequent section.

### 5.2.1 Basic Autoencoders

An autoencoder (a.k.a. auto-associator or diabolo network) is an unsupervised feed-forward neural network which is trained to reconstruct a given input from its latent distributed representation (Rumelhart et al., 1986b; Bengio, 2009). The architecture of the basic autoencoder is shown in Figure 5.1. It consists of an encoder  $f_{\theta}$  which maps an input vector  $\mathbf{x}^{(i)}$  to a hidden (*latent*) representation  $\mathbf{y}^{(i)} = f_{\theta}(\mathbf{x}^{(i)}) = s(\mathbf{W}\mathbf{x}^{(i)} + \mathbf{b})$ , with  $s$  being a non-linear activation function, such as a sigmoid function, and  $\mathbf{W}$  and  $\mathbf{b}$  being the weight matrix and an offset vector, respectively. A decoder  $g_{\theta'}$  then aims to reconstruct input  $\mathbf{x}^{(i)}$  from  $\mathbf{y}^{(i)}$ , i.e.  $\hat{\mathbf{x}}^{(i)} = g_{\theta'}(\mathbf{y}^{(i)}) = s(\mathbf{W}'\mathbf{y}^{(i)} + \mathbf{b}')$ . The training objective is the determination of parameters  $\hat{\theta} = \{\mathbf{W}, \mathbf{b}\}$  and  $\hat{\theta}' = \{\mathbf{W}', \mathbf{b}'\}$  that minimise the average reconstruction error over a set of input vectors  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ :

$$\hat{\theta}, \hat{\theta}' = \arg \min_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^n L(\mathbf{x}^{(i)}, g_{\theta'}(f_{\theta}(\mathbf{x}^{(i)}))), \quad (5.1)$$

where  $L$  is a loss function, such as cross-entropy. Parameters  $\theta$  and  $\theta'$  can be optimised by gradient descent methods.

Autoencoders are a means to learn representations of some input by retaining useful features in the encoding phase which help to reconstruct (an approximation of) the input, whilst discarding useless or noisy ones.

Beyond their application in natural language processing and computer vision (Section 5.1), this type of architecture has been previously shown to account for phenomena in human category learning (Kurtz, 2007), in particular in infant category learning (e.g., Westermann and Mareschal, 2014; French et al., 2004; Mareschal et al., 2000). Mareschal et al. (2000) (inter alia) draw an analogy between autoencoders and infant concept learning by suggesting to interpret the latter as an iterative process, where an infant, during the period of stimulus fixation, encodes the stimulus into an internal representation and assesses this encoding by using it to predict the properties of the actually perceived stimulus. The infant iteratively updates and re-assesses this encoding until predicted and actual properties are congruent. In this sense, the network reconstruction error is related to looking-time of an infant, that is, the less familiar an object, the higher the error and looking-time of the infant, respectively.

Common to the instances of autoencoders employed in the above-mentioned work is the use of a bottleneck hidden layer producing an *under-complete* representation of the input by having a smaller number of units than the input layer. In this setting autoencoders are similar to techniques such as principal component analysis (PCA, Jolliffe, 2002) since they both can be leveraged to reduce the dimensionality of often high-dimensional input data. Unlike PCA which performs a linear transformation of the input, autoencoders (with non-linearities and a cross-entropy loss function) are able to model non-linearities and are thus potentially more powerful for capturing complex structure of the input. Furthermore, autoencoders may benefit from using multiple hidden layers (Japkowicz et al., 2000; Hinton and Salakhutdinov, 2006; Vincent et al., 2010).

The use of a bottleneck hidden layer to produce under-complete representations of the input is one strategy of guiding parameter learning towards useful representations. The literature describes further strategies, such as constraining the hidden layer to yield sparse representations (Ranzato et al., 2006), or *denoising*.

## 5.2.2 Denoising Autoencoders

The training criterion with denoising autoencoders is the reconstruction of clean input  $\mathbf{x}^{(i)}$  given a corrupted version  $\tilde{\mathbf{x}}^{(i)}$  (Vincent et al., 2008, 2010). The reconstruction error

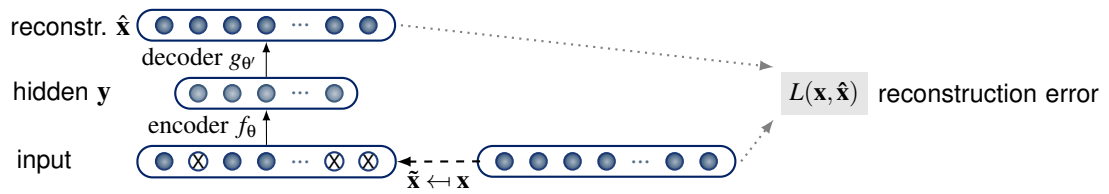


Figure 5.2: Denoising autoencoder whose hidden layer  $\mathbf{y}$  yields a latent representation given a corrupted version  $\tilde{\mathbf{x}}$  of some input  $\mathbf{x}$ .

for an input  $\mathbf{x}^{(i)}$  with loss function  $L$  then is:

$$err(\mathbf{x}^{(i)}, \tilde{\mathbf{x}}^{(i)}) = L(\mathbf{x}^{(i)}, g_\theta(f_\theta(\tilde{\mathbf{x}}^{(i)}))) \quad (5.2)$$

One possible corruption process is *masking noise*, where the corrupted version  $\tilde{\mathbf{x}}^{(i)}$  results from randomly setting a fixed proportion  $\nu$  of units of  $\mathbf{x}^{(i)}$  to 0. Figure 5.2 gives an illustration.

The underlying idea of denoising autoencoders is that if a latent representation is capable of reconstructing the actual input from its corruption, it presumably has learned to capture the regularities and interrelations of the structure of the input and can therefore be deemed a good representation. From a cognitive perspective, denoising can be construed as learning to activate knowledge about interrelated factors when being exposed to partial information about a concept. An example for this is the ability of humans to recognise objects that are partially occluded or which are depicted in corrupted images (Vincent et al., 2008).

### 5.2.3 Stacked Autoencoders

Several (denoising) autoencoders can be used as building blocks to form a deep neural network (Bengio et al., 2006; Vincent et al., 2010). For that purpose, the autoencoders are often pre-trained layer by layer, with the current layer being fed the latent representation yielded by the previous, already pre-trained, autoencoder as input. When stacking *denoising* autoencoders, input corruption is only applied during pre-training. Encodings propagated to the next autoencoder are obtained from uncorrupted input. Using this unsupervised pre-training procedure, initial parameters are found which approximate a good solution. Subsequently, the original input layer and hidden representations of all the autoencoders are stacked yielding a deep network.

The parameters of this network can then be optimised (*fine-tuned*) with respect to the objectives at hand. More precisely, a supervised criterion can be imposed on top of

Visual		eat.seeds	has.beak	has.claws	has.handlebar	has.wheels	has.wings	is.yellow	of.wood		
	<i>canary</i>	0.05	0.24	0.15	0.00	-0.10	0.19	0.34	0.00		
	<i>trolley</i>	0.00	0.00	0.00	0.30	0.32	0.00	0.00	0.25		
Textual		bird:n	breed:v	cage:n	chirp:v	fly:v	track:n	ride:v	run:v	rail:n	wheel:n
	<i>canary</i>	0.16	0.19	0.39	0.13	0.13	0.00	0.00	0.00	0.00	-0.05
	<i>trolley</i>	-0.40	0.00	0.00	0.00	0.00	0.14	0.16	0.33	0.17	0.20

Table 5.1: Examples of attribute-based representations provided as input to our autoencoders. (Some attributes are abbreviated for space reasons.)

the last hidden layer such as the minimisation of a prediction error on a supervised task (Bengio, 2009). Another approach is to unfold the stacked autoencoders and fine-tune their parameters with respect to the minimisation of the global reconstruction error (Hinton and Salakhutdinov, 2006). Alternatively, a semi-supervised criterion can be used (Ranzato and Szummer, 2008; Socher et al., 2011) through combination of the unsupervised training criterion (global reconstruction) with a supervised criterion, that is, the prediction of some target given the latent representation.

## 5.3 Grounded Semantic Representations with Autoencoders

To learn meaning representations of single words from textual and visual input, we employ stacked (denoising) autoencoders. Both input modalities are vector-based representations of words or, more precisely, the objects they refer to (e.g., *canary*, *trolley*). The vector dimensions correspond to textual and visual attributes, examples of which are shown in Table 5.1. We obtain these input vectors automatically with the methods described in Chapter 4 (Sections 4.2 and 4.3).

### 5.3.1 Architecture

We determined the architecture including the number of hidden layers and the activation function, as well as the training procedure in preliminary experiments, in which we, additionally to optimising for the training objective, took into account the performance of the model on a subset of the free word association norms collected by Nelson et al. (1998)<sup>5</sup> (see Section 3.4 for details on this dataset and task). Specifically,

<sup>5</sup>Available at <http://w3.usf.edu/FreeAssociation> (last accessed in April 2015).

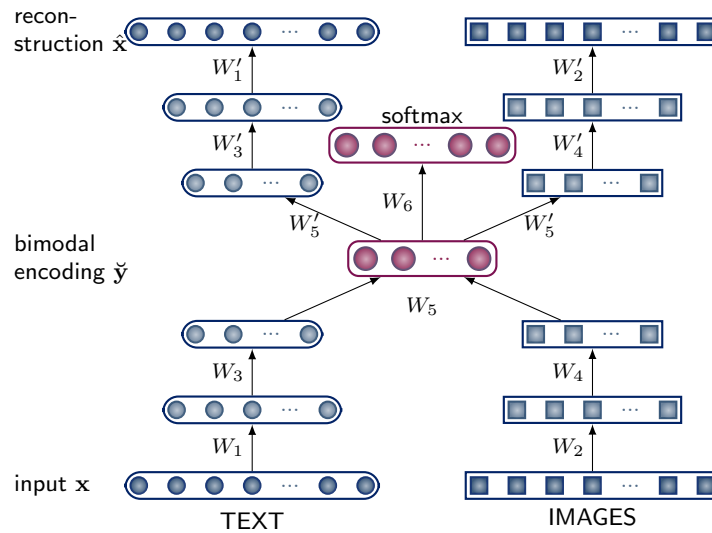


Figure 5.3: Stacked autoencoder trained with semi-supervised objective. Input to the model are single-word vector representations obtained from text and images. Vector dimensions correspond to textual and visual attributes, respectively (see Table 5.1). The edges are labelled with the weight matrices to be learned (bias vectors are omitted for the sake of clarity).

we used correlation analysis to monitor the correlation between model cue-associate cosine similarities (see Chapter 2, Section 2.2.1) and human probabilities. We describe the architecture of the best model in the following.

We first pre-train a stack of two autoencoders (AEs) for each modality separately. Then, we join the modalities by feeding the latent representations (*encodings*) induced by their respective second AE simultaneously to another AE. Its hidden layer  $\check{y}$  yields word representations that capture the meaning of words across both modalities. In the final training phase, we stack all layers and unfold them in order to fine-tune this bimodal stacked autoencoder (SAE). Figure 5.3 illustrates the architecture of the model. As can be seen from the figure, we additionally add a softmax-layer on top of the bimodal encoding layer (shown in the center of Figure 5.3, labelled as softmax), which outputs predictions with respect to the object label of an input (e.g., *dog*, *baseball*). It serves as a supervised training criterion in addition to the unsupervised reconstruction objective during fine-tuning, with the aim of guiding the learning towards descriptive and discriminative (bimodal) representations that capture the structure of the input patterns within and across the two modalities and discriminate between different objects.

The influence of object labels on categorisation and word learning has been the

subject of many studies which observed that labels can affect prelinguistic category formation during infancy (Plunkett et al., 2008; Sloutsky and Fisher, 2012, *inter alia*). A reason for this might be the formation of similar representations for objects that share the same label, which in turn causes infants to perceive dissimilar objects with an identical label as more similar (Westermann and Mareschal, 2014).

Adopting a progressive representation learning strategy and modality integration in the form of pre-training each layer separately is expedient for technical reasons, as outlined in the previous section (Section 5.2.3). It can also be interpreted as allowing the model to first learn meaning representations by detecting correlations between attributes within a modality, and later on across modalities, gradually enriching the representations. This is not unreasonable in view of studies suggesting that the ability of infants to integrate information from multiple sources during category learning develops with increasing age (see Westermann and Mareschal, 2014, for a review).

After training, a word is represented by its encoding in the bimodal layer, corresponding to a vector  $\check{y}$  of distributed unit activations (shown in the center of Figure 5.3). Recall that an individual unit of  $\check{y}$  does not represent a nameable attribute, but it is rather part of a pattern formed by the interplay between the visual and linguistic characteristics of the word it represents (cf. Chapter 2, Section 2.3). Two words can then be compared on the basis of their encoding vectors (e.g., by measuring their cosine similarity, Equation 2.4 in Chapter 2), and the more their activation patterns coincide, the more similar the words are assumed to be.

### 5.3.2 Model Details

For both modalities, we use the hyperbolic tangent function as activation function for encoder  $f_{\theta}$  and decoder  $g_{\theta'}$  and an entropic loss function for  $L$ . The weights of each autoencoder (AE) are tied, i.e.  $\mathbf{W}' = \mathbf{W}^T$ . We employ denoising AEs for pre-training the textual modality.

Regarding the visual autoencoder, we derive a new (‘denoised’) target vector to be reconstructed for each input vector  $\mathbf{x}^{(i)}$ , and treat  $\mathbf{x}^{(i)}$  itself as corrupted input. The target vector is derived as follows: each object  $o$  (or concept) in our data is represented by multiple images. Each image in turn is rendered in a visual attribute vector  $\mathbf{x}^{(i)}$ . The target vector is the weighted aggregation of  $\mathbf{x}^{(i)}$  and the centroid  $\mathbf{x}^{(o)}$  of all attribute vectors collectively representing object  $o$ . This denoising procedure compensates for prediction errors made by the attribute classifiers on individual images. Moreover, not



all attributes which are true for a concept are necessarily observable from a relevant image. Attribute predictions for individual images therefore introduce corruption with respect to the overall *concept* they represent.

The bimodal autoencoder is fed with the concatenated second hidden encodings of the visual and textual modalities as input and maps these to a joint hidden layer  $\check{\mathbf{y}}$  of  $B$  units. We normalise both unimodal input encodings to unit length. Again, we use tied weights for the bimodal autoencoder. We also actively encourage the autoencoder to detect dependencies between the two modalities while learning the mapping to the bimodal hidden layer, and therefore apply masking noise to one modality with a masking factor  $\nu$  (see Section 5.2.2 on denoising autoencoders), so that the corrupted modality optimally has to rely on the other modality in order to reconstruct its missing input features. Motivation for this is to simulate the information available to an infant during word learning, where the child typically can see the object and how it is referred to, but it may not have gained (rich) conceptual knowledge about it yet.

In the final step, we build a bimodal stacked autoencoder (SAE) with all pre-trained autoencoders and fine-tune their parameters with respect to a semi-supervised criterion. That is, we unfold the stacked autoencoder (as shown in Figure 5.3) and furthermore add a softmax output layer on top of the bimodal layer  $\check{\mathbf{y}}$  that outputs predictions  $\hat{\mathbf{t}}$  with respect to the inputs' object labels (e.g., *boat*):

$$\hat{\mathbf{t}}^{(i)} = \frac{\exp(\mathbf{W}^{(6)}\check{\mathbf{y}}^{(i)} + \mathbf{b}^{(6)})}{\sum_{k=1}^O \exp(\mathbf{W}_k^{(6)}\check{\mathbf{y}}^{(i)} + \mathbf{b}_k^{(6)})}, \quad (5.3)$$

with weights  $\mathbf{W}^{(6)} \in \mathbb{R}^{O \times B}$ ,  $\mathbf{b}^{(6)} \in \mathbb{R}^{O \times 1}$ , where  $O$  is the number of unique object labels. The overall objective to be minimised is then the weighted sum of the reconstruction error  $L_r$  and the classification error  $L_c$ :

$$L = \frac{1}{n} \sum_{i=1}^n \left( \delta_r L_r(\mathbf{x}^{(i)}, \hat{\mathbf{x}}^{(i)}) + \delta_c L_c(\mathbf{t}^{(i)}, \hat{\mathbf{t}}^{(i)}) \right) + \lambda R \quad (5.4)$$

where  $\delta_r$  and  $\delta_c$  are weighting parameters that give different importance to the partial objectives,  $L_c$  and  $L_r$  are entropic loss functions, and  $R$  is a regularisation term with  $R = \sum_{j=1}^5 2\|\mathbf{W}^{(j)}\|^2 + \|\mathbf{W}^{(6)}\|^2$ , i.e. we use an L2 weight decay penalty (penalisation of the sum of squared weights). Finally,  $\hat{\mathbf{t}}^{(i)}$  is the object label vector predicted by the softmax function for input vector  $\mathbf{x}^{(i)}$ , and  $\mathbf{t}^{(i)}$  is the correct object label, represented as an  $O$ -dimensional *one-hot vector*<sup>6</sup>.

<sup>6</sup>In a one-hot vector (a.k.a. *1-of- $N$  coding*), exactly one element is one and the others are zero. In our case, the non-zero element corresponds to the object label.

### 5.3.3 Model Properties

Our model benefits from its deep learning architecture, obtaining meaning representations from multiple layers. The first layers operate on individual modalities, whereas the final hidden layer combines them to a bimodal representation. This architecture allows us to test different hypotheses with respect to word meaning. Specifically, we can disentangle the contribution of visual or textual information, for instance by representing words based on their unimodal encoding and contrasting them with their bimodal representation. Related bimodal models have used SVD (Bruni et al., 2014, Chapter 6, Section 6.1.2), LDA (Roller and Schulte im Walde, 2013), or kCCA (Silberer et al., 2013, Chapter 4, Section 4.4) to project the input data into a joint space *directly* (see also Chapter 2, Section 2.5 for the former two models). There is no hierarchy of representations with potentially increasing complexity, nor an intermediate unimodal representation naturally connecting the input to the bimodal representation. Similarly to models employing SVD or kCCA, our model can also perform dimensionality-reduction in the course of representation learning by mapping to lower-dimensional hidden layers. However, in contrast to SVD, this is performed non-linearly which we argue allows to model complex relationships between visual and textual data.

Furthermore, the semi-supervised architecture affords flexibility, allowing to adapt the model to specific tasks. For example, by setting the corruption parameter  $\nu$  for the textual modality to one and the weighting parameter for the reconstruction error,  $\delta_r$ , to zero, a standard object classification model for images can be trained. More importantly, it has the potential for inductive inference with respect to attributes of new objects (cf. Johns and Jones, 2012, Chapter 3, Section 3.3.2). Particularly training a model that infers a (missing) modality given the other modality can be done by setting the corruption parameter  $\nu$  close to one for either modality (cf., cross-modal learning, Section 2.5.2, and Ngiam et al., 2011). As our input consists of natural language attributes, the model would infer textual attributes given visual attributes and vice versa. For example, when presented with an unknown object (e.g., an unseen bird species) and assuming that the object is visually similar to other objects of the same category (e.g., other birds), the visual input representation will be similar to the seen examples of the same category and will be mapped closely to these examples in the bimodal semantic space. Reconstructing the input from the bimodal representation, that is, mapping the bimodal representation to the output layers, textual attributes not perceived as input can be inferred. We show this in an experiment in Chapter 7 (Section 7.1; see also

Griffin et al., 2013). This inference ability follows directly out of the model, without additional assumptions or modifications. Previous models either do not have a simple way of projecting one modality onto a joint space (e.g., Andrews et al., 2009), or altogether lack a mechanism of inferring missing modalities (see Silberer and Lapata, 2012, and Chapter 3).

## 5.4 Conclusions

We presented our bimodal stacked autoencoder (SAE) model for deriving visually grounded meaning representations by mapping textual and visual information into a bimodal space. Its architecture was designed with respect to our desiderata for models of perceptually grounded meaning representations (see Chapter 3, Section 3.5). In summary, the model has the following characteristics:

- (1) It learns to map the input modalities to a modality-unifying hidden layer in a joint fashion (see Figure 2.5 (d), Page 26). In the next chapter, we will assess the effectiveness of the derived bimodal representations and compare it to related models.
- (2) In contrast to other network models which learn word *embeddings* from randomly initialised input, our input vectors are meaningful (they are attribute-based). The model can therefore derive bimodal representations for out-of-vocabulary words, provided that there is attribute-based information for these.
- (3) It offers the possibility to map just one modality into the bimodal space. In Chapter 7, we will apply the SAE to infer textual attributes when only presented with visual input from out-of-vocabulary words (Section 7.1). Similarly, we will apply the bimodal representations obtained from visual input for a categorisation task (Section 7.2).

Finally, due to its semi-supervised architecture, it has the potential to be explicitly trained for cross-modal mapping or supervised tasks, such as image classification (Section 5.3.3).

# Chapter 6

## Experiments: Simulating Human Behaviour in Cognitive Tasks

In this chapter we evaluate our bimodal stacked autoencoder (SAE) model presented in the previous chapter for its ability to explain human behaviour. To this end, we evaluate the model against human judgements in three semantic tasks related to concept similarity. All tasks focus on essential phenomena of cognition for which any lexical semantic model should account.

Many cognitive tasks, such as semantic priming (Thompson-Schill et al., 1998; Jones et al., 2006) or association, are based on the ability to judge similarity. Moreover, estimating similarity of pairs or groups of words is crucial for many practical applications (e.g., document retrieval, Manning et al., 2008). Vector-based models aimed at representing the meaning of individual words are therefore commonly evaluated by their ability to measure the strength of semantic similarity between lexical units. Likewise, in our first experiment described in Section 6.1, we evaluate the capacity of our model to simulate human behaviour on judging similarity in meaning and, furthermore, appearance between pairs of nominal concepts.

Semantic categorisation, the mental grouping of objects and events into meaningful classes, is a classic topic in the field of cognitive science, central to perception, learning, and the use of language. Categories enable us to structure the world and inductively predict or infer properties of newly encountered objects by generalising pre-established knowledge from similar objects, i.e. that share the same category. We will assess how well our model can simulate this aspect of cognition on a concept categorisation task in Section 6.2.

The internal structure of categories is graded in that some members are rated more

representative for a specific category than others. For example, both, *pythons* and *cats* are exemplars of PETS. However, humans generally consider a *python* to be a less typical exemplar for a PET than a *cat*. In our third experiment, presented in Section 6.3, we simulate typicality ratings to evaluate how well our model can account for such graded category membership.

## 6.1 Experiment 5: Word Similarity

We first give details on the evaluation dataset we used for the word similarity task. Then, in Section 6.1.2, we explain how the model was trained and describe the approaches used for comparison with our own work.

### 6.1.1 Elicitation of Evaluation Dataset

In this experiment, we collected similarity ratings that capture the concepts contained in the attribute production norms of McRae et al. (2005, henceforth McRae norms). Although several relevant datasets exist, such as the widely used WordSim353 (Finkelstein et al., 2002, see Section 3.4.1) or the more recent Rel-122 norms (Szumlanski et al., 2013), they contain many abstract words, (e.g., *love–sex* or *arrest–detention*) which are not covered in the McRae norms. This is for a good reason, as most abstract words do not have discernible attributes, or at least attributes that participants would agree upon. The new dataset we created consists exclusively of nouns from the McRae norms, and contains similarity ratings not only for semantic similarity, but also for visual similarity. We hope that the dataset will be useful for the development and evaluation of grounded semantic space models.<sup>1</sup>

**Participants** We used Amazon Mechanical Turk (AMT) to obtain similarity ratings for the word pairs grouped into tasks. Each task was completed by five volunteers, all self-reported native English speakers. They were allowed to complete as many tasks as they wanted. A total of 46 subjects (27 women, 18 men, 1 unspecified, mean age: 38.5 years, age range: 18–67) took part in the experiment and completed between one and 147 tasks each. Participants were paid \$0.5 per completed task.

