THE UNIVERSITY *of* EDINBURGH

A STRUCTURED REPRESENTATION OF IMAGES FOR LANGUAGE GENERATION AND IMAGE RETRIEVAL

Desmond Elliott

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy to the University of Edinburgh

2014

For Jennifer

ABSTRACT

A photograph typically depicts an aspect of the real world, such as an outdoor landscape, a portrait, or an event. The task of creating abstract digital representations of images has received a great deal of attention in the computer vision literature because it is rarely useful to work directly with the raw pixel data. The challenge of working with raw pixel data is that small changes in lighting can result in different digital images, which is not typically useful for downstream tasks such as object detection. One approach to representing an image is automatically extracting and quantising visual features to create a bag-of-terms vector. The bag-of-terms vector helps overcome the problems with raw pixel data but this unstructured representation discards potentially useful information about the spatial and semantic relationships between the parts of the image. The central argument of this thesis is that capturing and encoding the relationships between parts of an image will improve the performance of extrinsic tasks, such as image description or search. We explore this claim in the restricted domain of images representing events, such as riding a bicycle or using a computer.

The first major contribution of this thesis is the Visual Dependency Representation: a novel structured representation that captures the prominent region–region relationships in an image. The key idea is that images depicting the same events are likely to have similar spatial relationships between the regions contributing to the event. This representation is inspired by dependency syntax for natural language, which directly captures the relationships between the words in a sentence. We also contribute a data set of images annotated with multiple human-written descriptions, labelled image regions, and gold-standard Visual Dependency Representations, and explain how the gold-standard representations can be constructed by trained human annotators.

The second major contribution of this thesis is an approach to automatically predicting Visual Dependency Representations using a graph-based statistical dependency parser. A dependency parser is typically used in Natural Language Processing to automatically predict the dependency structure of a sentence. In this thesis we use a dependency parser to predict the Visual Dependency Representation of an image because we are working with a discrete image representation – that of image regions. Our approach can exploit features from the region annotations and the description to predict the relationships between objects in an image. In a series of experiments using gold-standard region annotations, we report significant improvements in labelled and unlabelled directed attachment accuracy over a baseline that assumes there are no relationships between objects in an image.

Finally, we find significant improvements in two extrinsic tasks when we

represent images as Visual Dependency Representations predicted from gold-standard region annotations. In an image description task, we show significant improvements in automatic evaluation measures and human judgements compared to state-of-the-art models that use either external text corpora or region proximity to guide the generation process. In the query-by-example image retrieval task, we show a significant improvement in Mean Average Precision and the precision of the top 10 images compared to a bag-of-terms approach. We also perform a correlation analysis of human judgements against automatic evaluation measures for the image description task. The automatic measures are standard measures adopted from the machine translation and summarization literature. The main finding of the analysis is that unigram BLEU is less correlated with human judgements than Smoothed BLEU, Meteor, or skip-bigram ROUGE.

Lay Summary

Objects typically occur in the world in predictable configurations. In terms of the spatial relationships between objects, a photograph of a person riding a bicycle looks substantially different compared to a person repairing a bicycle. People tend to be above or on bicycles when they are riding them, but beside them when they are repairing them. In this thesis we propose a new type