---

<sup>1</sup>Available at <http://homepages.inf.ed.ac.uk/s1151656/resources.html>.

Word Pairs	Semantic	Visual	Word Pairs	Semantic	Visual
<i>bag–sack</i>	5.0	5.0	<i>bat_(baseball)–baton</i>	2.8	4.0
<i>pistol–revolver</i>	5.0	5.0	<i>bracelet–chain</i>	2.8	4.0
<i>couch–sofa</i>	5.0	5.0	<i>pencil–wand</i>	1.8	4.0
<i>airplane–jet</i>	5.0	5.0	<i>bullet–thimble</i>	1.0	3.0
<i>frog–toad</i>	5.0	5.0	<i>closet–elevator</i>	1.5	2.8
<i>hornet–wasp</i>	4.8	4.8	<i>banner–scarf</i>	1.3	2.7
<i>curtains–drapes</i>	5.0	4.8	<i>shield–tray</i>	1.2	2.6
<i>colander–strainer</i>	5.0	4.8	<i>cantaloupe–plum</i>	4.5	1.8
<i>gloves–mittens</i>	5.0	4.2	<i>clarinet–keyboard_(musical)</i>	4.3	1.3
<i>cup–mug</i>	5.0	4.3	<i>car–scooter</i>	4.0	1.7
<i>blouse–shirt</i>	4.8	5.0	<i>gun–missile</i>	4.0	1.0
<i>missile–rocket</i>	4.8	5.0	<i>screwdriver–wrench</i>	3.6	1.4
<i>tortoise–turtle</i>	4.8	5.0	<i>microwave–skillet</i>	3.5	1.3
<i>gun–shotgun</i>	4.8	5.0	<i>airplane–truck</i>	3.4	1.2

Table 6.1: Mean semantic and visual similarity ratings for the concepts of the McRae norms with varying degrees of similarity. Averaged across experiment participants.

**Materials and Design** Initially, we created all possible pairings over the concepts of the McRae norms and computed the semantic relatedness of the corresponding WordNet (Fellbaum, 1998) synsets using Patwardhan and Pedersen’s (2006) WordNet-based measure. We opted for this specific measure as it achieves high correlation with human ratings and has a high coverage on our nouns. Next, we randomly selected 30 pairs for each concept under the assumption that they are representative of the full variation of semantic similarity. More specifically, for each individual concept, we ordered the pairs according to their relatedness scores and assigned them into bins. We then iteratively sampled pairs from the bins of all concepts in such a way that each bin contributed as equally as possible to the final set of overall concept pairs. This resulted in 7,576 pairs.<sup>2</sup> We split the pairs into overall 255 tasks; each task consisted of 32 pairs covering examples of weak to very strong semantic relatedness, and furthermore contained at most one instance of each target concept. Two control pairs from Miller and Charles (1991) were included in each task to potentially help identify and eliminate data from participants who assigned random scores.<sup>3</sup>

<sup>2</sup>For two concepts, *yacht* and *wall*, only 25 pairs were created.

<sup>3</sup>To obtain exactly 32 pairs per task, we used three filler pairs if necessary, which we removed again after the annotation.

**Procedure** Participants were first presented instructions that explained the task and gave examples. They were asked to rate a pair on two dimensions, visual and semantic similarity using a 5-point Likert scale (1 = *highly dissimilar* and 5 = *highly similar*). All 32 word pairs comprising one task were presented on a single web page and participants were required to scroll down in order to process through the list of pairs. Note that they were not provided with images depicting the concepts. In order to judge visual (and semantic) similarity they were required to use their visual knowledge of concepts. Our instructions are given in Appendix B.

**Results** Examples of the stimuli and elicited mean ratings are shown in Table 6.1. The similarity data was post-processed so as to identify and remove outliers. Similarly to previous work (Szumlanski et al., 2013), we considered an outlier to be any individual whose mean pairwise correlation coefficient (Spearman’s  $\rho$ ) fell outside two standard deviations from the mean correlation. 11.5% of the annotations were detected as outliers and removed. After outlier removal, we further examined how well the participants agreed in their similarity judgements. We measured inter-subject agreement as the average pairwise correlation coefficient between the ratings of all annotators for each task. For semantic similarity, the mean correlation was  $\rho = 0.76$  (Min = 0.34, Max = 0.97, StD = 0.11) and for visual similarity  $\rho = 0.63$  (Min = 0.19, Max = 0.90, StD = 0.14). These results indicate that the participants found the task relatively straightforward and produced similarity ratings with a reasonable level of consistency. For comparison, Patwardhan and Pedersen’s (2006) measure achieved a coefficient of  $\rho = 0.56$  on the dataset for semantic similarity and  $\rho = 0.48$  for visual similarity. The correlation between the average ratings of the AMT annotators and the Miller and Charles (1991) dataset was  $\rho = 0.91$ .

### 6.1.2 Experimental Setup

**Data** We learned meaning representations for the concepts of the McRae norms which are contained in the VISA dataset (see Section 4.2.3). As shown in Figure 5.3 (Chapter 5, Page 91), our bimodal stacked autoencoder (SAE) model takes as input two (real-valued) vectors representing the visual and textual modalities. Vector dimensions correspond to textual and visual attributes, respectively. We maintained the partition of the VISA image data into training, validation, and test set and acquired visual vectors for each of the sets by means of our attribute classifiers (see Chapter 4, Section 4.2.4).

We used the visual vectors of the training and development set for training the autoencoders, and the vectors for the test set for evaluation. Visual vectors were scaled to the  $[-1, 1]$  range. We derived textual attribute vectors by means of Strudel (Baroni et al., 2010) as explained in Chapter 4 (Section 4.3.1). Specifically, we ran Strudel on a 2009 dump of the English Wikipedia of about 800M words.<sup>4</sup> We only retained the ten attributes with highest log-likelihood ratio scores for each target word, amounting to a total of 2,362 dimensions for the textual vectors. Analogously to the visual representations, association scores were scaled to the  $[-1, 1]$  range.

**Model Parameters** Model hyper-parameters<sup>5</sup> were optimised on the same subset of the free word association norms collected by Nelson et al. (1998)<sup>6</sup> as we used in the experiments which assessed the effectiveness of our attribute-based representations (i.e. 435 word pairs), described in Chapter 4 (Section 4.4).

During training we used correlation analysis ( $\rho$ ) to monitor the degree of linear relationship between model cue-associate cosine similarities (see Chapter 2, Section 2.2.1) and human probabilities. The best autoencoder on the word association task obtained a correlation coefficient of  $\rho = 0.33$ . This model has the following architecture: the textual denoising autoencoder (see Figure 5.3, Page 91, left-hand side) consists of 700 hidden units which are then mapped to the second hidden layer with 500 units (the corruption parameter was set to  $v = 0.1$ ; see Chapter 5, Section 5.2.2); the visual autoencoder (see Figure 5.3, right-hand side) has 170 and 100 hidden units, in the first and second layer, respectively. The 500 textual and 100 visual hidden units feed a bimodal autoencoder containing 500 units, and masking noise was applied to the textual modality with  $v = 0.2$ . The weighting parameters for the joint training objective of the stacked autoencoder were set to  $\delta_r = 0.8$  and  $\delta_c = 1$  (see Equation (5.4), Page 93).

We used the meaning representations obtained from the output of the bimodal layer for the experiment.

**Comparison Models** We compare our SAE against unimodal autoencoders based solely on textual and visual input (left- and right-hand sides in Figure 5.3, Page 91,

---

<sup>4</sup>The corpus is downloadable from <http://wacky.sslmit.unibo.it/doku.php?id=corpora> (last accessed in April 2015).

<sup>5</sup>We performed random search over combinations of hyper-parameter values, including also the number of unimodal hidden layers and the number of units in each layer.

<sup>6</sup>Available at <http://w3.usf.edu/FreeAssociation> (last accessed in April 2015).



respectively). We also compare our model against a concatenation model as well as two latent inference approaches which differ in their modality integration mechanisms (both perform joint integration illustrated by Figure 2.5 (c) on Page 26, however). The first one is based on kernel canonical correlation analysis (kCCA, Hardoon et al., 2004) with a linear kernel (see Chapter 3, Section 3.3.3, for details on kCCA). The second one emulates Bruni et al.'s (2014) integration mechanism based on singular value decomposition (SVD, see below). All these models run on the same data and are given input identical to our model, namely attribute-based textual and visual representations.

We furthermore report results obtained with Bruni et al.'s (2014) full model as well as results of the state-of-the-art word embeddings learned with Mikolov et al.'s (2013b) continuous skip-gram model. We describe these models in the following.

**Bruni Model** Bruni et al. (2014) use SVD (Golub and Reinsch, 1970) to integrate co-occurrence-based textual representations of words with visual representations constructed on the basis of low-level image features. As we explained in Chapter 2, SVD is a mathematical technique used for reducing the dimensionality of vector spaces (Section 2.2.1). Specifically, Bruni et al. concatenate the (normalised) textual and visual vectors to create one matrix, where the rows correspond to words, and the columns correspond to either textual or visual features. They then perform SVD and retain the  $k$  largest singular values. Re-multiplication of the factorisation gives a bimodal semantic space with rank  $k$ . We apply this integration mechanism on our textual and visual attribute-based input vectors for direct comparison of the integration mechanisms.

In order to build Bruni et al.'s (2014) full model<sup>7</sup> we used their publicly available system (Bruni et al., 2013). Their textual modality is represented by a 30K-dimensional word-word co-occurrence matrix extracted from text corpora.<sup>8</sup> The entries of the matrix correspond to the weighted co-occurrence frequency of a target word (rows) and a context word (columns). Two words were considered co-occurring if one of them occurred in the window of two content words on each side of the other word. The visual modality is represented by bag-of-visual-words histograms built on the basis of clustered SIFT (Lowe, 2004) descriptors (see Chapter 4, Section 4.2.4.2 for details on bag-of-visual-words).

Bruni et al.'s (2014) and our model first of all differ with respect to the source

---

<sup>7</sup>The authors call the model *Feature Level fusion* since similarity scores are estimated from the bimodal representation arising from fusing the textual and visual representations in contrast to combining similarity scores estimated from each modality separately.

<sup>8</sup>We thank Elia Bruni for providing us with their data.

corpora from which visual and textual information is derived. For the textual representations, they use the larger ukWaC (2 billion tokens) in addition to WaCkypedia (800 million tokens),<sup>9</sup> we rely solely on the latter. In order to extract visual information, Bruni et al. exploit the ESP dataset (von Ahn and Dabbish, 2004). This dataset comprises 100K images randomly downloaded from the internet and tagged by humans (see Chapter 3, Section 3.2, for more details on ESP). The average number of images per tag is 70. The core difference between Bruni et al.’s model and ours lies in the visual input representation. Bruni et al. represent an image as a histogram over not nameable visual words corresponding to quantised SIFT descriptors, and a concept as the weighted sum of the histograms of all images tagged with the concept name. We also leverage information extracted from images. However, we represent an image and consequently a concept by means of interpretable attributes.

**Skip-gram Model** Mikolov et al.’s (2013a; 2013b) skip-gram model is a shallow neural network architecture that learns distributed vector representations for words (and phrases) from text corpora. The network architecture consists of an input layer encoding a word  $w_t$  with a one-hot vector, a continuous embedding layer to which the input is projected, and an output layer which is fed by the embedding layer and which encodes words  $w_{t-c}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+c}$ . The latter correspond to the context words within a window of  $c$  words surrounding word  $w_t$ . Training objective is to find word embeddings (representations) from which the context words can be predicted.

In contrast to our model, representations are directly learned from experience approximated by a huge amount of text data. In our evaluation, we used the 300-dimensional vectors trained on part of the Google News dataset which comprises 100B words.<sup>10</sup> They were trained using negative sampling, where the objective is the distinction between the target, i.e. a correct context word of word  $w_t$ , from randomly sampled negative examples using logistic regression. This objective is reported to be especially valuable for the learning of representations for frequent words. The model furthermore employed sub-sampling of frequent words, which is geared towards learning improved representations for less frequent words (Mikolov et al., 2013b).

---

<sup>9</sup>Both corpora are available at <http://wacky.sslmit.unibo.it/doku.php?id=corpora> (last accessed in May 2015).

<sup>10</sup>The vectors and the code are both available at <https://code.google.com/p/word2vec/> (last accessed in April 2015).

Models	Semantic Similarity				Visual Similarity			
	T	V	T+V	CI	T	V	T+V	CI
McRae	0.71	0.49	0.68	(-.018/+0.018)	0.58	0.52	0.61	(-.021/+0.020)
Attributes	0.63	0.62	0.71	(-.014/+0.013)	0.49	0.57	0.60	(-.018/+0.017)
SAE	0.67	0.61	0.72	(-.014/+0.014)	0.55	0.60	0.65	(-.016/+0.017)
SVD	—	—	0.70	(-.014/+0.015)	—	—	0.59	(-.018/+0.018)
kCCA	—	—	0.58	(-.020/+0.017)	—	—	0.56	(-.019/+0.019)
Bruni	—	—	0.50	(-.023/+0.027)	—	—	0.44	(-.025/+0.024)
skip-gram	0.73	—	—	(-.014/+0.013)	0.56	—	—	(-.020/+0.020)

Table 6.2: Correlation of model predictions against similarity ratings for the noun pairs of the McRae norms (using Spearman’s  $\rho$ ).

**Evaluation** We evaluate the models on the word similarity dataset gathered in the elicitation study described in Section 6.1.1. With each model, we measure the cosine similarity of the given word pairs and correlate these predictions with the mean human similarity ratings using Spearman’s  $\rho$ .

### 6.1.3 Results and Discussion

Table 6.2 presents our results on the word similarity task. As an indicator to how well automatically extracted attributes can approach the effectiveness of clean human generated attributes, we also report results of a distributional model induced from the McRae norms (see the row labelled McRae in the table). Each noun is represented as a vector with dimensions corresponding to attributes elicited by participants of the norming study. Vector components are set to the (normalised) frequency with which participants generated the corresponding attribute. We show results for three models, using all attributes except those classified as visual (columns labelled T), only visual attributes (V), and all available attributes (T+V). The concatenation model (see row Attributes in Table 6.2) is based on the concatenation (T+V) of textual attributes (which we obtain from Strudel) and visual attributes (obtained from our classifiers; columns T and V, respectively). The automatically obtained textual and visual attribute vectors serve as input to SVD, kCCA, and our bimodal stacked autoencoder (SAE). The third row in the table presents three variants of our model trained on textual and visual attributes only (T and V, respectively) and on both modalities jointly (T+V).

#	Pair	#	Pair	#	Pair
1	<i>pliers–tongs</i>	8	<i>pistol–rifle</i>	15	<i>cedar–oak</i>
2	<i>cathedral–church</i>	9	<i>cloak–robe</i>	16	<i>bull–ox</i>
3	<i>cathedral–chapel</i>	10	<i>nylons–trousers</i>	17	<i>dress–gown</i>
4	<i>pistol–revolver</i>	11	<i>cello–violin</i>	18	<i>bolts–screws</i>
5	<i>chapel–church</i>	12	<i>cottage–house</i>	19	<i>salmon–trout</i>
6	<i>airplane–helicopter</i>	13	<i>horse–pony</i>	20	<i>oven–stove</i>
7	<i>dagger–sword</i>	14	<i>gun–rifle</i>	21	<i>iguana–tortoise</i>

Table 6.3: Word pairs with highest semantic and visual similarity according to SAE model. Pairs are ranked from highest to lowest similarity.

We report 95% confidence intervals based on 5000 bootstraps for the results obtained with the bimodal models (T+V) and skip-gram (columns labelled CI). Recall that participants were asked to provide ratings on two dimensions, namely semantic and visual similarity. We would expect the textual modality to be more dominant when modelling semantic similarity and conversely the perceptual modality to be stronger with respect to visual similarity. This is borne out in our unimodal SAEs. The textual SAE correlates better with semantic similarity judgements ( $\rho = 0.67$ ) than its visual equivalent ( $\rho = 0.61$ ). And the visual SAE correlates better with visual similarity judgements ( $\rho = 0.60$ ) compared to the textual SAE ( $\rho = 0.55$ ). Interestingly, the bimodal SAE (T+V) is better than the unimodal variants on both types of similarity judgements, semantic and visual. An explanation could be that both modalities contribute complementary information and that the SAE model is able to extract a shared representation which improves generalisation performance across tasks by learning them jointly. Since the bimodal SAE has one more layer than its unimodal variants, however, it cannot be ruled out that the improved effectiveness is by virtue of its higher complexity. The bimodal autoencoder (SAE, T+V) outperforms all other bimodal models on both similarity tasks. It yields a correlation coefficient of  $\rho = 0.72$  on semantic similarity and  $\rho = 0.65$  on visual similarity. Human agreement on the former task is 0.76 and 0.63 on the latter. Table 6.3 shows examples of word pairs with highest semantic and visual similarity according to the SAE model.

We also observe that simply concatenating textual and visual attributes (Attributes, T+V) performs competitively with SVD and better than kCCA. This indicates that

the attribute-based representation is a powerful predictor on its own. With respect to models that do not make use of attributes, we see that Bruni et al. (2014) is outperformed by all other attribute-based systems (see columns T and T+V in Table 6.2). Interestingly, skip-gram is the best performing model on the semantic similarity task (see column T, first block), but falls short on the visual similarity task.

## 6.2 Experiment 6: Concept Categorisation

Concept learning and categorisation have been subject to many experimental studies and simulation approaches (see, e.g., Goldstone et al., 2012, for an overview). Existing models typically focus on a single modality, either perception or language. For example, perceptual information is represented in form of hand-coded (binary) values on a few dimensions (e.g., colour or shape (Anderson, 1991; Vanpaemel et al., 2005), artificial stimuli (Griffiths et al., 2007a; Sanborn et al., 2006), geometric shapes (Austerweil and Griffiths, 2010; McKinley and Nosofsky, 1995)) or by real-object images (e.g., Hsu et al., 2012). And linguistic representations are often derived from large text corpora (e.g., Fountain and Lapata, 2011; Frermann and Lapata, 2014). Very few approaches exist that use both, perception and language (Bruni et al., 2014; Westermann and Mareschal, 2014). Furthermore, many models focus on adult categorisation, assuming categories have already been formed. In this experiment, we induce semantic categories following a clustering-based approach which uses the bimodal word representations learned by our model. The clustering approach we employ is not performed in an incremental fashion in contrast to related work (Fountain and Lapata, 2011; Frermann and Lapata, 2014).

### 6.2.1 Experimental Setup

**Data** We evaluate model output against a gold standard set of categories created by Fountain and Lapata (2010). The dataset contains a classification, produced by human participants, of the nouns from the McRae norms (McRae et al., 2005) into (possibly multiple) semantic categories (40 in total).<sup>11</sup> We transformed the dataset into hard categorisations by assigning each noun to its most typical category as extrapolated from human typicality ratings (for details see Fountain and Lapata, 2010). Furthermore, we excluded 82 nouns which we used for optimising the clustering, as described below.

---

<sup>11</sup>The dataset can be downloaded from <http://homepages.inf.ed.ac.uk/s0897549/data/> (last accessed in May 2015).

Category	Words
STICK-LIKE UTENSILS	<i>baton, ladle, peg, spatula, spoon</i>
RELIGIOUS BUILDINGS	<i>cathedral, chapel, church</i>
WIND INSTRUMENTS	<i>clarinet, flute, saxophone, trombone, trumpet, tuba</i>
AXES	<i>axe, hatchet, machete, tomahawk</i>
FURNITURE W/ LEGS	<i>bed, bench, chair, couch, desk, rocker, sofa, stool, table</i>
FURNITURE W/O LEGS	<i>bookcase, bureau, cabinet, closet, cupboard, dishwasher, dresser</i>
LIGHTINGS	<i>candle, chandelier, lamp, lantern</i>
ENTRY POINTS	<i>door, elevator, gate</i>
WRITING/BRISTLED DEVICES	<i>brush, comb, crayon, paintbrush, pen, pencil</i>
UNGULATES	<i>bison, buffalo, bull, calf, camel, cow, donkey, elephant, goat, horse, lamb, ox, pig, pony, sheep</i>
BIRDS	<i>crow, dove, eagle, falcon, hawk, ostrich, owl, penguin, pigeon, raven, stork, vulture, woodpecker</i>

Table 6.4: Examples of clusters produced by CW using the representations obtained from the SAE model.

**Method** To obtain a clustering of nouns, we used Chinese Whispers (CW, Biemann, 2006), a randomised, agglomerative graph-clustering algorithm. CW assumes that semantic information is organised like a network, where words are nodes and the weight of an (undirected) edge linking two nodes denotes the similarity of the respective words. In the categorisation setting, CW partitions the nodes of the weighted graph into disjunct groups in a bottom-up approach: At the beginning, each word (i.e. node) forms an own, basic-level category. All words are then iteratively processed for a few repetitions in which each word is assigned to the category (i.e. cluster) of the most similar neighbour words, as determined by the maximum sum of (edge) weights between the word and the neighbour nodes pertaining to the same category. CW is a non-parametric model, it induces the number of clusters as well as which words belong to these clusters from the data. In our experiments, we initialised CW with different graphs resulting from different vector-based representations of the McRae nouns. CW can optionally apply a minimum weight threshold which we optimised using the categorisation dataset from Baroni et al. (2010). The latter contains a classification of 82 McRae nouns into 10 categories.

Category	Words
REPTILES	<i>alligator, crocodile, frog, iguana, platypus, rattlesnake, salamander, toad, tortoise</i>
KNITWEAR	<i>carpet, gloves, mat, mittens, pillow, scarf, shawl, socks, sweater</i>
TREES or TREE-LIKE PLANTS	<i>birch, broccoli, cedar, oak, parsley, pine, vine, willow</i>
ANIMALS (W/ BEAK or BLACK&WHITE)	<i>crow, dolphin, dove, eagle, goose, pelican, penguin, pigeon, raven, seagull, skunk, swan, whale, woodpecker</i>

Table 6.5: Examples of clusters produced by CW using the representations obtained from the visual SAE model.

Category	Words
LIGHT-RELATED DEVICES	<i>candle, chandelier, lamp, lantern, microscope, mirror, projector</i>
SOUND-RELATED DEVICES	<i>radio, stereo, tape, telephone</i>
ROCKS/ROCK CONSTRUCTIONS (SEA)FOOD	<i>brick, fence, marble, pyramid, stone, wall catfish, cod, crab, eel, lobster, mushroom, octopus, salmon, shrimp, squid, trout, tuna</i>
SANITARY WARE	<i>bathtub, bin, faucet, pipe, sink, tank, toilet</i>

Table 6.6: Examples of clusters produced by CW using the representations obtained from the textual SAE model.

**Model Parameters** We use the SAE model described in Experiment 5 (Section 6.1). Some performance gains could be expected if (hyper-)parameter optimisation took place separately for each task. However, we wanted to avoid overfitting, and show that our parameters are robust across tasks and datasets. We furthermore compare to the same models as in Experiment 5 (see Section 6.1.2), and again estimate the similarity of two words with the cosine similarity.

**Evaluation** We evaluate the clustering solution  $\mathcal{S}$  produced by CW using the F-score measure introduced in the SemEval 2007 task (Agirre and Soroa, 2007); it is the harmonic mean of precision and recall. Precision is defined as the number of correct members of a cluster  $S \in \mathcal{S}$  divided by the number of items in the cluster. Recall is the number of correct cluster members divided by the number of items in the gold-standard category  $G \in \mathcal{G}$ . The F-score of the entire set of clusters  $\mathcal{S}$ , evaluated against

Models	T	V	T+V
McRae	0.52	0.31	0.42
Attributes	0.35	0.37	0.33
SAE	0.36	0.35	0.43
SVD	—	—	0.39
kCCA	—	—	0.37
Bruni	—	—	0.34
skip-gram	0.37	—	—

Table 6.7: F-score results on concept categorisation.

gold standard  $\mathcal{G}$ , is:

$$F_S = \sum_{G \in \mathcal{G}} \frac{|G|}{|S|} F(G), \quad (6.1)$$

where  $F(G)$  is the maximum F-score of category  $G$  obtained at any cluster.

## 6.2.2 Results and Discussion

Our results on the categorisation task are given in Table 6.7. In this task, simple concatenation of visual and textual attributes does not yield improved performance over the individual modalities (see row *Attributes* in Table 6.7). In contrast, all bimodal models are better (SVD and SAE) than or equal (kCCA) to their unimodal equivalents and skip-gram. The SAE outperforms both kCCA and SVD by a large margin delivering clustering performance similar to McRae’s human-produced norms. Table 6.4 shows examples of clusters produced by CW when using vector representations provided by the SAE model, and Tables 6.5 and 6.6 list clusters obtained from its visual and textual equivalent, respectively. Note that we added the cluster labels manually for illustration purposes.

## 6.3 Experiment 7: Typicality Ratings

An important finding in the study of natural language concepts is that categories show graded category-membership structure. For example, humans generally judge a *trout* to be a better example of the category FISH than *eel*. In the same way, an *apple* intuitively seems to be a better example of the category FRUIT than *olives*. Several experi-



mental studies underline the pervasiveness of typicality (or “goodness of example”) in a wide variety of cognitive tasks such as priming (Rosch, 1977), sentence verification (McCloskey and Clucksberg, 1979), and inductive reasoning (Rips, 1975). Because of its importance, typicality is also an evaluation criterion for models of categorisation and concept representation. Any such model should be able to give an account of the graded category structure and correctly predict differences in the typicality of category members. We therefore assess our SAE model on a typicality rating task (O’Connor et al., 2009) where the model is presented with instances of a category and must predict the degree to which the instances are typical amongst members of that category.

### 6.3.1 Experimental Setup

**Data** Our experiments use the dataset created by O’Connor et al. (2009). They collected typicality ratings by presenting participants with a category name and an instance of that category, and asking them to rate the goodness of the instance as an example for the category using a 9-point scale (where 9 means the member is a very good example, and 1 means it is a very poor example). Categories were presented in blocks. Typicality ratings for each category-instance pair were then averaged across 21 participants. The dataset contains typicality judgements for 33 categories, however we use the same 20 categories (611 category–instance items) which they used in their typicality experiments.

**Method** Assuming that all members of a category are known, the task is to determine each member’s typicality. We estimate the degree to which a member is representative of its category by measuring its similarity to the prototype of the category. The prototype is simply the mean of the semantic representations of its members (i.e. bimodal encodings in our case). We evaluate the models by correlating their predictions against elicited typicality ratings.

**Model Parameters** We use the SAE model and all comparison models described in Experiment 5 (Section 6.1.2). In addition, we compare against O’Connor et al. (2009) who model typicality ratings by means of an attribute-based attractor network. Their network learns concept and category representations using the attributes of the McRae norms<sup>12</sup> as target output. The learned representations correspond to the output

---

<sup>12</sup>They excluded taxonomic attributes from the norms.

Models	Typicality		
	T	V	T+V
McRae	0.45	0.14	0.41
Attributes	0.37	0.17	0.40
SAE	0.42	0.24	0.43
SVD	–	–	0.38
kCCA	–	–	0.19
Bruni	–	–	0.42
skip-gram	0.40	–	–

Table 6.8: Mean Spearman’s  $\rho$  between gold typicality ratings and model produced ratings over 20 categories used in O’Connor et al. O’Connor et al. (2009).

activations of the network. It is similar to our SAE model in that it yields semantic representations from attributes, however in our case the attributes are *learned* from data. We again used the cosine similarity to estimate the similarity between two words.

### 6.3.2 Results and Discussion

Our results are summarised in Table 6.8. As can be seen, the SAE falls slightly behind the human-produced McRae textual attributes (see row McRae, column T in the table) and is consistently better at predicting typicality ratings compared to all other bimodal models except Bruni’s model, for which this is the first experiment where it comes close to the effectiveness of the SAE. Interestingly, the SAE unimodal representations also outperform comparison models based on a single modality. Textual SAE (see column T in Table 6.8) is better than textual Attributes and skip-gram. And visual SAE (see column V) is better than McRae and Attributes. O’Connor et al.’s (2009) attractor network yields a correlation coefficient of 0.39 on the same dataset using human-produced attributes.

We also provide a more detailed comparison to O’Connor et al.’s (2009) model in Table 6.9 where we show typicality correlation coefficients for individual categories. Note that their model (AttrNN), perhaps counterintuitively, yields negative correlations for some categories, whereas this is not the case for the SAE. Although we would expect AttrNN to perform overall better given that it is trained on human-produced attributes, we observe that SAE does better on most categories.

Category	AttrNN	SAE	Category	AttrNN	SAE
FURNITURE	<b>0.76</b>	0.75	BIRD	<b>0.62</b>	0.24
APPLIANCE	<b>0.69</b>	0.64	INSECT	<b>0.55</b>	0.13
WEAPON	<b>0.63</b>	0.57	VEGETABLE	<b>0.47</b>	0.29
UTENSIL	<b>0.50</b>	0.32	FISH	0.38	<b>0.50</b>
CONTAINER	0.49	<b>0.71</b>	ANIMAL	0.12	<b>0.38</b>
CLOTHING	0.46	<b>0.55</b>	PET	0.08	<b>0.23</b>
MUSICAL	0.44	<b>0.56</b>	MAMMAL	0.02	<b>0.12</b>
TOOL	<b>0.38</b>	0.16	CARNIVORE	0.61	<b>0.66</b>
VEHICLE	-0.02	<b>0.45</b>	HERBIVORE	-0.14	<b>0.14</b>
FRUIT	<b>0.73</b>	0.30	PREDATOR	-0.05	<b>0.27</b>

Table 6.9: Correlation (Pearson’s  $r$ ) between model and human typicality ratings for 20 categories used in O’Connor et al. (2009). Comparison between their attractor network and SAE.

An analysis of the performance of the unimodal SAEs on individual categories (correlation coefficients not shown) reveals that the textual SAE is better at judging typicality for categories whose formation underlie functional or behavioural commonalities (e.g., APPLIANCE, CARNIVORE, MAMMAL, MUSICAL INSTRUMENT, PET, WEAPON). This is not surprising in view of Strudel’s bias (from which we obtained the textual input representations) towards attributes denoting actions, functions, or situations (Baroni et al., 2010). Furthermore, the bimodal SAE is better than or equal to both unimodal SAEs on many individual categories (e.g., ANIMAL, APPLIANCE, CONTAINER, TOOL, VEHICLE). This indicates, as we have already noted in the previous experiments, that the model yields representations which capture shared and complementary information provided by the two modalities.

## 6.4 Conclusions

We assessed the ability of our SAE model, presented in Chapter 5, to account for human behaviour in cognitive tasks related to word similarity. Its visual and textual input modalities were approximated by attribute-based representations (see Chapter 4). We evaluated the SAE in comparison to other attribute-based models applying different

integration mechanisms (kCCA, SVD, concatenation), an approach which uses SVD to integrate visual and textual input based on bag-of-(visual-)words-representations (Bruni), and a text-based neural network model (skip-gram). We found that our model gave in most cases a better fit to behavioural data (see Tables 6.2, 6.7, 6.8).

Specifically, we demonstrated the effectiveness of SAE's integration mechanism, performing overall better than other bimodal models augmented with the same input (kCCA, SVD, concatenation). Furthermore, in direct comparison of our visual attribute-based representations (Attributes) to human-produced attribute norms (McRae), the former achieved consistently better results across all tasks (see columns V in Tables 6.2, 6.7, 6.8). This indicates that we can utilise the attribute classifiers (Section 4.2.4) to derive visual representations which are not less informative than human-produced visual attributes.

# Chapter 7

## Image-related Tasks

In the previous chapter we made the claim that the learning of semantic representations benefits from language *and* vision. This was experimentally demonstrated on language-based tasks, where visually grounded models of word meaning simulated human behaviour better than their unimodal counterparts. A question that naturally arises is whether such bimodal models can also benefit vision-related tasks. We address this question in this chapter and present two experiments in which only visual input in the form of images is provided. In both experiments we use our attribute-centric, bimodal stacked autoencoder (SAE). In the first experiment, linguistic information needs to be explicitly inferred by the SAE. Specifically, the task is to produce visual and textual attributes for concepts when only presented with their images. The second experiment uses images as a stand-in for real-world objects for a task on visual category learning, where a system has to indicate whether a given object belongs to a category represented by a few example objects. This task does not rely on linguistic information in any way. However, it allows us to test whether bimodal image representations implicitly representing linguistic information are nevertheless useful.

### 7.1 Experiment 8: Generation of Attributes

In this section, we conduct a small-scale experiment which evaluates how accurately our approach to visually grounded meaning representations (Chapters 4 and 5) can produce visual and textual information for *new* concepts. By new concepts we mean those which neither the attribute classifiers (Section 4.2.4) nor the SAE model (Section 5.3) have encountered during training since they are not covered in the VISA dataset (Section 4.2.3). In Chapter 3 attribute norms were used as an approximation of the percep-

Concept	Visual Attributes
<i>one-armed_bandit</i>	is_box_shaped has_keys has_symbols has_speakers is_rectangular has_keyboard has_shelves made_of_wood made_of_metal has_handle has_legs has_racks has_windows made_of_plastic
<i>scabbard</i>	has_guard has_blade is_silver made_of_steel made_of_metal has_shaft is_T_shaped
<i>beer_glass</i>	is_deep has_lid is_transparent has_semicircular_handle is_concave is_cylindrical
<i>lychee</i>	has_skin has_stalks has_leaves has_seeds is_pink has_green_top has_peel is_red is_round has_layers has_pit is_small is_green comes_in_bulbs is_orange is_yellow

Table 7.1: Examples of concepts not covered by VISA and their visual attributes obtained by deriving the centroid of the attribute vectors of individual images (Equation (4.1), Page 69). Attributes are ordered in decreasing order of their scores.

tual modality and combined with the textual modality to infer perceptual information for concepts contained in the norms. In the present experiment, we obtain visual and textual information for new concepts using the visual modality, as approximated by images depicting them. We focus on the following two questions:

- (1) Can our visual attributes and the attribute classifiers generalise to new concepts, yielding accurate descriptions of their visual properties?
- (2) Can the SAE model infer textual information for new concepts when presented only with their predicted visual attributes?

To address these questions, we automatically obtain attributes for new concepts and evaluate them against human-produced attribute norms.

### 7.1.1 Data

As an approximation of a gold standard, we use the CSLB norms (Devereux et al., 2013, see Chapter 3, Section 3.1). The norms cover 639 basic-level concepts (e.g., *frog*, *shoe*, *flower*), 416 of them are also in the McRae norms<sup>1</sup>. Our test set is based on the data provided for the ImageNet Large Scale Visual Recognition Challenge (ILSRVC, Russakovsky et al., 2014). This dataset comprises 1.2 million images from ImageNet

<sup>1</sup>The list of overlapping concepts between the McRae norms and CSLB norms can be found in the supplementary material of Devereux et al. (2013).

CSLB attributes	ILSVRC attributes		
	$\gamma = 0.0$	$\gamma = 0.1$	$\gamma = 0.2$
is_cylindrical	is_cylindrical	is_cylindrical	—
is_long	is_long	is_long	—
is_white	is_white	—	—
(is_thin)	—	—	—
(has_roots)	—	—	—
has_a_stalk_stem	has_stalks	has_stalks	has_stalks
has_leaves	has_leaves	—	—
is_green	is_green	is_green	is_green
has_layers	—	—	—
(is_green_and_white)	—	—	—
—	has_flowers    has_peel	has_flowers    has_peel	
	is_yellow    has_top	has_top    (has_core	
	has_green_top    has_skin	has_ferrule	
	is_small    (has_core	has_pointed_end)	
	has_ferrule		
	has_pointed_end)		
<b>Recall</b>	0.86	0.57	0.28
<b>abs. Recall</b>	0.60	0.40	0.20
<b>Precision</b>	0.46	0.57	1.00

Table 7.2: Example of comparing CSLB attributes and generated attributes for concept *leek*. Attributes in parentheses were not found among the set of attributes of the respective other 'norm', and '—' denote attributes only listed in one 'norm' for the concept.

(Deng et al., 2009) for 1,000 subordinate-level synsets (e.g., *spring frog*, *running shoe*, *dahlia*; see Section 4.2.2 for details on ImageNet).

Since the CSLB and the ILSVRC concepts are not on the same level of abstraction, we leverage WordNet's taxonomy (Section 4.2.2) in order to find correspondencies between them. Specifically, we retrieve all hypernyms for an ILSVRC synset. For example, the hypernyms for the synset *marmoset* are *New World monkey*, *monkey*, *primate*, *mammal*, and so on. Then we map the synset to the hypernym found in the CSLB concepts. For example, the ILSVRC synsets *marmoset* and *baboon* are both mapped to the CSLB concept *monkey*. Excluding the CSLB concepts covered by VISA and those not contained in the ILSVRC data yields a test set of 62 target concepts, for each of which we sample approximately 600 images.

### 7.1.2 Visual Attribute Generation

In order to generate a list of visual attributes for each target concept, we derive visual attribute vectors using Equation (4.1) (Page 69) and subsequently keep all attributes

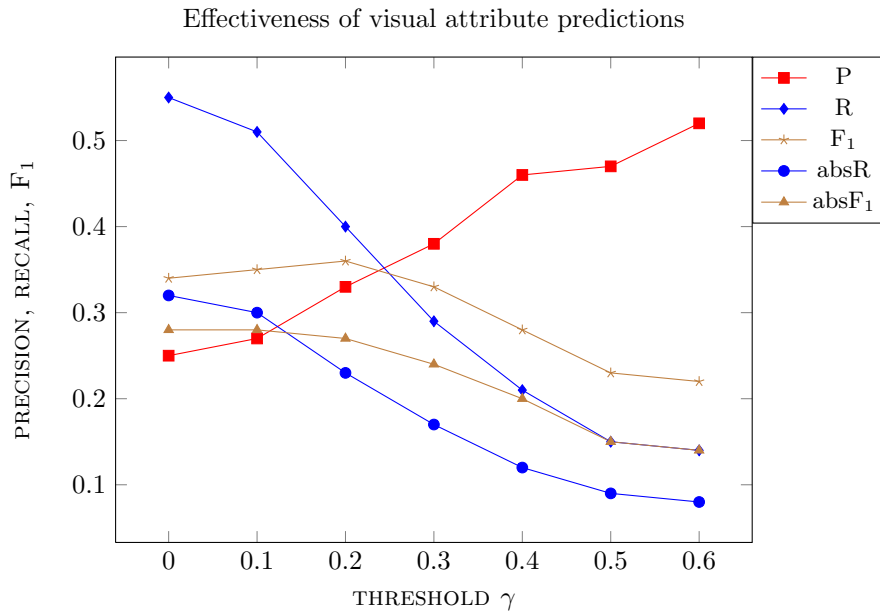


Figure 7.1: Effectiveness of the prediction of visual attributes for CSLB concepts unknown to VISA. Predictions were obtained by means of visual attribute classifiers.

with a score greater than a cutoff  $\gamma \in \mathbb{R}$ . Examples of concepts and the obtained visual attributes are shown in Table 7.1. Since the attributes of the CSLB norms and our attributes do not necessarily coincide, we lemmatise them using the Stanford CoreNLP toolkit (Manning et al., 2014) and automatically align them across all target concepts. For cases where there is no string match<sup>2</sup>, we choose the attribute with the lowest Jaccard distance  $d < 1.0$  to the source attribute. For example, `has_a_hingeCSLB` is mapped to `has_hinges`, `is_a_cylinderCSLB` to `is_cylindrical`, `has_a_long_handleCSLB` is mapped to `has_long_handle`, and `has_four_tinesCSLB` to `has_prongs`, whereas `has_controlsCSLB` is left unaligned.

### 7.1.3 Textual Attribute Generation

We generate a list of textual (Strudel) attributes for each of the target concepts, represented by a visual attribute vector, by means of our SAE model, which we used in the experiments presented in the previous chapter (see Section 6.1, Page 100, for a description of its training data and its parameters). More specifically, the SAE receives the visual centroid vectors of the target concepts as input and maps them to the textual

<sup>2</sup>We ignore concrete numbers, for example, `has_4_legsCSLB` and `has_2_legs` are considered a match. Furthermore, we make use of the alternative attribute strings provided in the CSLB norms for each attribute (e.g., `has four prongsCSLB` has the alternative `has four tinesCSLB`).



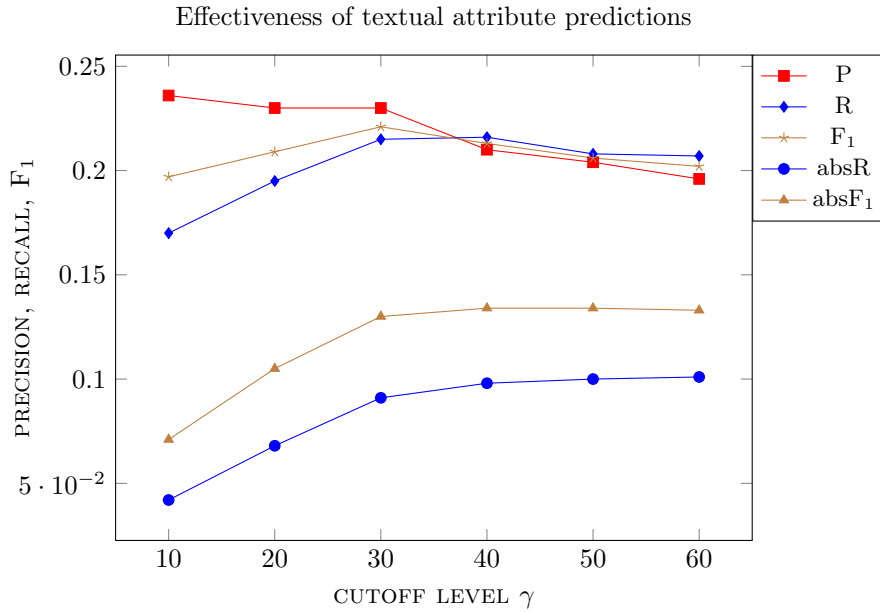


Figure 7.2: Effectiveness of the prediction of textual attributes for CSLB concepts unknown to VISA. Predictions were obtained by projecting visual input vectors to textual attributes using the SAE model.

output layer, while leaving the textual input layer unclamped. To compensate for some potential bias towards certain textual attributes in the VISA training data, we discard all predicted attributes for a target concept with an output value lower than a threshold  $\zeta$ . We estimate  $\zeta$  on the VISA training images, by first obtaining textual attribute predictions for each VISA concept in the same way as explained above, i.e. by feeding the visual centroid vectors into the SAE while leaving the textual input unclamped, and obtaining textual attribute predictions  $\hat{\mathbf{X}} \in \mathbb{R}^{N \times A}$  as output. Threshold  $\zeta(a)$  for an individual attribute  $a$  is then computed as the median of all its output values  $\hat{\mathbf{x}}_a$ <sup>3</sup> plus 2 times the Median Absolute Deviation (MAD; Leys et al., 2013):

$$\zeta(a) := \text{median}(\hat{\mathbf{x}}_a) + 2 \text{MAD}(\hat{\mathbf{x}}_a), \quad (7.1)$$

where  $\hat{\mathbf{x}}_a$  contains the predicted scores for attribute  $a$  for all VISA concepts, and MAD is computed as

$$\text{MAD}(\hat{\mathbf{x}}_a) = b \text{median}(|\hat{\mathbf{x}}_a - \text{median}(\hat{\mathbf{x}}_a)|), \quad (7.2)$$

with  $b = 1.4826$ , assuming normally distributed data. From the remaining attributes we select the  $\gamma \in \mathbb{N}$  textual attributes with highest score as final predicted attribute list for a concept.

<sup>3</sup>For simplicity, we use  $a$  to denote both an attribute and its index.

CSLB attributes are not thresholded, but only non-visual attributes are considered. Our textual attributes are aligned with CSLB non-visual attributes by applying the same alignment procedure as for the visual attributes. For example, *is\_dangerous*<sub>CSLB</sub> is mapped to *dangerous:j*, *is\_sat\_on*<sub>CSLB</sub> to *sit:v*, and *is\_a\_crustacean*<sub>CSLB</sub> to *crustacean:n*, whereas *does\_clean\_floors*<sub>CSLB</sub> is left unaligned as no correspondence could be found.

#### 7.1.4 Evaluation measures

We evaluate the generated attribute lists against the gold standard using average precision, recall and their harmonic mean, i.e.  $F_1$ -score. The true positives are the attributes contained in both our derived lists and in the CSLB norms for an individual concept. We report two recall scores: one assumes the gold positives to be all attributes listed for a concept by the CSLB norms, and the other limits the gold positives to attributes that also appear in the derived attribute lists, and could thus be potentially listed for a concept. We also report the effectiveness for varying cutoff levels  $\gamma$ . Table 7.2 shows an example of how we compare visual attributes across the two norms.

#### 7.1.5 Results and Discussion

Figure 7.1 presents the results on the visual attribute prediction task varying cutoff levels from  $\gamma = 0$  to 0.6. Recall that  $\gamma$  is imposed on each concept's visual attribute vector. The best  $F_1$ -score is obtained with  $\gamma = 0.2$  ( $P = 0.33$ ,  $R = 0.40$ ,  $F_1 = 0.36$ ). With respect to absolute recall, (i.e. when considering all CSLB attributes including those which could not be aligned to VISA attributes)  $\gamma = 0.1$  yielded the best tradeoff ( $P = 0.27$ ,  $\text{absR} = 0.30$ ,  $F_1 = 0.22$ ). The oracle recall, i.e. the proportion of CSLB attributes which could be aligned with the VISA attributes, is  $R = 0.60$ . Although there is a large margin between model and oracle recall, the results on this intrinsic experiment demonstrate that automatically derived visual attribute norms are a promising approximation of human generated visual concept information, which is particularly valuable for concepts for which such information is not available.

The results on predicting textual attributes are given in Figure 7.2 for cutoff levels ranging from 10 predicted attributes per concept to 60. The best  $F_1$ -score is obtained when considering  $\gamma = 30$  attributes ( $F_1 = 0.22$ ,  $P = 0.23$ ). The best absolute  $F_1$ -score is fairly low ( $F_1 = 0.13$ ) with an absolute recall of 0.10, whereas oracle recall is 0.60.

Recall that the set of textual attributes has not been predefined, but was extracted automatically from a text corpus by means of Strudel (Baroni et al., 2010), and the

Concept	Textual Attributes	
	Inferred (SAE)	Extracted (Strudel)
<i>currant</i>	<i>fruit:n sugar:n pickled:j flavor:n cultivate:v</i>	<i>bun:n fruit:n pastry:n cake:n dough:n salad:n pick:v sour:j juice:n ripe:j</i>
<i>needle</i>	<i>become:v use:v cut:v sharpen:v wrist:n end:n</i>	<i>syringe:n stitch:n insert:v compass:n yarn:n inject:v skin:n loop:n vein:n deflect:v</i>
<i>jellyfish</i>	<i>silver:j color:v fisherman:n catch:v swim:v fish:v carpet:n white:j ocean:n fishing:n</i>	<i>tentacle:n sting:n bloom:n gene:n venom:n hydrozoan:n water:n bell:n feed:v creature:n</i>
<i>newspaper</i>	<i>glossy:j storey:n eat:v publish:v diner:n interview:n travel:v wait:v dinner:n column:n</i>	<i>publish:v article:n editor:n write:v circulation:n print:v column:n journalist:n publication:n headline:n</i>

Table 7.3: Examples of concepts not covered by VISA and their 10 textual attributes inferred by the SAE model (left column) or extracted by running Strudel (Baroni et al., 2010) on Wikipedia (right column). Attributes are ordered in decreasing order of their scores.

attributes which apply for an individual concept are predicted by the SAE model solely on the basis of visual attribute predictions extracted from images. As a consequence, errors can be attributed to various factors. An analysis shows that errors are not only made by the SAE, which might predict plainly wrong attributes for a concept (e.g., *fish:v* was predicted for *bikini*), or wrong attributes which are related in meaning to the concept’s category (*seedless:j* for *mango*). In addition, the disparity of the two underlying attribute sets hampers the alignment of their attributes. For example, *wear:v* (predicted for *mask*) and *don:v* (*suit*) had not been aligned to the CSLB attribute *is worn<sub>CSLB</sub>*, and *appliance:n* was predicted for *can\_opener*, but CSLB lists the related attributes *is\_a\_utensil<sub>CSLB</sub>*, *is\_a\_tool<sub>CSLB</sub>* for this concept. On the other hand, many predicted attributes were assessed as wrong since they do not have a correspondence in the CSLB norms, mostly because they are related to the concept in a very broad sense (e.g. *fried:j* for *mussel*, *aroma:n* for *currant*, *frequency:n* for *television*, *animated:j* for *monkey*, *cultivar:v* for *daisy*), or because they are visual and therefore not considered in this evaluation (*carapace:n* for *mussel*, *bark:n*, *green:j* for *elm*). It is generally correct to consider these cases as errors when addressing the task of automatically generating attribute norms. However, a human-based evaluation could provide a fairer answer to how well the SAE is able to output textual information from visual input.

To determine an upper bound on how well textual attributes obtained with Strudel can approximate the gold standard norms, we ran the system on the Wikipedia corpus which we had used to train the SAE model (see Section 6.1.2). For 53 concepts of the 62 test concepts, Strudel extracted at least one textual attribute. We evaluated the output for cutoff levels ranging from the 10 highest scored attributes per concept to 60, and obtained the best  $F_1$ -score when considering 10 attributes ( $F_1 = 0.42$ ,  $P = 0.62$ ,  $R = 0.32$ ,  $\text{absR} = 0.11$ ,  $\text{absF}_1 = 0.19$ ). Strudel’s precision is more than twice as high as the precision achieved with our inferred textual attributes ( $P = 0.23$ ). Table 7.3 gives four example concepts and their 10 highest scored inferred and extracted textual attributes, respectively. As suggested by the large difference in precision, the extracted attributes are more closely related to the corresponding concept (e.g., *bun:n*, *fruit:n* for concept *currant*) than the inferred attributes (e.g., *fruit:n*, *sugar:n*). We hypothesise that the SAE learned inter-modal associations between visual attributes and *category*-specific textual attributes through the mediation of textual information injected during training. Since the examined concepts are not even part of our SAE training data, it is not surprising, though, that the inferred attributes do not capture their *concept*-specific semantic details.

## 7.2 Experiment 9: Visual Category Learning

A continuously increasing body of work as focussed on tasks lying at the boundaries of computer vision and natural language processing. Examples include image annotation, image description generation (see, e.g. Yatskar et al., 2014, and the references therein; Kulkarni et al., 2013), image retrieval for text queries, and so on. Recent methods tackling these tasks by jointly modelling image and text data were discussed in Sections 2.5 and 5.1.2. In contrast to these applications, topics that lie at the core of computer vision research can be approached without the involvement of language, such as image classification<sup>4</sup> Subject of the latter is the retrieval of images using *images* as query. Even though the need of linguistic information is not apparent for these tasks, they might still benefit from its integration, assuming that it renders images more like the way humans perceive them (e.g., Zhang et al., 2013; Frome et al., 2013).

---

<sup>4</sup>Image classification tasks expect systems to list the class names of the objects present in images, but these names are mere symbols and could as well be not interpretable by humans (e.g., Russakovsky et al., 2014; Everingham et al., 2014).

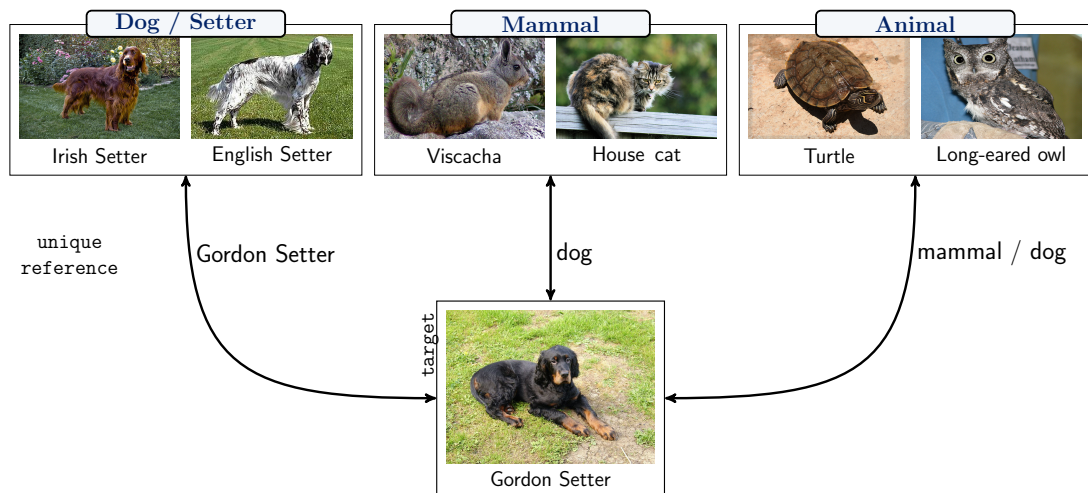


Figure 7.3: Example of the variants one can use in order to refer to an object. In the context of the leftmost pair of images representing category *dog* or *Setter*, the target object in the image in the bottom center can be referred to by the category name (*dog* or *Setter*, see label above images), or, for the purpose of a unique reference, by *Gordon Setter*. With varying contexts (image pairs in the middle and on the left, respectively), possible references for the target object also vary according to the level of abstraction.

Recently, new tasks have been introduced which require the development of algorithms that account more for how humans learn concepts and refer to real-world objects (Jia et al., 2013; Ordonez et al., 2013; Wang et al., 2014). Borrowing from research on infant category learning (e.g., Quinn and Eimas, 1986; Behl-Chadha, 1996) in cognitive science, Jia et al. (2013) define a *visual category learning*<sup>5</sup> task which taps into the ability of humans to learn new categories from just a few example objects (e.g., Xu and Tenenbaum, 2007). The task simulates corresponding experiments by using images as a stand-in for real-world objects. Given a set of example images representing a category, the system needs to infer an appropriate level of generalisation which enables it to decide whether the object present in a new image belongs to the category. A useful application scenario for this task may be image retrieval, where a user has some example images at query time representing a single category and wants to find other images from the same category (Torresani et al., 2014). Figure 7.3 illustrates the problem. Let us assume a user defines a query by means of the leftmost pair of images, representing (the subordinate-level) category *Setters*. It should be rather straightforward for a system to return the image (*Gordon Setter*) in the bottom center (*target*)

<sup>5</sup>Jia et al. (2013) call the task *visual concept learning*. We will use *category* instead in order to stay in line with our use of the term *concept* to refer to basic-level objects.

henceforth) as a retrieval hit. A query comprising the image pair in the middle is more difficult – it would require the system to generalise from the given subordinate-level or basic-level categories (*viscacha* or *house cat*, respectively) to the superordinate-level category *mammal* in order to correctly recognise the target as a retrieval hit. The right-most query represents a category of an even higher level of abstraction, requiring the system to infer the category *animal* from the given examples.

In the following section, we will describe our approach for Jia et al.’s (2013) task on visual category learning. Our research questions can be summarised as follows:

- (1) Are image representations derived from visual and linguistic information beneficial for Jia et al.’s visual category learning task?
- (2) Is our attribute-centric approach to meaning representations able to generalise to unseen concepts (i.e. those which the attribute classifiers have not seen during training)?
- (3) To which extent are representations from the SAE useful for the present task (i.e. without using them within a model specifically tailored for the task)?

## 7.2.1 Visual Category Learning

Category learning is also known as concept learning (Quinn and Einas, 1986; Tenenbaum, 1999; Ashby and Maddox, 2011). We use the former term in order to stay in line with our use of the term *concept* to specifically denote the mental representation (knowledge) of basic-level objects, such as *turtle*. Henceforth, we deviate from our notion of the term *category* given in Chapter 1 (Section 1.3), and use it to denote a set of objects at any level of categorisation, i.e. subordinate-level (e.g., *terrapin*), basic-level (e.g., *turtle*), or superordinate-level (e.g., *animal*) categories.

The task defined by Jia et al. (2013) can be formulated as follows:

**Task definition.** *Given a set of example images representing a category as well as a series of new images, indicate for each new image (query), whether it belongs to the same category as the example images.*

Note that the system is not required to output an explicit object label. Implicitly,

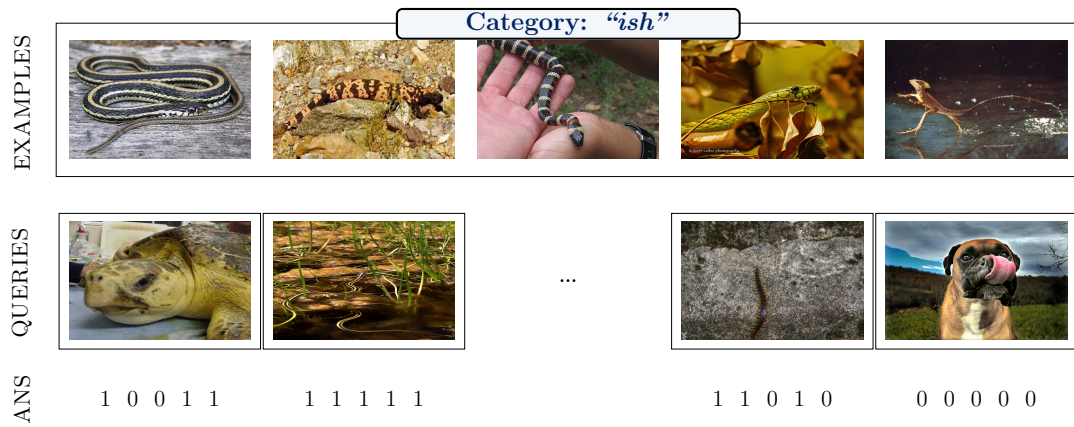


Figure 7.4: Example test item. The five example images (top) represent a category (synset  $\{\textit{reptile}, \textit{reptilian}\}$ , sampled as level 3 (‘super-basic-level’ category) from synset  $\{\textit{garter snake}, \textit{grass snake}\}$ ). For each of the 20 query images (of which four images are shown in the center), the task is to give a score denoting whether the query image is a member of the given category. Model answers are compared against human answers collected from five annotators per query (boolean answers; bottom). Categories are labelled with fantasy words (e.g., “ish”; top).

though, the name of the category transcends to all new query images (i.e. objects) which belong to the category. Figure 7.4 gives an example of the task.

Jia et al. (2013) approached this task with a Bayesian generalisation model which uses probabilistic predictions from image classifiers as input. The classifiers, at the time reportedly state-of-the-art, were trained on the ILSVRC2010 data (Russakovsky et al., 2014) with linear multinomial logistic regression using 160K-dimensional feature vectors obtained from a pipeline system (Lin et al., 2011). The ILSVRC2010 data comprises 1.2 million images categorised into 1,000 ImageNet/WordNet leaf node classes (*synsets*; see Chapter 4, Section 4.2.2 for details on ImageNet and WordNet). The core of their Bayesian model is a normalised confusion matrix  $\mathbf{A}$  which they obtained on top of these classifiers. Entry  $A_{j,i}$  of the matrix denotes the probability that the true synset is  $j$  given the classifier predicted synset  $i$ . The model operates on a hierarchy of nodes, where the set of nodes is denoted by  $\mathcal{H}$ . The hierarchy corresponds to the part of the WordNet/ImageNet taxonomy that comprises the 1,000 ILSVRC synsets as leaf nodes, denoted by  $\mathcal{S} \subset \mathcal{H}$ . The probability that a query image  $\mathbf{x}_q$  belongs to the category  $Cat$  represented by  $N$  example images,  $\mathcal{X}$ , is then given by the following

equation:

$$P(\mathbf{x}_q \in \text{Cat}|\mathcal{X}) = \sum_{h \in \mathcal{H}} P(\mathbf{x}_q|h) P(h|\mathcal{X}), \quad (7.3)$$

where  $P(\mathbf{x}_q|h)$  denotes the probability that query  $\mathbf{x}_q$  belongs to the category rooted at  $h$  (e.g.,  $\{\text{reptile}\}$ ), and is given by

$$P(\mathbf{x}_q|h) = \sum_{j=1}^{|\mathcal{S}|} A_{j\hat{y}_q} \mathbf{1}_h(j), \quad (7.4)$$

where  $\hat{y}_q$  is the classifier prediction for  $\mathbf{x}_q$ , and  $\mathbf{1}_h(j)$  the indicator function denoting whether node  $j$  is a member of hypothesis  $h$ . The posterior distribution of a category rooted at  $h$  given example images  $\mathbf{x}_i \in \mathcal{X}$  with classifier predictions  $\hat{y}_i$  is

$$P(h|\mathcal{X}) \propto P(h) \prod_{i=1}^N P(\mathbf{x}_i|h) = P(h) \frac{1}{|h|^N} \prod_{i=1}^N \sum_{j=1}^{|\mathcal{S}|} A_{j\hat{y}_i} \mathbf{1}_h(j). \quad (7.5)$$

The prior probability of  $h$  was set to  $P(h) \propto \frac{|h|}{\sigma^2} \exp(\frac{-|h|}{\sigma})$ , with  $\sigma = 200$ .

Jia et al.’s two best-performing baselines represent example and query images by the  $L_1$ -normalised output of the image classifiers. In their *prototype model* (PM), the score of a query is computed as its  $\chi^2$  distance to the closest example image. The *histogram of classifier outputs* model (HC) computes the score of a query as its  $\chi^2$  distance to histogram of classifier outputs, aggregated over the examples.

## 7.2.2 Method

As discussed earlier in this chapter, linguistic information is not required for addressing the task of visual category learning. We could therefore apply a method that uses our visual attribute representations of images only. However, this would not give an answer to the question of whether the task benefits from linguistic information. Hence our method leverages the SAE architecture introduced in Chapter 5 which is trained to integrate textual and visual information of objects.

Specifically, we first train the SAE (see Section 5.3) on visual and textual attribute-based representations of images which depict real-world objects. We obtain these input representations from our visual attribute classifiers and Strudel, respectively (see Chapter 4). At test time, we are presented with a set of example images from a category and a query image of an object. Our goal is to compute a score which denotes whether the object is an instance of the category. To this end, we derive a representation for each image from its SAE encoding and use this for membership scoring as we describe



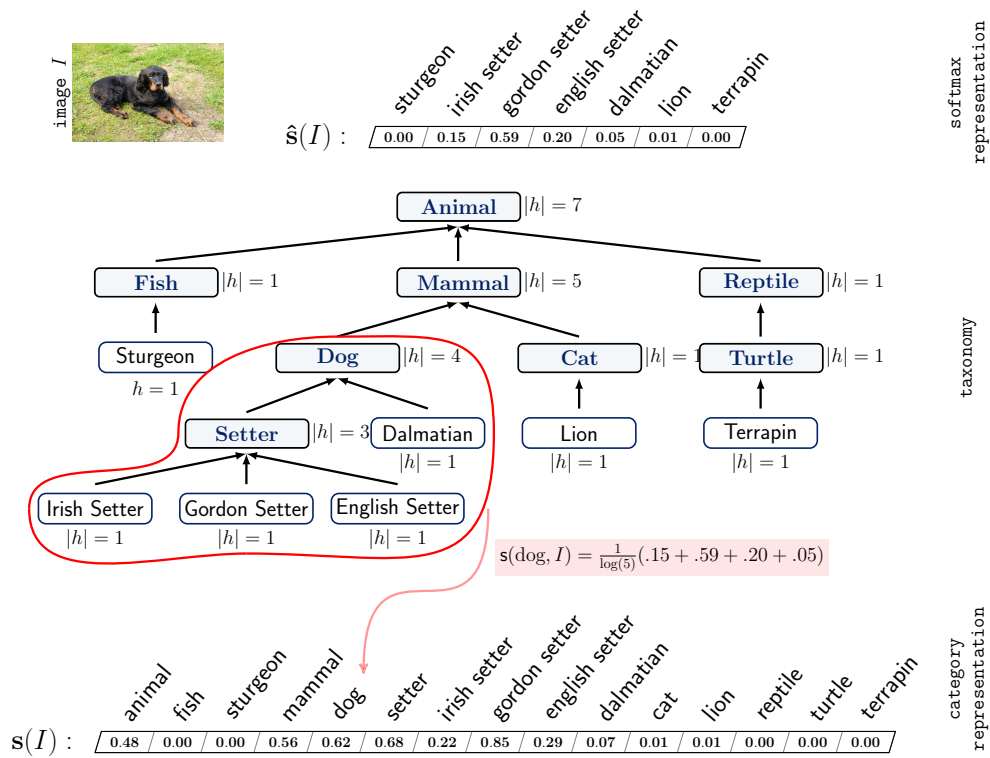


Figure 7.5: Example computation of the category-based representation  $s(I)$  for image  $I$  on the basis of its softmax representation  $\hat{s}(I)$  and a small taxonomy.

below. Note that since the object labels of all test images are unknown, we do not have textual attribute information at test time. The SAE is therefore fed with the images' visual attribute vectors while its textual input layer is set to zero.

A straightforward approach to deriving a representation from the resulting SAE encoding is to directly use its *bimodal* encoding (see Figure 7.6). A more sophisticated way is to derive a category-based representation similarly to Jia et al. (2013), by using a taxonomy as follows (see Figure 7.5 for an example of the procedure using a taxonomy of 15 nodes):

**Category-based representation** We assume that we have access to a taxonomy, such as WordNet (Fellbaum, 1998). Let  $x$  denote an image represented by its encoding  $\hat{s}(x)$  obtained from the softmax layer of the SAE.  $\hat{s}(x)$  denotes a probability distribution over all object labels  $o \in \mathcal{O}$  (the leaf nodes in the taxonomy):  $\hat{s}(x) = (P(o_j|x))_{j=1, \dots, |\mathcal{O}|}$ . We derive a category-based vector representation  $s(x)$  whose entries correspond to all (leaf and internal) nodes  $h \in \mathcal{H}$  of the taxonomy (e.g., *dog* is an

internal node in Figure 7.5), thus covering different possible levels of categorisations. The value of component  $h$ ,  $s_h(x)$ ,<sup>6</sup> is the weighted accumulation of the predictions of the leaf nodes  $o$  which are (direct or indirect) hyponyms of category  $h$  (i.e. the object labels subsumed by  $h$ ), that is,

$$s_h(x) = \frac{1}{\log(|h| + 1)} \sum_{o \in O} P(o|x) \mathbf{1}_h(o), \quad (7.6)$$

where  $|h|$  is the number of object labels subsumed by  $h$ , and  $\mathbf{1}_h(o)$  is the indicator function denoting whether leaf node  $o$  is a member of  $h$  (see example shaded in red in Figure 7.5). The weighting factor causes score  $s_h(x)$  to favour basic-level categories which is in compliance to the level of abstraction humans tend to choose (Mervis and Rosch, 1981). (An example of the resulting representation  $\mathbf{s}(x)$  is shown at the bottom of Figure 7.5.)

**Membership Scoring** Given a representation for all example images of a category and a query image, we compute a score indicating whether the query is a member of the category. We distinguish between two scoring paradigms which are similar to the baselines employed by Jia et al. (2013)<sup>7</sup>

**PM** Prototype Model

The score for a query is its cosine similarity to the most similar example.

**AM** Aggregation Model

The score for a query is its cosine similarity to the centroid of the category, computed as the average representation of all examples.

**Comparison to Jia et al. (2013)** Our category-based approach in combination with the AM scoring paradigm is technically akin to Jia et al.’s Bayesian model: Equation (7.4) (for  $P(x_q|h)$ ) is similar to the second factor of Equation (7.6) (for  $s_h(x)$ ), with the difference that Jia et al. use a confusion matrix while we directly use the probability of an object class given an image. Instead of Jia et al.’s prior for the posterior distribution (Equation (7.5)), we use a weighting term (first factor in Equ. (7.6)) and apply it to both the query and example images. Furthermore, we compute the average of the examples’ category-based representations (the AM scoring paradigm) instead of Jia et al.’s multiplication of the examples’ scores. Finally, we compute the final score using

<sup>6</sup>For simplicity, we use the symbol  $h$  to denote both, an index of  $\mathbf{s}$  and its corresponding category.

<sup>7</sup>The exact application of Jia et al.’s (2013) baselines performed worse than our PM and AM models.

the cosine similarity, which corresponds—figuratively—to applying Jia et al.’s function (Equation (7.3)) to the normalised category-based representations.

## 7.2.3 Experimental Setup

### 7.2.3.1 Evaluation Data

We evaluate the models on the test data created by Jia et al. (2013).<sup>8</sup> It comprises 4,000 tasks of twenty test items each. Every task defines a category of a particular degree of generalisation, represented by five example images, and twenty query images. For each of the 1,000 ILSVRC synsets (e.g., {*garter snake*, *grass snake*}) there are four tasks, one pertaining to the subordinate-level category (the synset, e.g., {*garter snake*, *grass snake*}), one basic-level (e.g., {*colubrid snake*, *colubrid*}), one ‘super-basic-level’ (e.g., {*reptile*, *reptilian*}) and one superordinate-level category (e.g., {*vertebrate*}). Figure 7.4 exemplifies a task, where five example queries were sampled from synset {*reptile*} pertaining to the ‘super-basic-level’ category of {*grass snake*}. The figure shows four of the twenty query images (see center of Figure 7.4).

All images were sampled from the ILSVRC2010 test images. Jia et al. used the WordNet/ImageNet taxonomy to determine the different generalisation levels per synset and sample the corresponding example and query images. Please refer to Jia et al. (2013) and the supplemental material for details on how the categories and images were selected.

The authors collected human judgements for all tasks by means of AMT, where they asked the participants to decide, for each query image, whether it belongs to the category represented by the five example images. They obtained binary judgements from five AMT participants per task (see bottom of Figure 7.4).

### 7.2.3.2 Model Parameters

In order to train the SAE model, it received both the visual and the textual modality as input. For the visual modality we followed Jia et al. (2013) in their use of the ILSVRC2010 training images (Russakovsky et al., 2014) and sampled 400 images for each of the 1,000 synsets. We obtained visual vectors for each image by means of our attribute classifiers (see Chapter 4, Section 4.2.4), and scaled them to the  $[-1, 1]$  range.

For the textual modality we relied on the same output as we used in Chapter 6 and that was obtained by running Strudel (Baroni et al., 2010) on a 2009 dump of the

---

<sup>8</sup>We are grateful to Jia Yangqing for providing us with the data.

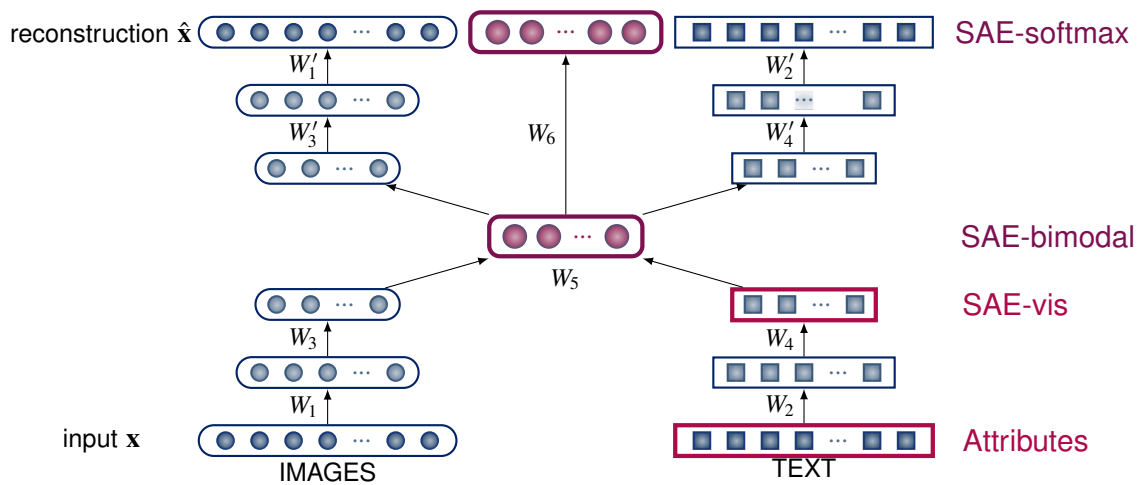


Figure 7.6: Illustration of SAE model and the representations obtained on different levels of the SAE.

English Wikipedia (see Section 6.1.2 for details). Since many of the ILSVRC synsets are often fine-grained (e.g., *siamang*), not all of them could be found in the Strudel data. We therefore leveraged WordNet and retrieved as textual representation of each synset the Strudel representation of its most specific hypernym which was available in the Strudel data (e.g., *ape*). In cases where a hypernym synset consisted of several lemmas (e.g. *trumpet*, *horn*), the corresponding vectors were averaged. The rationale for this procedure is that humans might not have gained (linguistic) knowledge of a certain object of a subordinate-level category, such as a *siamang*, but they can visually recognise the more general, probably basic-level, category of which it is a member (*ape*) and can activate their knowledge of this category. The values of the textual vectors were scaled to the  $[-1, 1]$  range.

For training the models, we set the corruption parameter (see Section 5.2.2) for pre-training the textual denoising autoencoder to 0.2, and to 0.99 for fine-tuning the SAE. The size of the bimodal layer was set to 150 units.

For the category-based representation we used two different taxonomies: First, we leveraged WordNet (Fellbaum, 1998) and obtained a subtaxonomy of 1,640 nodes which contains all ILSVRC synsets as leaves. Second, we automatically built a taxonomy by clustering the bimodal representations of the training data. That is, we computed the centroid representation of each synset on the basis of the encodings of the training data. We clamped both, the textual and visual vectors to the input layers of the SAE. We then performed agglomerative clustering on these embeddings to obtain

a hierarchical tree, using Cluto (Zhao et al., 2005).<sup>9</sup>

### 7.2.3.3 Comparison Models

In our experiments, we compare the performance of the image representations obtained by the category-based approach (using WordNet or the clusters) and by different layers of the SAE (see the highlighted layers and their labels in Figure 7.6). The baseline model represents images with their visual attribute vectors obtained from the attribute classifiers (*Attributes* in Figure 7.6). Mapping these vectors onto the second hidden visual layer gives the encodings *SAE-vis*. *SAE-bimodal* is the bimodal encoding layer, and *SAE-softmax* the encoding layer that outputs a probabilistic prediction with respect to the object label for a given input. Note that all encodings take into account linguistic information during training except for *Attributes*, i.e. the visual input vectors. The two category-based models obtain image representations using the *SAE-softmax* layer and either WordNet or the clusters (see Equation (7.6), Page 126). During test time, only the image represented by visual attribute predictions is given as input to the models, while the textual vectors are set to zero.

We furthermore compare our models to Jia et al.'s (2013) visually-grounded Bayesian model developed for the present task as well as their two best-performing baselines.

### 7.2.3.4 Evaluation

For comparison reasons, we adopt the evaluation measures applied by Jia et al. (2013). That is, we evaluate the performance of a model by comparing its answer scores to the ground truth data using average precision (AP, Everingham et al., 2014) as well as  $F_1$ -score at the point of intersection between the precision and recall curves (Jia et al., 2013). We additionally report results obtained in a binary setting in which the system has to provide a binary score indicating whether a query image is or is not member of the category. We evaluate the models by means of 4-fold cross-validation, where each training fold was used to determine a model's decision threshold.

## 7.2.4 Results

The overall results are given in Table 7.4. All our prototype-based models (\*-PM in the first block in Table 7.4) perform worse than their corresponding variants based on aggregated example representations (\*-AM, second block). This is not unexpected, as

---

<sup>9</sup>The command was `vcluster.exe -clmethod=agglo <encodingfile> <tokenfile> 500`.

Models	AP	F <sub>1</sub>	F <sub>1</sub> (binary)
Attributes–PM	49.1	54.9	54.8
SAE-vis–PM	54.0	56.0	58.3
SAE-bimodal–PM	56.6	57.4	60.1
SAE-softmax–PM	56.7	57.4	59.8
Attributes–AM	50.9	55.7	55.7
SAE-vis–AM	56.4	57.5	60.2
<b>SAE-bimodal–AM</b>	59.0	59.4	61.9
<b>SAE-softmax–AM</b>	57.5	60.0	62.3
<b>SAE-WN–AM</b>	<b>62.3</b>	<b>61.6</b>	<b>62.8</b>
SAE-cluster–AM	58.8	59.2	60.8
Jia–PM	61.74	56.07	—
Jia–HC	60.58	56.82	—
Jia–VG (actual model)	<b>72.82</b>	<b>66.97</b>	—
Human Performance	—	75.5	—

Table 7.4: Results on the visual category learning task as defined by Jia et al. (2013). We report average precision (AP) and F<sub>1</sub>. Models ending with –PM apply the prototype paradigm for membership scoring, and –AM stands for the aggregation paradigm. Results on human performance were reproduced.

the most specific category to which the object present in a query image belongs might not be present in one of the example images, but can yet belong to the same (possibly superordinate) category (cf. the example in Figure 7.3).

Our best performing baseline models are SAE-bimodal–AM and SAE-softmax–AM which are based on the bimodal encoding layers (see Figure 7.6). Attributes–\* perform worst. These are the only models which did not receive any linguistic information during training. Recall that at test time no linguistic information was input into the SAE either. One reason for the difference in effectiveness of the models may be their increasing complexity. However, SAE-bimodal and SAE-softmax perform en par, even though the latter has a higher complexity due to its additional layer. Taken together, these results suggest that the SAE learned beneficial associations between visual attributes and categories through the mediation of textual information present during training.

We report results with the category-based representations, derived using hierarchical information, only for the better performing scoring paradigm based on aggregated example representations (third block in Table 7.4). SAE-WN-AM, which uses the WordNet taxonomy, performs better than all baseline models, including the two baselines by Jia et al. (Jia-PM and Jia-HC). Yet, its difference in  $F_1$ -score is marginal compared to SAE-softmax-AM, which leverages only object label predictions. This gives an answer to our third question regarding using the image representations provided by the SAE directly: these representations are relatively strong on their own, even though they do not exploit hierarchical information. Furthermore, we observe that the taxonomy automatically obtained by clustering the bimodal encodings of the training data (SAE-cluster-AM) does not result in any performance gains, which indicates that the produced clustering does not provide more useful information over and above the SAE encodings (bimodal-AM, softmax-AM) themselves. We could systematically study to which extent it is possible to induce a more useful taxonomy on the basis of the SAE model, but we leave this for future work.

Compared to the models by Jia et al. (2013), we observe that all SAE-\*-AM models perform better than Jia-PM and Jia-HC in terms of  $F_1$ -score (fourth block in Table 7.4). Recall that Jia et al.’s models are very similar to the SAE-softmax-\* models, with the main difference being their use of state-of-the-art image classifiers but no integration of linguistic information during training. Overall, Jia-VG is the best performing model. However, the difference in  $F_1$ -score is only 7 percentage points (to SAE-softmax-AM) and 5.4 points (to SAE-WN-AM), despite the fact that this model, firstly, was developed for this specific task and directly exploits hierarchical information from ImageNet/WordNet (as does SAE-WN-AM), secondly, uses state-of-the-art image classifiers, and thirdly, additionally benefits from a confusion matrix learned on top of the image classifiers. Their model does not make use of direct linguistic information, such as textual attributes, to which our SAE-model is exposed during training.

We performed experiments with a confusion matrix which we obtained as described by Jia et al. However, in our case, the use of the matrix did yield comparable or even inferior results. A reason for this might be that our attribute-based image representations (of 414 dimensions) are simply less effective in discriminating between the subordinate-level target classes compared to the 160K-dimensional image vectors used by Jia et al.

We report human performance in the last line in Table 7.4. Following Jia et al. (2013), it was estimated by randomly sampling one human participant per task and

Models	In VISA				Not in VISA			
	All levels		Level 0		All levels		Level 0	
	AP	F <sub>1</sub>	AP	F <sub>1</sub>	AP	F <sub>1</sub>	AP	F <sub>1</sub>
SAE-bimodal-AM	56.5	57.3	56.0	57.0	60.5	60.6	61.6	61.6
SAE-softmax-AM	55.2	57.8	57.3	58.8	58.8	61.3	62.1	62.8
SAE-WN-AM	56.7	58.6	55.6	56.7	60.8	62.3	62.3	61.7
SAE-cluster-AM	55.9	56.7	56.6	56.8	59.6	59.8	61.4	60.9
Human Performance	—	74.2	—	79.8	—	76.2	—	80.3

Table 7.5: Results on category learning task as defined by Jia et al. Distinction between known and unknown objects in the VISA dataset. We report microaverage precision (AP) and F<sub>1</sub>.

comparing her prediction against the others.

**Comparison between seen and unseen objects** To answer our second question on whether our attribute-centric approach can generalise to unseen object classes, we evaluated our best performing models by contrasting the synsets available in the VISA dataset (see Section 4.2.3) with those unknown to the dataset.

We considered all synset matches as known (13.7% of the 1,000 leaf synsets), and those synsets whose subsumed words are in VISA (e.g., synset {*chard*, *Swiss chard*, *spinach beet*, *leaf beet*} is not in VISA, but *spinach* is). This resulted in a set of 628 unseen synsets. The evaluation data only provides the synsets of the example images representing a task of the subordinate-level category (i.e. the images all depict one of the 1,000 leaf synsets). As a consequence, example images of higher-ordinate categories (e.g., {*reptile*}) could still depict an object known to VISA, since the examples are sampled from the leaf synsets subsumed by the category. We therefore report the results obtained for the categories of all levels and for the subordinate-level categories only.

Recall that the attribute classifiers which we use to obtain visual representations were trained on the concepts in VISA. We would thus expect to achieve higher performance for the synsets that are also available in VISA, as the dataset might be missing attributes necessary to sufficiently describe the unknown synsets. As shown in Table 7.5, this is not the case. Similarly to the human performance, our models achieve even higher performance on the tasks for unknown synsets (column labelled Not in



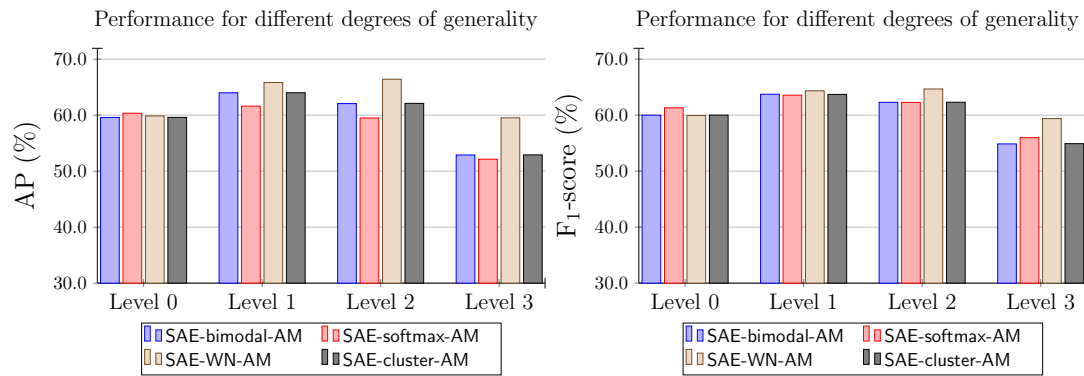


Figure 7.7: The performance of the best models for the different levels of categorisation. Shown is the AP (left) and the F<sub>1</sub>-score (right).

VISA in Table 7.5). This indicates that our attribute-centric representation approach is indeed able to generalise to unseen objects.

**Comparison between different levels of abstraction** For a deeper insight into the effectiveness of the learned meaning representations, we evaluated the performance of our best models grouped according to different degrees of generality of the categories. Figure 7.7 shows the results. Interestingly, the model which uses the object predictions as image representations (SAE-softmax-AM) performs better than the other models only for subordinate-level categories (level 0). With an increasing level of abstraction, the model that utilises the WordNet taxonomy (SAE-WN-AM) becomes superior to the other models (level 1 to level 3). Best results are obtained on level 1 and level 2, which roughly correspond to basic-level and 'super-basic-level' categories. This gives a more conclusive answer to our third question: even though the general-purpose meaning representations (SAE-bimodal-AM) are effective on the task of learning visual categories and are at least as good as representations based on object predictions (SAE-softmax-AM), they fall short in comparison to models that are more tailored to inferring the correct level of abstraction (SAE-WN-AM).

## 7.3 Conclusions

We addressed the question whether bimodal models can benefit image-related tasks and presented two experiments in which only images were provided. We used our attribute-centric, stacked bimodal autoencoder (SAE) in both experiments which was

trained on visual and textual information, but was presented with visual information only at test time.

The task of the first experiment was to produce visual and textual attributes for new concepts represented by images. We generated visual attributes using our attribute classifiers. These attribute predictions were input to our SAE in order to infer textual attributes. We evaluated the attribute predictions against the CSLB attribute norms (Devereux et al., 2013). The second task on visual category learning was to indicate whether an image is an instance of a category represented by a few example images. Model output was evaluated against human judgements.

In summary, the experimental results demonstrated the ability of our attribute classifiers to generalise to new concepts. Moreover, we showed that the SAE is able to activate linguistic information from purely visual input. Firstly, it inferred textual attribute descriptions for new concepts from their visual representations, and secondly, the representations obtained by the bimodal autoencoder proved more effective on the visual categorisation task compared to its visual input representations or unimodal autoencoders.

# Chapter 8

## Conclusions

We conclude the thesis with a summary of the main findings in light of the claims put forward in the introduction, and discuss avenues for further research.

### 8.1 Main Findings

This thesis presented an approach to grounding lexical meaning representations by integrating linguistic and visual information. The key aspects of our approach are two-fold. Firstly, information is rendered in natural language attributes for both modalities, and is obtained automatically from text and image data, respectively (Chapter 4). Secondly, our grounding method is based on a deep stacked autoencoder architecture (SAE) which learns bimodal meaning representations from visual and linguistic input in a joint manner by means of a semi-supervised criterion (Chapter 5). Feeding the attribute-based representations as input to the SAE yields a bimodal framework which we used to test the main claims of this thesis.

Our first claim was that the integration of visual and textual information of concrete concepts yields meaning representations which more closely approximate human conceptual knowledge compared to purely text-based models. The ability of our model to simulate human behaviour was evaluated on three semantic tasks related to concept similarity. Experimental results showed that the SAE model yields an overall better fit with behavioural data than unimodal (textual or visual) models (Chapter 6). The claim was further supported in Chapter 3 in similar evaluation settings, where we used existing methods to integrate standard distributional models with human-produced attributes as a proxy for the visual modality.

Our second claim stated that the visual modality can be approximated by information rendered in natural language attributes and extracted from images. We first demonstrated that we can automatically obtain visual attribute predictions from images which can be used as a substitute for human-produced visual attributes in visual grounding models (Chapter 4, Section 4.4). In Chapter 6 (Section 6.1), we further showed models using our visual attributes were dominating in predicting visual similarity of concepts, whereas textual attributes were dominating in predicting semantic similarity judgements. Also, in Chapter 7, we demonstrated the generalisation ability of our visual attribute classifiers in two tasks, i.e. generating visual attributes for unseen concepts (Section 7.1), and in visual (image-based) categorisation (Section 7.2).

Our third claim stated that the visual and textual modalities are interrelated and that it is therefore beneficial to use joint integration methods which derive bimodal meaning representations by finding and exploiting their associations. This claim was empirically validated in Chapters 3, 4, and 6, where we experimentally compared joint integration methods to a simple concatenation approach. We found that joint models can better simulate human behaviour on different cognitive tasks related to word similarity and association. Furthermore, in Chapter 6 we demonstrated that our modelling approach yields bimodal representations which simulate human behaviour more effectively than all bimodal comparison methods across all tasks.

The benefits of a joint integration mechanism from a more practical perspective were addressed in Chapter 7 in image-related tasks, where our model was trained on visual and textual information, but was presented with visual information only at test time. In Section 7.1, we generated textual attributes for concepts when provided with their images, and showed that our model can activate linguistic information from purely visual input. That the latter can be useful even in purely image-based tasks is indicated in the visual categorisation task (Section 7.2), where representations obtained by the bimodal autoencoder proved more effective compared to visual input representations or unimodal autoencoders.

## 8.2 Future Work

In this section we discuss avenues for future work. Specifically, we focus on how the visual and textual modalities could be further improved and highlight applications

which could benefit from our approach.

**Visual Modality** As outlined in Chapter 4 (Section 4.1) and experimentally demonstrated (in Chapters 4, 6, and 7), there are good reasons for adopting an attribute-centric approach in the context of the visual grounding problem. A shortcoming, however, is that the information visual attributes can capture is limited to knowledge that can be easily verbalised. For example, information about extraordinary shapes or patterns, relations between parts of objects, and so on, cannot be provided. Furthermore, in order to learn the visual attribute classifiers, we trained support vector machines on rather shallow image representations (based on bag-of-visual-words), which themselves may be missing relevant information from the images. Even though our attribute classifiers proved to be sufficient for approximating the visual modality, the training method and therefore probably their accuracy falls short of what might be possible with state-of-the-art methods. In Section 5.1 we mentioned the success of convolutional neural networks (CNNs) for computer vision tasks (e.g., image classification) which is due to their ability to learn powerful feature representations from (almost) raw pixels values (Krizhevsky et al., 2012). CNNs typically consist of a stack of alternating convolutional<sup>1</sup> and (optionally) pooling<sup>2</sup> layers as lower layers, and fully connected upper layers, with a supervised layer on top.

Related work on visually grounded representations (Kiela and Bottou, 2014) used the features yielded by the last hidden layer of a CNN trained for object classification. An appealing alternative would be to apply a training paradigm which retains our attribute-centric approach, e.g., by simply training attribute classifiers with CNN features (Escorcia et al., 2015), which could result in more accurate classifiers.

Note that the SAE model is not attribute-specific, but could be used to derive bimodal meaning representations on the basis of any text and image features, including distributed input representations. We could therefore apply the SAE to integrate visual vectors obtained from a CNN with word representations learned by means of, e.g., Mikolov et al.'s (2013b) skip-gram model. As a side effect, this would allow for a more direct and fairer comparison of attribute-based image vectors and distributed

---

<sup>1</sup>In a convolutional layer, each unit is connected to a subset of contiguous units of the previous layer, where the weights (convolution kernel or filter) are shared by all units. There may be several different filters in each layer. A convolutional layer acts as feature extractor (e.g., it detects edges in the lower layers).

<sup>2</sup>Each unit in a pooling layer is connected to a subset of contiguous units of the previous layer from which it subsamples its output (e.g., by taking the maximum). The purpose of pooling layers is variance reduction.

state-of-the-art representations. In our experiments in Chapter 6, we compared models using our attribute-based approach only to a model which integrates bag-of-visual-words (BoVW) representations, which was clearly outperformed by the former.

With respect to the way information is exploited from image data, an extension would be possible by, in analogy to the distributional hypothesis, including information explicitly extracted from outside the bounding boxes of objects (see also Bruni et al., 2012b).

**Textual Modality** We employed Strudel, an off-the-shelf method (Baroni et al., 2010) for acquiring textual attributes. The latter are typically single words and provide rather general information. For example, for the word *blender*, Strudel extracts the attributes *use:v*, *drink:n* or *liquid:n*, whereas the norms of McRae et al. (2005) contain *used\_for\_making\_drinks|mixing\_liquids* or *used\_for\_chopping\_food*. Future work could focus on the development of methods for deriving more informative textual attributes. We could for example learn extraction patterns, or feature representations, with indirect supervision provided by McRae et al.'s (2005) non-visual attributes. Specifically, we could learn a mapping from textual information (e.g., a target word's context words or textual attributes from Strudel) to human-produced attributes (a simple example is *use*, *drink*, *liquid* which corresponds to *used\_for\_making\_drinks*).

**Concepts** In the evaluation of our model on cognitive tasks (Chapter 6) we focussed on concepts known to the VISA dataset. A natural and obvious extension of the presented work is the inclusion of unseen concepts. An extension to other concrete nouns is straightforward and only a minor step into the direction of a full modelling account of visually grounded semantic representations. More intriguing is to enhance our work to action verbs, which raises new research questions, for example: How can we extract and capture visual information from images with respect to the motion and the participants involved in the action referred to by a verb?

In this context, it is necessary to point out that a model of semantic representation should also account for abstract concepts which make up a large part of the vocabulary. However, this is outside the scope of the thesis.

**Applications** In this thesis, we have we only scratched the surface of image-based applications which could potentially benefit from linguistic information (Chapter 7, Section 7.2). In the future we could apply our bimodal model to zero-shot learning

(i.e. the classification of objects for which there are no training examples; see Sections 2.5.2 and 4.2.1), or entry-level categorisation (i.e. the prediction of the word people most likely use to refer to a depicted object; Ordonez et al., 2013). Furthermore, as discussed in this thesis, visual attributes enable a system to give information about depicted objects of classes unknown to image classifiers. A less straightforward application would therefore be some form of hybrid object classification, in which objects are labelled with their closest class and additionally described with distinctive attributes with respect to that class (e.g., *horse* and *stripes* for the unknown class *zebra*).

# Appendix A

## VisA Dataset

### A.1 Concepts and Synsets in VisA

Concept	Synset ID	Words denoting synset
accordion	n02672831	accordion, piano_accordion, squeeze_box
airplane	n02691156	airplane, aeroplane, plane
alligator	n01698434	alligator, gator
ambulance	n02701002	ambulance
anchor	n02709367	anchor, ground_tackle
ant	n02219486	ant, emmet, pismire
apartment	n02726305	apartment, flat
apple	n07739125	apple
apron	n02730930	apron
armour	n02739668	armor, armour
ashtray	n02747802	ashtray
asparagus	n07719213	asparagus
avocado	n07764847	avocado, alligator_pear, avocado_pear, aguacate
axe	n02764044	ax, axe
bag	n02774152	bag, handbag, pocketbook, purse
bagpipe	n02775483	bagpipe
ball	n02779435	ball
balloon	n02782093	balloon
banana	n07753592	banana
banjo	n02787622	banjo
banner	n02788021	banner, streamer
barn	n02793495	barn
barrel	n02795169	barrel, cask
basement	n02800497	basement, cellar
basket	n02801938	basket, handbasket
bat_(animal)	n02139199	bat, chiropteran
bat_(baseball)	n02799175	baseball_bat, lumber
bath tub	n02808440	bath tub, bathing_tub, bath, tub
baton	n02809605	baton

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).



Concept	Synset ID	Words denoting synset
bayonet	n02812949	bayonet
bazooka	n02813752	bazooka
beans	n07724943	bean, edible_bean
bear	n02131653	bear
beaver	n02363005	beaver
bed	n02818832	bed
bedroom	n02821627	bedroom, sleeping_room, sleep- ing_accommodation, chamber, bedchamber
beehive	n02822865	beehive, hive
beetle	n02164464	beetle
beets	n07719839	beet, beetroot
belt	n02827606	belt
bench	n02828884	bench
bike	n02834778	bicycle, bike, wheel, cycle
bin_(waste)	n02839910	bin
birch	n12281241	birch, birch_tree
biscuit	n07693972	biscuit
bison	n02410509	bison
blackbird	n01558594	blackbird, merl, merle, ouzel, ousel, Euro- pean_blackbird, Turdus_merula
blender	n02850732	blender, liquidizer, liquidiser
blouse	n02854926	blouse
blueberry	n07743544	blueberry
bluejay	n01580870	blue_jay, jaybird, Cyanocitta_cristata
boat	n02858304	boat
bolts	n02865665	bolt
bomb	n02866578	bomb
bookcase	n02870880	bookcase
book	n02870526	book
boots	n02872752	boot
bottle	n02876657	bottle
bouquet	n02879087	bouquet, corsage, posy, nosegay
bowl	n02881193	bowl
bow_(ribbon)	n02880189	bow, bowknot
bow_(weapon)	n02879718	bow
box	n02883344	box
bracelet	n02887970	bracelet, bangle
bra	n02892767	brassiere, bra, bandeau
bread	n07679356	bread, breadstuff, staff_of_life
brick	n02897820	brick
bridge	n02898711	bridge, span
broccoli	n07714990	broccoli
broom	n02906734	broom
brush	n02908217	brush
bucket	n02909870	bucket, pail

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
buckle	n02910353	buckle
budgie	n01821869	budgerigar, budgereegah, budgerygah, budgie, grass_parakeet, lovebird, shell_parakeet, Melopsittacus_undulatus
buffalo	n02410702	American_bison, American_buffalo, buffalo, Bison_bison
buggy	n02912557	buggy, roadster
building	n02913152	building, edifice
bullet	n02916350	bullet, slug
bull	n02403325	bull
bungalow	n02919792	bungalow, cottage
bureau	n03015254	chest_of_drawers, chest, bureau, dresser
bus	n02924116	bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus, passenger_vehicle
butterfly	n02274259	butterfly
buzzard	n01607962	buzzard, Buteo_buteo
cabbage	n07713895	cabbage, chou
cabinet	n02933112	cabinet
cabin	n02932400	cabin
cage	n02936714	cage, coop
cake	n07801508	oil_cake
calf	n01887896	calf
camel	n02437136	camel
camisole	n02944075	camisole
canary	n01533339	canary, canary_bird
candle	n02948072	candle, taper, wax_light
cannon	n02950632	cannon
canoe	n02951358	canoe
cantaloupe	n12164656	cantaloupe, cantaloup, cantaloupe_vine, cantaloup_vine, Cucumis_melo_cantalupensis
cap_(bottle)	n02954938	cap
cape	n02955767	cape, mantle
cap_(hat)	n02955065	cap
caribou	n02433925	caribou, reindeer, Greenland_caribou, Rangifer_tarandus
car	n02958343	car, auto, automobile, machine, motorcar
carpet	n04118021	rug, carpet, carpeting
carrot	n07730207	carrot
cart	n03484083	handcart, pushcart, cart, go-cart
catapult	n02981911	catapult, arbalest, arbalist, ballista, bricole, mangonel, onager, trebuchet, trebucket
caterpillar	n02309337	caterpillar
catfish	n02517442	catfish, siluriform_fish
cathedral	n02984203	cathedral, duomo

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
cat	n02121620	cat, true_cat
cauliflower	n07715103	cauliflower
cedar	n11623105	cedar, cedar_tree, true_cedar
celery	n07730406	celery
cellar	n02991847	cellar, wine_cellar
cello	n02992211	cello, violoncello
chain	n02999410	chain
chair	n03001627	chair
chandelier	n03005285	chandelier, pendant, pendent
chapel	n03007130	chapel
cheese	n07850329	cheese
cheetah	n02130308	cheetah, chetah, Acinonyx_jubatus
cherry	n07757132	cherry
chickadee	n01592084	chickadee
chicken	n01791625	chicken, Gallus_gallus
chimp	n02481823	chimpanzee, chimp, Pan_troglodytes
chipmunk	n02360282	chipmunk
chisel	n03020692	chisel
church	n03028079	church, church_building
clam	n01956481	clam
clamp	n03036866	clamp, clinch
clarinet	n03037709	clarinet
cloak	n03045337	cloak
clock	n03046257	clock
closet	n04550184	wardrobe, closet, press
coat	n03057021	coat
cockroach	n02233338	cockroach, roach
coconut	n07772935	coconut, cocoanut
cod	n02522399	cod, codfish
colander	n03066849	colander, cullender
comb	n03074855	comb
cork	n03108853	cork, bottle_cork
corkscrew	n03109150	corkscrew, bottle_screw
corn	n12144580	corn
cottage	n02919792	bungalow, cottage
couch	n04256520	sofa, couch, lounge
cougar	n02125311	cougar, puma, catamount, mountain_lion, painter, panther, Felis_concolor
cow	n01887787	cow
coyote	n02114855	coyote, prairie_wolf, brush_wolf, Canis_latrans
crab	n01976957	crab
cranberry	n07743902	cranberry
crane_(machine)	n03126707	crane
crayon	n03128248	crayon, wax_crayon
crocodile	n01697178	crocodile

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
crossbow	n03136369	crossbow
crowbar	n03138344	crowbar, wrecking_bar, pry, pry_bar
crow	n01579028	crow
crown	n03138669	crown, diadem
cucumber	n07718472	cucumber, cuke
cupboard	n03148324	cupboard, closet
cup	n03147509	cup
curtains	n03151077	curtain, drape, drapery, mantle, pall
cushion	n04198797	shock_absorber, shock, cushion
dagger	n03158885	dagger, sticker
dandelion	n12024176	dandelion, blowball
deer	n02430045	deer, cervid
desk	n03179701	desk
dish	n03206908	dish
dishwasher	n03207941	dishwasher, dish_washer, dishwashing_machine
dog	n02084071	dog, domestic_dog, Canis_familiaris
doll	n03219135	doll, dolly
dolphin	n02068974	dolphin
donkey	n02389559	domestic_ass, donkey, Equus_asinus
door	n03221720	door
dove	n01812337	dove
drapes	n03151077	curtain, drape, drapery, mantle, pall
dresser	n03015254	chest_of_drawers, chest, bureau, dresser
dress	n03236735	dress, frock
drill	n03239726	drill
drum	n03249569	drum, membranophone, tympan
duck	n01846331	duck
dunebuggy	n03256788	dune_buggy, beach_buggy
eagle	n01613294	eagle, bird_of_Jove
earmuffs	n03261603	earmuff
eel	n01444339	electric_eel, Electrophorus_electric
eggplant	n07713074	eggplant, aubergine, mad_apple
elephant	n02503517	elephant
elevator	n03281145	elevator, lift
elk	n02431785	wapiti, elk, American_elk, Cervus_elaphus_canadensis
emu	n01519873	emu, Dromaius_novaehollandiae, Emu_novaehollandiae
envelope	n03291819	envelope
falcon	n01610955	falcon
fan_(appliance)	n03320046	fan
faucet	n04559451	water_faucet, water_tap, tap, hydrant
fawn	n02430830	fawn
fence	n03327234	fence, fencing
finch	n01529672	finch

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
flamingo	n02007558	flamingo
flea	n02186153	flea
flute	n03372029	flute, transverse_flute
football	n03378765	football
fork	n03384167	fork
fox	n02118333	fox
freezer	n03170635	deep-freeze, Deepfreeze, deep_freezer, freezer
fridge	n03273913	electric_refrigerator, fridge
frog	n01640846	true_frog, ranid
garage	n03416489	garage
garlic	n07818277	garlic, ail
gate	n03427296	gate
giraffe	n02439033	giraffe, camelopard, Giraffa_camelopardalis
gloves	n03441112	glove
goat	n02416519	goat, caprine_animal
goldfish	n01443537	goldfish, Carassius_auratus
goose	n01855672	goose
gopher	n02358091	ground_squirrel, gopher, spermophile
gorilla	n02480855	gorilla, Gorilla_gorilla
gown	n03450230	gown
grapefruit	n07749969	grapefruit
grape	n07758680	grape
grasshopper	n02226429	grasshopper, hopper
grater	n03454885	grater
grenade	n03458271	grenade
groundhog	n02361587	groundhog, woodchuck, Marmota_monax
guitar	n03467517	guitar
gun	n03467984	gun
guppy	n01448594	guppy, rainbow_fish, Lebistes_reticulatus
hammer	n03481172	hammer
hamster	n02342885	hamster
hare	n02326432	hare
harmonica	n03494278	harmonica, mouth_organ, harp, mouth_harp
harp	n03495258	harp
harpoon	n03495671	harpoon
harpsichord	n03496296	harpsichord, cembalo
hatchet	n04449966	tomahawk, hatchet
hawk	n01605630	hawk
helicopter	n03512147	helicopter, chopper, whirlybird, eggbeater
helmet	n03513137	helmet
hoe	n03524574	hoe
honeydew	n07756325	honeydew, honeydew_melon
hook	n03532342	hook
hornet	n02213107	hornet

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
horse	n02374451	horse, Equus_caballus
hose_(leggings)	n03540267	hosiery, hose
hose	n03540090	hose
housefly	n02190790	housefly, house_fly, Musca_domestica
house	n03544360	house
hut	n03547054	hovel, hut, hutch, shack, shanty
hyena	n02117135	hyena, hyaena
iguana	n01677366	common_iguana, iguana, Iguana_iguana
inn	n03541696	hostel, hostelry, inn, lodge, auberge
jacket	n03590306	jacket
jar	n03593526	jar
jeans	n03594734	jean, blue_jean, denim
jeep	n03594945	jeep, landrover
jet	n03595860	jet, jet_plane, jet-propelled_plane
kettle	n03612814	kettle, boiler
keyboard_(musical)	n03614532	keyboard_instrument
key	n03613294	key
kite	n04284869	sport_kite, stunt_kite
knife	n03624134	knife
ladle	n03633091	ladle
lamb	n02412440	lamb
lamp	n03636248	lamp
lantern	n03640988	lantern
lemon	n07749582	lemon
leopard	n02128385	leopard, Panthera_pardus
lettuce	n07723559	lettuce
level	n03658858	level, spirit_level
lime	n07749731	lime
limousine	n03670208	limousine, limo
lion	n02129165	lion, king_of_beasts, Panthera_leo
lobster	n01982650	lobster
machete	n03699591	machete, matchet, panga
mackerel	n02624167	mackerel
magazine	n06595351	magazine, mag
mandarin	n07747951	mandarin, mandarin_orange
marble	n03721047	marble
mat	n03727837	mat
menu	n07565083	menu
microscope	n03760671	microscope
microwave	n03761084	microwave, microwave_oven
mink_(coat)	n03770954	mink, mink_coat
mink	n02442845	mink
minnow	n01442972	minnow, Phoxinus_phoxinus
mirror	n03773035	mirror

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
missile	n04008634	projectile, missile
mitten	n03775071	mitten
mixer	n03775199	mixer
mole_(animal)	n01889520	mole
moose	n02432983	elk, European_elk, moose, Alces_alces
moth	n02283201	moth
motorcycle	n03790512	motorcycle, bike
mouse_(computer)	n03793489	mouse, computer_mouse
mouse	n02330245	mouse
mug	n03797390	mug
mushroom	n07734744	mushroom
muzzle	n03803284	muzzle
napkin	n03201895	dinner_napkin
necklace	n03814906	necklace
nectarine	n07751148	nectarine
nightgown	n03824381	nightgown, gown, nightie, night-robe, night-dress
nightingale	n01560105	nightingale, Luscinia_megarhynchos
nylons	n03836976	nylons, nylon_stocking, rayons, rayon_stocking, silk_stocking
oak	n12268246	oak, oak_tree
octopus	n01970164	octopus, devilfish
olive	n12301445	olive
onions	n07722217	onion
orange	n07747607	orange
oriole	n01575745	Old_World_oriole, oriole
ostrich	n01518878	ostrich, Struthio_camelus
otter	n02444819	otter
oven	n03862676	oven
owl	n01621127	owl, bird_of_Minerva, bird_of_night, hooter
ox	n02403003	ox
paintbrush	n03876231	paintbrush
pajamas	n03877472	pajama, pyjama, pj's, jammies
pan	n03880323	pan
panther	n02128925	jaguar, panther, Panthera_onca, Felis_onca
pants	n02854739	bloomers, pants, drawers, knickers
parakeet	n01821203	parakeet, parrakeet, parroket, paraquet, paroquet, parroquet
parka	n03891051	parka, windbreaker, windcheater, anorak
parsley	n07819896	parsley
partridge	n01797886	ruffed_grouse, partridge, Bonasa_umbellus
peach	n07751004	peach
peacock	n01806143	peacock
pear	n07767847	pear
peas	n07725531	green_pea, garden_pea

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
peg	n03051249	clothespin, clothes_pin, clothes_peg
pelican	n02051845	pelican
pencil	n03908204	pencil
penguin	n02055803	penguin
pen	n03906997	pen
pepper	n07815956	white_pepper
pepper	n07815839	black_pepper
perch	n02555863	perch
pheasant	n01803078	pheasant
piano	n03928116	piano, pianoforte, forte-piano
pickle	n07824988	pickle
pier	n03934042	pier
pigeon	n01811909	pigeon
pig	n02395406	hog, pig, grunter, squealer, Sus_scrofa
pillow	n03938244	pillow
pineapple	n07753275	pineapple, ananas
pine	n11608250	pine, pine_tree, true_pine
pin	n03940256	pin
pipe_(plumbing)	n03206158	discharge_pipe
pistol	n03948459	pistol, handgun, side_arm, shooting_iron
plate	n03960490	plate
platypus	n01873310	platypus, duckbill, duckbilled_platypus, duck-billed_platypus, Ornithorhynchus_anatinus
pliers	n03966976	pliers, pair_of_pliers, plyers
plug_(electric)	n03968293	plug, male_plug
plum	n07751451	plum
pony	n02382437	pony
porcupine	n02348173	Canada_porcupine, Erethizon_dorsatum
potato	n07710616	potato, white_potato, Irish_potato, murphy, spud, tater
pot	n03990474	pot
projector	n04009552	projector
pumpkin	n07735510	pumpkin
pyramid	n13917690	truncated_pyramid
python	n01743605	python
rabbit	n02324045	rabbit, coney, cony
raccoon	n02508021	raccoon, racoon
racquet	n04039381	racket, racquet
radio	n06277135	radio, radiocommunication, wireless
radish	n07735687	radish
raft	n04045397	raft
raisin	n07752664	raisin
rake	n04050066	rake
raspberry	n07745466	raspberry
rat	n02331046	rat

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).



Concept	Synset ID	Words denoting synset
rattlesnake	n01754876	rattlesnake, rattler
raven	n01579260	raven, Corvus_corax
razor	n04057047	razor
revolver	n04086273	revolver, six-gun, six-shooter
rhubarb	n07713267	pieplant, rhubarb
rice	n07804323	rice
rifle	n04090263	rifle
robe	n04097866	robe
robin	n01558993	robin, American_robin, Turdus_migratorius
rocker	n04098513	rocker
rocket	n04099175	rocket, rocket_engine
rock	n09416076	rock, stone
rooster	n01792158	cock, rooster
rope	n04108268	rope
ruler	n04118776	rule, ruler
sack	n04122825	sack, poke, paper_bag, carrier_bag
saddle	n04123740	saddle
sailboat	n04128499	sailboat, sailing_boat
salamander	n01629276	salamander
salmon	n02534734	salmon
sandals	n04133789	sandal
sardine	n02533209	pilchard, sardine, Sardina_pilchardus
saxophone	n04141076	sax, saxophone
scarf	n04143897	scarf
scissors	n04148054	scissors, pair_of_scissors
scooter	n03791053	motor_scooter, scooter
screwdriver	n04154565	screwdriver
screws	n04153751	screw
seagull	n02041246	gull, seagull, sea_gull
seal	n02076196	seal
shack	n03547054	hovel, hut, hutch, shack, shanty
shawl	n04186455	shawl
shed	n04187547	shed
sheep	n02411705	sheep
shelves	n04190052	shelf
shield	n04192698	shield, buckler
ship	n04194289	ship
shirt	n04197391	shirt
shoes	n04199027	shoe
shotgun	n04206356	shotgun, scattergun
shovel	n04208210	shovel
shrimp	n01986806	shrimp
sink	n02998563	cesspool, cesspit, sink, sump
skateboard	n04225987	skateboard

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
skillet	n03400231	frying_pan, frypan, skillet
skirt	n04230808	skirt
skis	n04228054	ski
skunk	n02445715	skunk, polecat, wood_pussy
skyscraper	n04233124	skyscraper
sledgehammer	n03731695	maul, sledge, sledgehammer
sled	n04235291	sled, sledge, sleigh
sleigh	n04235291	sled, sledge, sleigh
slippers	n04241394	slipper, carpet_slipper
snail	n01944390	snail
socks	n04254777	sock
sofa	n04256520	sofa, couch, lounge
spade	n04266486	spade
sparrow	n01527347	hedge_sparrow, sparrow, dunnock, Prunella_modularis
spatula	n04269944	spatula
spear	n04270891	spear, lance, shaft
spider	n01772222	spider
spinach	n07736692	spinach
spoon	n04284002	spoon
squid	n01971280	squid
squirrel	n02355227	squirrel
starling	n01576695	starling
stereo	n04315948	stereo, stereophony, stereo_system, stereo-phonic_system
stick	n04317420	stick
stone	n09416076	rock, stone
stool_(furniture)	n04326896	stool
stork	n02002075	stork
stove	n04330340	stove, kitchen_stove, range, kitchen_range, cooking_stove
strainer	n04332243	strainer
strawberry	n07745940	strawberry
submarine	n04347754	submarine, pigboat, sub, U-boat
subway	n04349306	subway_train
swan	n01858441	swan
sweater	n04370048	sweater, jumper
swimsuit	n04371563	swimsuit, swimwear, bathing_suit, swimming_costume, bathing_costume
sword	n04373894	sword, blade, brand, steel
table	n04379243	table
tack	n04383130	tack
tangerine	n07748416	tangerine
tank_(army)	n04389033	tank, army_tank, armored_combat_vehicle, armoured_combat_vehicle

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
tank_(container)	n04388743	tank, storage_tank
tape_(scotch)	n02992795	cellulose_tape, Scotch_tape, Sellotape
tap	n04559451	water_faucet, water_tap, tap, hydrant
taxi	n02930766	cab, hack, taxi, taxicab
telephone	n04401088	telephone, phone, telephone_set
tent	n04411264	tent, collapsible_shelter
thermometer	n03043423	clinical_thermometer, mercury-in-glass_clinical_thermometer
thimble	n04423845	thimble
tie	n03815615	necktie, tie
tiger	n02129604	tiger, Panthera_tigris
toad	n01645776	true_toad
toaster	n04442312	toaster
toilet	n04446276	toilet, lavatory, lav, can, john, privy, bathroom
tomahawk	n04449966	tomahawk, hatchet
tomato	n07734017	tomato
tongs	n04450749	tongs, pair_of_tongs
tortoise	n01670092	tortoise
toy	n04461879	toy
tractor	n04465501	tractor
trailer	n04467099	trailer, house_trailer
train	n04468005	train, railroad_train
tray	n04476259	tray
tricycle	n04482393	tricycle, trike, velocipede
tripod	n04485082	tripod
trolley	n04397027	tea_cart, teacart, tea_trolley, tea_wagon
trombone	n04487394	trombone
trousers	n03688605	long_trousers, long_pants
trout	n02537085	trout
truck	n04490091	truck, motortruck
trumpet	n03110669	cornet, horn, trumpet, trump
tuba	n02804252	bass_horn, sousaphone, tuba
tuna	n02626762	tuna, tunny
turkey	n01794158	turkey, Meleagris_gallopavo
turnip	n07735803	turnip
turtle	n01663401	sea_turtle, marine_turtle
typewriter	n04505036	typewriter
umbrella	n04507155	umbrella
unicycle	n04509417	unicycle, monocycle
van	n04520170	van
veil	n03502331	head_covering, veil
vest	n04531873	vest, waistcoat
vine	n13100677	vine
violin	n04536866	violin, fiddle
vulture	n01616318	vulture

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets (continues on next page).

Concept	Synset ID	Words denoting synset
wagon	n04543158	wagon, waggon
wall	n03252637	dry_wall, dry-stone_wall
walnut	n07771212	walnut
walrus	n02081571	walrus, seahorse, sea_horse
wand	n04549629	wand
wasp	n02212062	wasp
whale	n02062744	whale
wheelbarrow	n02797295	barrow, garden_cart, lawn_cart, wheelbarrow
wheel	n04574999	wheel
whip	n04577769	whip
whistle	n04579667	whistle
willow	n12724942	willow, willow_tree
woodpecker	n01838598	woodpecker, peckerwood, pecker
worm	n01922303	worm
wrench	n04606574	wrench, spanner
yacht	n04610013	yacht, racing_yacht
yam	n07712267	yam
zebra	n02391049	zebra
zucchini	n07716358	zucchini, courgette

Table A.1: List of concepts, their WordNet synset IDs, and the corresponding synsets.

## A.2 Annotation interface

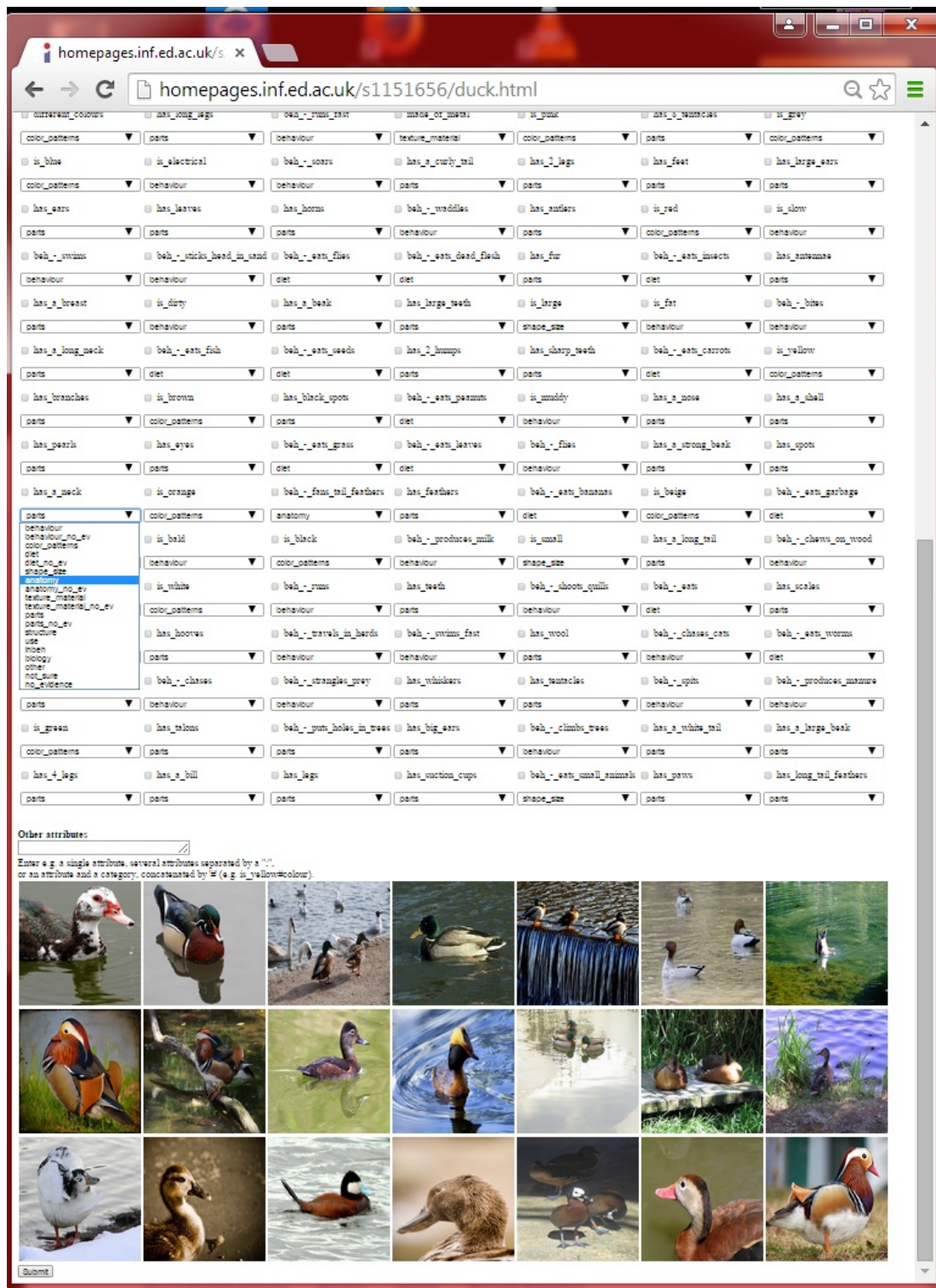


Figure A.1: Interface for the annotation of the concept *duck* with its visual attributes on the basis of example images. The list of visual attributes only contains the McRae attributes listed for at least one concept of the currently processed category. For example, the current list contains only the attributes listed for the concepts of the category BIRD, which excludes, for example, the attributes `made_of_wood` or `has_leaves`.

# Appendix B

## Instructions to Participants in Word Similarity Study

In this experiment you will be presented with a series of word pairs. The words in each pair are separated by a dash ('—'). Your task is to rate the pair along two dimensions. First, you will rate the **degree** to which the two words have the same **meaning (semantic similarity)**. And then, you will rate the **degree** to which the objects the words refer to **look the same (visual similarity)**. You will make these judgements by choosing a rating from **1 (highly dissimilar)** to **5 (highly similar)**.

### Rating Examples

Before taking part in the experiment, please read carefully the examples below. They illustrate word pairs with varying degrees of similarity in appearance and meaning and provide explanations for the provided ratings.

Word Pair	Rating Semantic;Visual	Explanation
butter — cheese:	3;3	<i>Butter</i> and <i>cheese</i> are both types of dairy products. They are produced and used differently. They might be rated as vaguely similar. Some types of <i>cheese</i> can have a similar appearance as <i>butter</i> , so they could be rated as visually vaguely similar.
bathroom — towel:	2;1	The words <i>bathroom</i> and <i>towel</i> are somehow related, as <i>towels</i> can be found in bathrooms. However, their meanings don't have anything in common. <i>Towels</i> and <i>bathrooms</i> don't have a similar appearance either.

aquarium -- cage:	3;2	<i>Aquariums</i> and <i>cages</i> are vaguely similar as they have a loosely common function (they are both enclosures in which usually animals are kept). <i>Aquariums</i> are made of glass and <i>cages</i> are often made of bars or wires and can be of different shapes. Visually they are thus different.
aquarium -- water:	2;2	<i>Aquariums</i> are filled with <i>water</i> , so water can be seen as being part of an aquarium, which makes the two words somehow related, but yet different in meaning. <i>Water</i> is a visually salient part of an <i>aquarium</i> , but nevertheless it looks different.
semolina -- sand:	1;5	<i>Semolina</i> is food whereas <i>sand</i> consists of rock and mineral particles. As the meanings of the words don't have anything in common, they would be rated as highly dissimilar. Visually, however, <i>semolina</i> and <i>sand</i> look almost identical, so they would be rated high in terms of their appearance.

### Notes

Some words are ambiguous, i.e., they have more than one meaning. In such cases, we will disambiguate the word for you, indicating the meaning we are interested in within parentheses. For example, *mouse (computer)* is an electronic device, whereas *mouse (animal)* is a rodent. Please make your similarity judgements with respect to the provided meaning.

Please do not forget to accept the HIT before you begin to work on it.

# Glossary

## **AMT**

Amazon Mechanical Turk 4, 9, 60, 97, 99, 127

## **AP**

interpolated average precision 67

## **attribute**

Property of a concept xi, xiii, 2, 4, 8, 11–13, 15, 16, 19, 20, 22, 23, 25, 30–32, 34–36, 39, 40, 45, 47–49, 51, 52, 54, 60, 63, 69–74, 97

## **BNC**

British National Corpus 15–17, 40

## **BoVW**

bag-of-visual-words 24, 65, 137, 138

## **CCA**

Canonical correlation analysis. A data analysis and dimensionality reduction method (Hotelling, 1936). 29, 38

## **CNN**

Convolutional neural network 24, 57, 84, 137

## **CW**

Efficient graph clustering algorithm that clusters undirected, weighted graphs (Biemann, 2006). 106–108

## **HOG**

Histogram of Oriented Gradients. Feature descriptors used in computer vision and image processing generated by counting occurrences of gradient orientation in spatial regions of the image (Dalal and Triggs, 2005). 56, 66

## **ILSVRC**

The ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2014) 114



**kCCA**

Kernel canonical correlation analysis. A kernelised data analysis and dimensionality reduction method (Hardoon et al., 2004). 38, 44, 45, 48, 74–77, 79, 80, 94, 101, 103, 108, 112

**LAB**

Uniform colour scale recommended by the International Commission on Illumination (CIE) in 1976 (see, e.g., ?). 65

**McRae norms**

The semantic attribute production norms by McRae et al. (2005). 30, 31, 40, 41, 43, 45, 47, 48, 51, 60–64, 70, 75, 76, 78, 97–99, 103, 105, 109, 114

**PoS**

part-of-speech 12, 22, 59, 71, 72, 74, 83

**SAE**

bimodal stacked autoencoder xi, xiv, 69, 71, 91, 93, 95, 96, 99, 100, 103, 104, 106–111, 113, 114, 116, 117, 119, 122, 124, 128–131, 135

**SIFT**

An approach to extract descriptions of visual image features that are invariant to scaling and rotation and partially to changes in illumination and affine transformations (Lowe, 2004). 24, 56, 60, 101, 102

**StD**

standard deviation 99

**SVD**

Singular value decomposition. A mathematical standard technique for reducing the dimensionality of vector spaces. 14, 17–19, 22, 26, 94, 101, 103, 108, 112

**SVM**

support vector machine 64, 65, 137

**VSM**

vector space model 13, 14, 16, 18, 26

# Bibliography

- Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado. 7, 41
- Agirre, E. and Soroa, A. (2007). SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 7–12, Prague, Czech Republic. 107
- Anderson, J. R. (1991). The Adaptive Nature of Human Categorization. *Psychological Review*, 98(3):409–429. 105
- Andrews, M., Vigliocco, G., and Vinson, D. (2009). Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review*, 116(3):463–498. xi, 6, 23, 26, 27, 29, 34, 36, 39, 41, 42, 51, 76, 95
- Arel, I., Rose, D., and Karnowski, T. P. (2010). Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]. *IEEE Computational Intelligence Magazine*, 5(4):13–18. 81
- Ashby, F. G. and Maddox, W. T. (2011). Human Category Learning 2.0. *Annals of the New York Academy of Sciences*, 1224(1):147–161. 122
- Austerweil, J. L. and Griffiths, T. L. (2010). Learning Hypothesis Spaces and Dimensions through Concept Learning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 73–78, Austin, Texas. 105
- Bansal, M., Gimpel, K., and Livescu, K. (2014). Tailoring Continuous Word Representations for Dependency Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 809–815. 22
- Barbu, E. (2008). Combining Methods to Learn Feature-norm-like Concept Descriptions. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 9–16, Hamburg, Germany. 69
- Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching Words and Pictures. *The Journal of Machine Learning Research*, 3:1107–1135. 27

- Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't Count, Predict! A Systematic Comparison of Context-counting vs. Context-predicting Semantic Vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, Baltimore, Maryland. 23
- Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2010). Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science*, 34(2):222–254. <http://clic.cimec.unitn.it/strudel/>. 6, 9, 15, 16, 54, 70, 71, 73, 75, 100, 106, 111, 118, 119, 127, 138
- Barsalou, L. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences*, 22:577–609. i, 2, 23
- Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, 59:617–845. 2, 23
- Bassok, M. (1990). Transfer of Domain-specific Problem-solving Procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(3):522–533. 6
- Behl-Chadha, G. (1996). Basic-level and Superordinate-like Categorical Representations in Early Infancy. *Cognition*, 60(2):105–141. 121
- Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127. 87, 90
- Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828. 82, 83
- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155. 2, 21
- Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2006). Greedy Layer-Wise Training of Deep Networks. In *Advances in Neural Information Processing Systems 19*, pages 153–160, Vancouver, Canada. 89
- Bergsma, S. and Goebel, R. (2011). Using Visual Information to Predict Lexical Preference. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 399–405, Hissar, Bulgaria. 27
- Bergsma, S. and Van Durme, B. (2011). Learning Bilingual Lexicons using the Visual Similarity of Labeled Web Images. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1764–1769, Barcelona, Spain. 27
- Bespalov, D., Bai, B., Qi, Y., and Shokoufandeh, A. (2011). Sentiment classification based on supervised latent n-gram analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 375–382. ACM. 22

- Biemann, C. (2006). Chinese Whispers – an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In *Proceedings of TextGraphs: the 1st Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80, New York City, NY. 106, 156
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022. 19, 29, 34
- Borga, M. (2001). Canonical Correlation - a Tutorial. 38
- Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., and Pascual, L. (2004). Cross-linguistic Analysis of Vocabulary in Young Children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development*, 75(4):1115–1139. i, 2, 23
- Bruni, E., Boleda, G., Baroni, M., and Tran, N. (2012a). Distributional Semantics in Technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 136–145, Jeju Island, Korea. 24, 25
- Bruni, E., Bordignon, U., Liska, A., Uijlings, J., and Sergienya, I. (2013). VSEM: An open library for visual semantics representation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–192, Sofia, Bulgaria. 101
- Bruni, E., Tran, G., and Baroni, M. (2011). Distributional Semantics from Text and Images. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 22–32, Edinburgh, UK. 24, 25
- Bruni, E., Tran, N., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49:1–47. 18, 24, 26, 27, 94, 101, 102, 105
- Bruni, E., Uijlings, J., Baroni, M., and Sebe, N. (2012b). Distributional Semantics with Eyes: Using Image Analysis to Improve Computational Representations of Word Meaning. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1219–1228, Nara, Japan. 24, 26, 138
- Bullinaria, J. A. and Levy, J. P. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39(3):510–526. 17
- Bullinaria, J. A. and Levy, J. P. (2012). Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming, and SVD. *Behavior Research Methods*, 44(3):890–907. 2, 14, 15, 18
- Chen, H., Gallagher, A., and Girod, B. (2012). Describing Clothing by Semantic Attributes. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, pages 609–623, Florence, Italy. <http://purl.stanford.edu/tb980qz1002>. 58

- Cireşan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column Deep Neural Networks for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3642–3649, Providence, RI, USA. 84
- Cohen, J. and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ. 44
- Collins, A. M. and Loftus, E. F. (1975). A Spreading-activation Theory of Semantic Processing. *Psychological Review*, 82(6):407. 12
- Collobert, R. and Weston, J. (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, Helsinki, Finland. 21, 22, 83
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537. 1, 22, 82, 83
- Cree, G. S. and Armstrong, B. C. (2012). Computational Models of Semantic Memory. In Spivey, M., McRae, K., and Joanisse, M., editors, *Cambridge Handbook of Psycholinguistics*, pages 259–282. Cambridge University Press. 20
- Cree, G. S., McRae, K., and McNorgan, C. (1999). An Attractor Model of Lexical Conceptual Processing: Simulating Semantic Priming. *Cognitive Science*, 23(3):371–414. 2, 12, 20
- Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 886–893, San Diego, CA. 56, 66, 156
- Dale, R., Somers, H. L., and Moisl, H., editors (2000). *Handbook of Natural Language Processing*. Marcel Dekker, Inc., New York, NY, USA. 4
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407. 14, 17
- Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, Florida. <http://www.image-net.org>. 4, 24, 32, 57, 58, 59, 60, 61, 115
- Deng, L. (2014). A Tutorial Survey of Architectures, Algorithms, and Applications for Deep Learning. *APSIPA Transactions on Signal and Information Processing*. 83
- Devereux, B., Pilkington, N., Poibeau, T., and Korhonen, A. (2009). Towards Unrestricted, Large-Scale Acquisition of Feature-Based Conceptual Representations from Corpus Data. *Research on Language and Computation*, 7(2-4):137–170. 70

- Devereux, B. J., Tyler, L. K., Geertzen, J., and Randall, B. (2013). The Centre for Speech, Language and the Brain (CSLB) Concept Property Norms. *Behavior Research Methods*. <http://csl.psychol.cam.ac.uk/propertynorms/>. 10, 12, 13, 31, 114, 134
- Dhall, A., Goecke, R., Joshi, J., Wagner, M., and Gedeon, T. (2013). Emotion Recognition in the Wild Challenge (EmotiW) Challenge and Workshop Summary. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 371–372, Sydney, Australia. 84
- Duan, K., Parikh, D., Crandall, D., and Grauman, K. (2012). Discovering Localized Attributes for Fine-grained Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474–3481, Washington, DC, USA. 58
- Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74. 73
- Erk, K. and Padó, S. (2008). A Structured Vector Space Model for Word Meaning in Context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii. 13, 15
- Escorcia, V., Carlos Niebles, J., and Ghanem, B. (2015). On the Relationship between Visual Attributes and Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1256–1264, Boston, MA. 137
- Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2014). The PASCAL Visual Object Classes Challenge - a Retrospective. *International Journal of Computer Vision*. accepted. 120, 129
- Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2012). The PASCAL Visual Object Classes Challenge 2012 Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop>. 32
- Everingham, M. and Van Gool, L. and Williams, C. K. I. and Winn, J. and Zisserman, A. (2008). The PASCAL Visual Object Classes Challenge 2008 Results. <http://www.pascal-network.org/challenges/VOC/voc2008/workshop>. 57
- Fan, R., Chang, K., Hsieh, C., Wang, X., and Lin, C. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874. 64
- Farah, M. J. and McClelland, J. L. (1991). A Computational Model of Semantic Memory Impairment: Modality Specificity and Emergent Category Specificity. *JOURNAL OF EXPERIMENTAL PSYCHOLOGY: GENERAL*, 120(4):339–357. 20
- Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). Describing Objects by their Attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, Miami Beach, Florida. <http://vision.cs.uiuc.edu/attributes/>. 57, 58, 62, 64, 65

- Fei-Fei, L., Fergus, R., and Perona, P. (2007). Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *Computer Vision and Image Understanding*, 106(1):59–70. 32
- Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press. 58, 59, 70, 98, 125, 128
- Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010). Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645. 56
- Feng, F., Li, R., and Wang, X. (2013). Constructing Hierarchical Image-tags Bimodal Representations for Word Tags Alternative Choice. In *Proceedings of the ICML 2013 Workshop on Challenges in Representation Learning*, Atlanta, Georgia. 85, 86
- Feng, Y. and Lapata, M. (2010). Visual Information in Semantic Representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99, Los Angeles, California. 24, 26, 27
- Feng, Y. and Lapata, M. (2013). Automatic Caption Generation for News Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812. 27
- Ferrari, V. and Zisserman, A. (2007). Learning Visual Attributes. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 433–440, Cambridge, Massachusetts. MIT Press. 57
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116–131. xiii, 7, 41, 46, 97
- Fountain, T. and Lapata, M. (2010). Meaning Representation in Natural Language Categorization. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 1916–1921, Amsterdam, The Netherlands. 105
- Fountain, T. and Lapata, M. (2011). Incremental Models of Natural Language Category Acquisition. In Carlson, C., Hölscher, and Shipley, T., editors, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Austin, Texas. 105
- French, R. M., Mareschal, D., Mermillod, M., and Quinn, P. C. (2004). The Role of Bottom-Up Processing in Perceptual Categorization by 3-to 4-Month-Old Infants: Simulations and Data. *Journal of Experimental Psychology: General*, 133(3):382–397. 88
- Frermann, L. and Lapata, M. (2014). Incremental Bayesian Learning of Semantic Categories. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–258, Gothenburg, Sweden. 105

- Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013). DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems 26*, pages 2121–2129. 27, 84, 86, 120
- Glenberg, A. M. and Kaschak, M. P. (2002). Grounding Language in Action. *Psychonomic Bulletin and Review*, 9(3):558–565. i, 2, 23
- Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning*, pages 513–520, Bellevue, Washington, USA. 83
- Goldstone, R. L., Kersten, A., and Cavalho, P. F. (2012). Concepts and Categorization. In Healy, A. F. and Proctor, R. W., editors, *Comprehensive Handbook of Psychology*, volume 4: Experimental psychology, pages 607–630. Wiley, New Jersey. 105
- Goldstone, R. L. and Son, J. Y. (2005). Similarity. In Holyoak, K. J. and Morrison, R. G., editors, *The Cambridge Handbook of Thinking and Reasoning*, pages 13–36. Cambridge University Press. 6
- Goller, C. and Küchler, A. (1996). Learning Task-Dependent Distributed Representations by Backpropagation Through Structure. In *In Proceedings of the 1996 IEEE International Conference on Neural Networks (Volume:1)*, pages 347–352, Washington, DC. 83
- Golub, G. and Reinsch, C. (1970). Singular Value Decomposition and Least Squares Solutions. *Numerische Mathematik*, 14(5):403–420. 14, 17, 101
- Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., and Lazebnik, S. (2014). Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. In *Proceedings of the 13th European Conference on Computer Vision*, pages 529–545, Zurich, Switzerland. 27
- Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. (2013). Challenges in Representation Learning: A Report on Three Machine Learning Contests. In Lee, M., Hirose, A., Hou, Z.-G., and Kil, R., editors, *Neural Information Processing*, volume 8228 of *Lecture Notes in Computer Science*, pages 117–124. Springer Berlin Heidelberg. 84, 86
- Graves, A., Mohamed, A.-R., and Hinton, G. (2013). Speech Recognition with Deep recurrent Neural Networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, Vancouver, Canada. 84
- Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA. 6, 14



- Griffin, G., Holub, A., and Perona, P. (2007). Caltech-256 Object Category Dataset. Technical report, California Institute of Technology. [http://www.vision.caltech.edu/Image\\_Datasets/Caltech256/](http://www.vision.caltech.edu/Image_Datasets/Caltech256/). 32
- Griffin, L. D., Wahab, M. H., and Newell, A. J. (2013). Distributional Learning of Appearance. *PLoS ONE*, 8(2):1–15. 95
- Griffiths, T. L., Canini, K. R., Sanborn, A. N., and Navarro, D. J. (2007a). Unifying Rational Models of Categorization via the Hierarchical Dirichlet Process. In *Proceedings of the Annual Conference of the Cognitive Science Society*, volume 29, pages 323–328, Nashville, Tennessee. 105
- Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007b). Topics in Semantic Representation. *Psychological Review*, 114(2):211–244. 14, 41, 42, 76
- Grondin, R., Lupker, S., and Mcrae, K. (2009). Shared Features Dominate Semantic Richness Effects for Concrete Concepts. *Journal of Memory and Language*, 60(1):1–19. 13
- Hardoon, D. R., Szedmak, S. R., and Shawe-Taylor, J. R. (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664. 29, 38, 51, 74, 101, 157
- Harris, Z. S. (1954). Distributional Structure. *Word*, 10(2–3):146–162. i, 2, 13
- Hill, F. and Korhonen, A. (2014). Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 255–265, Doha, Qatar. 24, 26
- Hinton, G., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554. 20, 81
- Hinton, G. E. (1981). Implementing Semantic Networks in Parallel Hardware. In Hinton, G. E. and Anderson, J. A., editors, *Parallel Models of Associative Memory*, pages 161–187. Erlbaum, Hillsdale, NJ. 19, 20
- Hinton, G. E. (1986). Learning Distributed Representations of Concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 1–12, Amherst, Massachusetts. 19, 20
- Hinton, G. E. (2007). Learning Multiple Layers of Representation. *Trends in Cognitive Sciences*, 11(10):428–434. 81
- Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507. 20, 81, 88, 90
- Hinton, G. E. and Shallice, T. (1991). Lesioning an Attractor Network: Investigations of Acquired Dyslexia. *Psychological Review*, 98:74–95. 20

- Hintzman, D. L. (1986). "Schema Abstraction" in a Multiple-Trace Memory Model. *Psychological Review*, 93(4):411–428. 6
- Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780. 84
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:312–377. 29, 38, 156
- Hsu, A. S., Martin, J. B., Sanborn, A. N., and Griffiths, T. L. (2012). Identifying Representations of Categories of Discrete Items Using Markov Chain Monte Carlo with People. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 485–490, Sapporo, Japan. 105
- Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 873–882, Jeju Island, Korea. 21, 22
- Huang, J. and Kingsbury, B. (2013). Audio-visual Deep Learning for Noise Robust Speech Recognition. In *Proceedings 38th International Conference on Acoustics, Speech, and Signal Processing*, pages 7596–7599, Vancouver, Canada. 85
- Huiskes, M. J. and Lew, M. S. (2008). The MIR Flickr Retrieval Evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 39–43, Vancouver, Canada. <http://press.liacs.nl/mirflickr/>. 32, 58
- Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., and Daumé III, H. (2014). A Neural Network for Factoid Question Answering over Paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 633–644, Doha, Qatar. 22
- Japkowicz, N., Jose Hanson, S., and Gluck, M. A. (2000). Nonlinear Autoassociation Is Not Equivalent to PCA. *Neural Computation*, 12(3):531–545. 88
- Jia, Y., Abbott, J. T., Austerweil, J., Griffiths, T., and Darrell, T. (2013). Visual Concept Learning: Combining Machine Vision and Bayesian Generalization on Concept Hierarchies. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 1842–1850. Curran Associates, Inc. 10, 121, 122, 123, 124, 125, 126, 127, 129, 131
- Johns, B. T. and Jones, M. N. (2012). Perceptual Inference through Global Lexical Similarity. *Topics in Cognitive Science*, 4(1):103–120. xi, 24, 25, 27, 29, 36, 37, 38, 39, 45, 51, 52, 94
- Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer. 88
- Jones, M. N., Kintsch, W., and Mewhort, D. J. K. (2006). High-dimensional Semantic Space Accounts of Priming. *Journal of Memory and Language*, 55(4):534–552. 96

- Jones, M. N., Willits, J. A., and Dennis, S. (2015). Models of Semantic Memory. In Busemeyer, J., Townsend, J., Wang, Z., and Eidels, A., editors, *Oxford Handbook of Computational and Mathematical Psychology*, pages 232–254. Oxford University Press. 12, 20
- Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, c., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., Mirza, M., Jean, S., Carrier, P.-L., Dauphin, Y., Boulanger-Lewandowski, N., Aggarwal, A., Zumer, J., Lamblin, P., Raymond, J.-P., Desjardins, G., Pascanu, R., Warde-Farley, D., Torabi, A., Sharma, A., Bengio, E., Côté, M., Konda, K. R., and Wu, Z. (2013). Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pages 543–550, Sydney, Australia. 84
- Karpathy, A. and Fei-Fei, L. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, Boston, MA. 86
- Kelly, C., Devereux, B., and Korhonen, A. (2010). Acquiring Human-like Feature-based Conceptual Representations from Corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 61–69, Los Angeles, California. 70
- Kelly, C., Devereux, B., and Korhonen, A. (2012). Semi-supervised Learning for Automatic Conceptual Property Extraction. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, pages 11–20, Montréal, Canada. 70
- Kelly, C., Devereux, B., and Korhonen, A. (2014). Automatic Extraction of Property Norm-Like Data From Large Text Corpora. *Cognitive Science*, 38(4):638–682. 70
- Kelly, C., Korhonen, A., and Devereux, B. (2013). Minimally Supervised Learning for Unconstrained Conceptual Property Extraction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 746–751, Berlin, Germany. 70
- Kiela, D. and Bottou, L. (2014). Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 36–45, Doha, Qatar. 24, 25, 27, 84, 137
- Kim, Y., Lee, H., and Provost, E. M. (2013). Deep Learning for Robust Feature Generation in Audiovisual Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3687–3691, South Brisbane, Australia. 85
- Kiros, R., Salakhutdinov, R., and Zemel, R. (2014a). Multimodal Neural Language Models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 595–603, Beijing, China. 27, 84

- Kiros, R., Salakhutdinov, R., and Zemel, R. (2014b). Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In *Deep Learning and Representation Learning Workshop: NIPS 2014*, Montréal, Canada. 22, 27, 84, 86
- Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing Crosslingual Distributed Representations of Words. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1459–1474, Mumbai, India. 2
- Kolodner, J. (1993). *Case-based Reasoning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 6
- Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. Technical report, Department of Computer Science, University of Toronto. 32
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc. 84, 86, 137
- Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2013). BabyTalk: Understanding and Generating Simple Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903. 120
- Kumar, N., Belhumeur, P. N., and Nayar, S. K. (2008). FaceTracer: A Search Engine for Large Collections of Images with Faces. In *European Conference on Computer Vision (ECCV)*, pages 340–353, Marseille, France. <http://www.cs.columbia.edu/CAVE/databases/facetracer/>. 58
- Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2011). Describable Visual Attributes for Face Verification and Image Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977. 58
- Kurtz, K. J. (2007). The Divergent Autoencoder (DIVA) Model of Category Learning. *Psychonomic Bulletin & Review*, 14(4):560–576. 88
- Laffont, P.-Y., Ren, Z., Tao, X., Qian, C., and Hays, J. (2014). Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics*, 33(4):149:1–149:11. 58
- Lampert, C. H., Nickisch, H., and Harmeling, S. (2009). Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, Miami Beach, Florida. <http://attributes.kyb.tuebingen.mpg.de/>. 57, 58, 62, 64
- Landau, B., Smith, L., and Jones, S. (1998). Object Perception and Object Naming in Early Development. *Trends in Cognitive Sciences*, 2(1):19–24. i, 2, 23
- Landauer, T. and Dumais, S. T. (1997). A Solution to Plato’s Problem: the Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211–240. 1, 2, 14, 15, 17, 18, 26

- Lazaridou, A., Bruni, E., and Baroni, M. (2014). Is this a Wampimuk? Cross-modal Mapping between Distributional Semantics and the Visual World. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland. 27
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition - Volume 2*, pages 2169–2178, Washington, DC, USA. 24, 57, 66
- Le, Q. V., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., and Ng, A. Y. (2013). Building High-level Features Using Large Scale Unsupervised Learning. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8595–8598, Vancouver, Canada. 84
- Lebret, R. and Collobert, R. (2014). Rehabilitation of Count-based Models for Word Vector Representations. *Proceeding of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, 9041:417–429. 23
- Lebret, R., Legrand, J., and Collobert, R. (2013). Is Deep Learning Really Necessary for Word Embeddings? In *NIPS Workshop on Deep Learning*, Stateline, NV, USA. 23
- LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1990). Handwritten Digit Recognition with a Back-propagation Network. In *Advances in Neural Information Processing Systems 2*, NIPS 1989, pages 396–404. Morgan Kaufmann Publishers. 20, 81
- Leong, C. W. and Mihalcea, R. (2011). Going Beyond Text: A Hybrid Image-Text Approach for Measuring Word Relatedness. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1403–1407, Chiang Mai, Thailand. 27
- Levy, O. and Goldberg, Y. (2014). Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. 22
- Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225. 23
- Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting Outliers: Do not use Standard Deviation around the Mean, use Absolute Deviation around the Median. *Journal of Experimental Social Psychology*, 49(4):764–766. 117
- Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Cao, L., and Huang, T. S. (2011). Large-scale Image Classification: Fast Feature Extraction and SVM Training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1689–1696, Colorado Springs, Colorado. 123

- Liu, J., Kuipers, B., and Savarese, S. (2011). Recognizing Human Actions by Attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3337–3344, Colorado Springs, Colorado. 57
- Lowe, D. G. (2004). Distinctive Image Features from Scale-invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110. 24, 56, 60, 101, 157
- Lund, K. and Burgess, C. (1996). Producing High-dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208. 2, 13, 14, 15
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY. 2, 14, 96
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. 116
- Mao, J., Xu, W., Yang, Y., Wang, J., and Yuille, A. L. (2014). Explain Images with Multimodal Recurrent Neural Networks. In *Deep Learning and Representation Learning Workshop: NIPS 2014*, Montréal, Canada. 27, 84, 86
- Mareschal, D., French, R. M., and Quinn, P. C. (2000). A Connectionist Account of Asymmetric Category Learning in Early Infancy. *Developmental Psychology*, 36(5):635–645. 88
- McCloskey, M. and Clucksberg, S. (1979). Decision Processes in Verifying Category Membership Statements: Implications for Models of Semantic Memory. *Cognitive Psychology*, 11:1–37. 109
- McKinley, S. C. and Nosofsky, R. M. (1995). Investigations of Exemplar and Decision Bound Models in Large, Ill-Defined Category Structures. *Journal of Experimental Psychology, Human Perception and Performance*, 21(1):128–48. 105
- McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic Feature Production Norms for a Large Set of Living and Nonliving Things. *Behavior Research Methods*, 37(4):547–559. xi, 2, 4, 5, 12, 13, 30, 40, 52, 60, 78, 97, 105, 138, 157
- McRae, K., de Sa, V. R., and Seidenberg, M. S. (1997). On the Nature and Scope of Featural Representations of Word Meaning. *Journal of Experimental Psychology: General*, 126(2):99–130. 20
- McRae, K. and Jones, M. (2013). Semantic Memory. In Reisberg, D., editor, *The Oxford Handbook of Cognitive Psychology*. Oxford University Press. 1, 11, 40
- Medin, D. L., Goldstone, R. L., and Markman, A. B. (1995). Comparison and Choice: Relations between Similarity Processes and Decision Processes. *Psychonomic Bulletin & Review*, 2(1):1–19. 6

- Medin, D. L. and Schaffer, M. M. (1978). Context Theory of Classification Learning. *Psychological Review*, 85(3):207–238. 6
- Mervis, C. B. and Rosch, E. (1981). Categorization of Natural Objects. *Annual Review of Psychology*, 32(1):89–115. 11, 126
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *Workshop at the International Conference on Learning Representations*. 22, 26, 102
- Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pages 1045–1048, Makuhari, Chiba, Japan. 84
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc. 22, 101, 102, 137
- Mikolov, T., tau Yih, W., and Zweig, G. (2013c). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. 22
- Miller, G. A. and Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1). 7, 98, 99
- Mitchell, J. (2011). *Composition in Distributional Models of Semantics*. PhD thesis, School of Informatics, University of Edinburgh. 17
- Mnih, A. and Hinton, G. E. (2009). A Scalable Hierarchical Distributed Language Model. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1081–1088. Curran Associates, Inc. 21
- Murphy, G. L. and Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review*, 92(3):289. 5
- Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). The University of South Florida Word Association, Rhyme, and Word Fragment Norms. <http://www.usf.edu/FreeAssociation/>. xiii, 40, 41, 49, 75, 76, 77, 78, 90, 100
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. (2011). Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 689–696, Bellevue, Washington, USA. 85, 94
- Nissim, M. and Markert, K. (2003). Syntactic Features and Word Similarity for Supervised Metonymy Resolution. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 56–63, Sapporo, Japan. 7

- Nosofsky, R. (1986). Attention, Similarity and the Identification-categorization Relationship. *Experimental Psychology: General*, 115:39–57. 6
- Novick, L. R. (1990). Representational Transfer in Problem Solving. *Psychological Science*, 1(2):128. 6
- O'Connor, C. M., Cree, G. S., and McRae, K. (2009). Conceptual Hierarchies in a Flat Attractor Network: Dynamics of Learning and Computations. *Cognitive Science*, 33(4):665–708. xiv, 13, 109, 110
- Ordonez, V., Deng, J., Choi, Y., Berg, A. C., and Berg, T. L. (2013). From Large Scale Image Categorization to Entry-Level Categories. In *IEEE International Conference on Computer Vision*, pages 2768–2775, Sydney, Australia. 121, 139
- Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., and Shafir, E. (1990). Category-based induction. *Psychological Review*, 97(2):185. 6
- Osherson, D. N., Stern, J., Wilkie, O., Stob, M., and Smith, E. E. (1991). Default Probability. *Cognitive Science*, 2(15):251–269. 57
- Paşca, M., Lin, D., Bigham, J., Lifchits, A., and Jain, A. (2006). Names and Similarities on the Web: Fact Extraction in the Fast Lane. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 809–816, Sydney, Australia. 14
- Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199. 2, 13, 14, 15
- Pantel, P. and Lin, D. (2002). Discovering Word Senses from Text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Alberta, Canada. 7
- Parikh, D. and Grauman, K. (2011). Relative Attributes. In *International Conference on Computer Vision (ICCV)*, pages 503–510, Barcelona, Spain. 57
- Patterson, G., Xu, C., Su, H., and Hays, J. (2014). The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *International Journal of Computer Vision*, 108(1-2):59–81. 57, 58
- Patwardhan, S. and Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8. 98, 99
- Pennacchiotti, M., De Cao, D., Basili, R., Croce, D., and Roth, M. (2008). Automatic Induction of FrameNet Lexical Units. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 457–465, Honolulu, Hawaii. 14



- Perfetti, C. (1998). The Limits of Co-occurrence: Tools and Theories in Language Research. *Discourse Processes*, 25(2&3):363–377. i, 2, 23
- Plunkett, K., Hu, J.-F., and Cohen, L. B. (2008). Labels Can Override Perceptual Categories in Early Infancy. *Cognition*, 106(2):665–681. 92
- Pollack, J. B. (1990). Recursive Distributed Representations. *Artificial Intelligence*, 46(1–2):77–105. 83
- Quinn, P. C. and Eimas, P. D. (1986). On Categorization in Early Infancy. *Merrill–Palmer Quarterly*, 32(4):331–363. 121, 122
- Ranzato, M., Poultney, C., Chopra, S., and LeCun, Y. (2006). Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems 19*, NIPS 2006, pages 1137–1144. 21, 81, 88
- Ranzato, M. and Szummer, M. (2008). Semi-supervised Learning of Compact Document Representations with Deep Networks. In *Proceedings of the 25th International Conference on Machine Learning*, pages 792–799, Helsinki, Finland. 90
- Rastegari, M., Diba, A., Parikh, D., and Farhadi, A. (2013). Multi-attribute Queries: To Merge or not to Merge? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3310–3317, Portland, Oregon. 58
- Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, pages 448–453, Montréal, Canada. 7
- Rifai, S., Dauphin, Y. N., Vincent, P., Bengio, Y., and Muller, X. (2011). The Manifold Tangent Classifier. In *Advances in Neural Information Processing Systems 24*, pages 2294–2302. 84
- Rips, L. J. (1975). Inductive Judgments about Natural Categories. *Journal of Verbal Learning and Verbal Behavior*, 14:665–681. 109
- Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., and Patterson, K. (2004). Structure and Deterioration of Semantic Memory: a Neuropsychological and Computational Investigation. *Psychological Review*, 111(1):205–235. 13, 20, 28
- Rogers, T. T. and McClelland, J. L. (2004). Semantic Cognition: A Parallel Distributed Processing Approach. In *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press. 12, 19, 23
- Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., and Schiele, B. (2010). What Helps Where - And Why? Semantic Relatedness for Knowledge Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 910–917, San Francisco, California. 58

- Roller, S. and Schulte im Walde, S. (2013). A Multimodal LDA Model integrating Textual, Cognitive and Visual Modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, Seattle, Washington. 24, 26, 27, 94
- Rosch, E. (1977). *Studies in Cross-cultural Psychology*, volume 1, chapter Human Categorization, pages 1–49. London: Academic Press. 109
- Rosch, E., Mervis, C., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic Objects in Natural Categories. *Cognitive Psychology*, 8:382–439. 5
- Rosch, E. and Mervis, C. B. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7(4):573–605. 6, 13
- Roth, M. and Lapata, M. (2015). Context-aware Frame-Semantic Role Labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460. 2, 22
- Rubenstein, H. and Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633. 7
- Rumelhart, D., McClelland, J., and Group, P. R. (1986a). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (V1 and V2)*, volume 1–2. MIT press, Cambridge. 19
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning Internal Representations by Error Propagation. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: Foundations, pages 318–362. MIT Press. 87
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. 84, 114, 120, 123, 127, 156
- Russakovsky, O. and Fei-Fei, L. (2010). Attribute Learning in Large-scale Datasets. In *European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes*, Crete, Greece. <http://www.image-net.org/download-attributes>. 57, 58
- Russell, B., Torralba, A., Murphy, K., and Freeman, W. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77:157–173. 32, 49, 58
- Sainath, T. N., Mohamed, A.-r., Kingsbury, B., and Ramabhadran, B. (2013). Deep Convolutional Neural Networks for LVCSR. In *IEEE 2013 International Conference on Acoustics, Speech and Signal Processing*, pages 8614–8618. IEEE. 84
- Salton, G. and Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523. 16
- Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA. 67

- Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2006). A More Rational Model of Categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 726–731, Vancouver, Canada. 105
- Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123. 2, 14
- Schwenk, H. (2007). Continuous Space Language Models. *Computer Speech & Language*, 21(3):492–518. 21
- Schwenk, H. (2010). Continuous-space Language Models for Statistical Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 93:137–146. 21
- Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47. 2, 14
- Sharma, G. and Jurie, F. (2011). Learning Discriminative Spatial Representation for Image Classification. In *Proceedings of the British Machine Vision Conference*, pages 6.1–6.11, Dundee, UK. <https://sharma.users.greyc.fr/hatdb/>. 58
- Silberer, C., Ferrari, V., and Lapata, M. (2013). Models of Semantic Representation with Visual Attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582, Sofia, Bulgaria. Association for Computational Linguistics. 10, 24, 25, 94
- Silberer, C. and Lapata, M. (2012). Grounded Models of Semantic Representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea. 10, 24, 27, 95
- Silberer, C. and Lapata, M. (2014). Learning Grounded Meaning Representations with Autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland. Association for Computational Linguistics. 10, 24, 25, 27
- Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556. <http://arxiv.org/pdf/1409.1556.pdf>. 84
- Sivic, J. and Zisserman, A. (2003). Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, Nice, France. 57, 65
- Sloman, S. A., Love, B. C., and Ahn, W.-K. (1998). Feature Centrality and Conceptual Coherence. *Cognitive Science*, 22(2):189–228. 11
- Sloutsky, V. M. and Fisher, A. V. (2012). Linguistic Labels: Conceptual Markers or Object Features? *Journal of Experimental Child Psychology*, 111(1):65–86. 92

- Smith, E. E., Shoben, E. J., and Rips, L. J. (1974). Structure and Process in Semantic Memory: A Featural Model for Semantic Decisions. *Psychological Review*, 81(3):214–241. 12
- Socher, R. (2014). *Recursive Deep Learning for Natural Language Processing and Computer Vision*. PhD thesis, Computer Science Department, Stanford University. 83
- Socher, R., Bauer, J., Manning, C. D., and Andrew Y., N. (2013a). Parsing with Compositional Vector Grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria. Association for Computational Linguistics. 22
- Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Y. (2013b). Zero-Shot Learning Through Cross-Modal Transfer. In *Advances in Neural Information Processing Systems 26*, pages 935–943. 22, 27, 86
- Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic Compositionality Through Recursive Matrix-vector Spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, Jeju Island, Korea. 22
- Socher, R., Karpathy, A., Le, Q. V., Manning, C., and Ng, A. (2014). Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics*, pages 113–124. 27, 86
- Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh, Scotland. 90
- Sohn, K., Shang, W., and Lee, H. (2014). Improved Multimodal Deep Learning with Variation of Information. In *Advances in Neural Information Processing Systems 27*, pages 2141–2149, Montréal, Canada. 85, 86
- Srivastava, N. and Salakhutdinov, R. (2012). Multimodal Learning with Deep Boltzmann Machines. In *Advances in Neural Information Processing Systems 25*, pages 2231–2239. 85
- Srivastava, N. and Salakhutdinov, R. (2014). Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learning Research*, 15:2949–2980. 85, 86
- Steyvers, M. (2010). Combining Feature Norms and Text Data with Topic Models. *Acta Psychologica*, 133(3):234–342. 24, 26
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112, Montréal, Canada. 84

- Szumanski, S., Gomez, F., and Sims, V. K. (2013). A New Set of Norms for Semantic Relatedness Measures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 890–895, Sofia, Bulgaria. 97, 99
- Taylor, K. I., Devereux, B. J., Acres, K., Randall, B., and Tyler, L. K. (2012). Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. *Cognition*, 122(3):363–374. 13
- Tellex, S., Katz, B., Lin, J., Fernandes, A., and Marton, G. (2003). Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 41–47, Toronto, Canada. 14
- Tenenbaum, J. B. (1999). Bayesian Modeling of Human Concept Learning. In Kearns, M., Solla, S., and Cohn, D., editors, *Advances in Neural Information Processing Systems 11*, pages 59–68. MIT Press. 122
- Thompson-Schill, S. L., Kurtz, K. J., and Gabrieli, J. D. E. (1998). Effects of Semantic and Associative Relatedness on Automatic Priming. *Journal of Memory and Language*, 38(4):440–458. 96
- Torresani, L., Szummer, M., and Fitzgibbon, A. (2014). Classes: A Compact Image Descriptor for Efficient Novel-Class Recognition and Search. In Cipolla, R., Battiato, S., and Farinella, G. M., editors, *Registration and Recognition in Images and Videos*, volume 532 of *Studies in Computational Intelligence*, pages 95–111. Springer Berlin Heidelberg. 121
- Turian, J., Ratinov, L., and Bengio, Y. (2010). Word Representations: A Simple and General Method for Semi-supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. 22
- Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188. 1, 14, 15, 18
- Tyler, L., Moss, H., Durrant-Peatfield, M., and Levy, J. (2000). Conceptual Structure and the Structure of Concepts: A Distributed Account of Category-Specific Deficits. *Brain and Language*, 75(2):195–231. 20
- Tyler, L. K. and Moss, H. E. (2001). Towards a Distributed Account of Conceptual Knowledge. *TRENDS in Cognitive Sciences*, 5(6):244–252. 13
- Vanpaemel, W., Storms, G., and Ons, B. (2005). A Varying Abstraction Model for Categorization. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 2277–2282, Stresa, Italy. 105
- Varma, M. and Zisserman, A. (2005). A Statistical Approach to Texture Classification from Single Images. *International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis*, 62(1–2):61–81. 65

- Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. (2013). Decoding with Large-Scale Neural Language Models Improves Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1387–1392, Seattle, Washington, USA. 21
- Vigliocco, G., Vinson, D. P., Lewis, W., and Garrett, M. F. (2004). Representing the Meanings of Object and Action Words: The Featural and Unitary Semantic Space Hypothesis. *Cognitive Psychology*, 48(4):422–488. 2, 12
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103. ACM. 88, 89
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 11:3371–3408. 83, 88, 89
- Vinson, D. P. and Vigliocco, G. (2008). Semantic Feature Production Norms for a Large Set of Objects and Events. *Behavior Research Methods*, 40(1):183–190. 2, 12, 13
- von Ahn, L. and Dabbish, L. (2004). Labeling Images with a Computer Game. In *Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '04*, pages 319–326, New York, NY, USA. ACM. 24, 32, 49, 58, 102
- Voorspoels, W., Vanpaemel, W., and Storms, G. (2008). Exemplars and Prototypes in Natural Language Concepts: A Typicality-based Evaluation. *Psychonomic Bulletin & Review*, 15:630–637. 13, 44
- Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology. <http://www.vision.caltech.edu/visipedia/CUB-200-2011.html>. 58
- Wang, J., Yan, F., Aker, A., and Gaizauskas, R. (2014). A Poodle or a Dog? Evaluating Automatic Image Annotation Using Human Descriptions at Different Levels of Granularity. In *Proceedings of the Third Workshop on Vision and Language*, pages 38–45, Dublin, Ireland. 121
- Weeds, J., Weir, D., and McCarthy, D. (2004). Characterising Measures of Lexical Distributional Similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1015–1021, Geneva, Switzerland. 18
- Westermann, G. and Mareschal, D. (2014). From Perceptual to Language-mediated Categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1634):20120391. 20, 88, 92, 105

- Weston, J., Bengio, S., and Usunier, N. (2010). Large Scale Image Annotation: Learning to Rank with Joint Word-image Embeddings. *Machine Learning*, 81(1):21–35. 27, 86
- Xu, F. and Tenenbaum, J. B. (2007). Word learning as Bayesian Inference. *Psychological Review*, 114(2):245–272. 121
- Xu, J. and Croft, W. B. (2000). Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems*, 18(1):79–112. 6
- Yarowsky, D. (1992). Word-sense Disambiguation Using Statistical Models of Roget’s Categories Trained on Large Corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, pages 454–460, Nantes, France. 6
- Yatskar, M., Galley, M., Vanderwende, L., and Zettlemoyer, L. (2014). See No Evil, Say No Evil: Description Generation from Densely Labeled Images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics*, page 110–120, Dublin, Ireland. 120
- Yee, E., Chrysikou, E. G., and Thompson-Schill, S. L. (2013). Semantic memory. In Ochsner, K. N. and Kosslyn, S., editors, *The Oxford Handbook of Cognitive Neuroscience*, volume 1: Core Topics. Oxford University Press. 5
- Zhang, H., Zha, Z.-J., Yang, Y., Yan, S., Gao, Y., and Chua, T.-S. (2013). Attribute-augmented Semantic Hierarchy: Towards Bridging Semantic Gap and Intention Gap in Image Retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 33–42, Barcelona, Spain. 120
- Zhang, X. and Lapata, M. (2014). Chinese Poetry Generation with Recurrent Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Doha, Qatar. 84
- Zhao, Y., Karypis, G., and Fayyad, U. (2005). Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168. 129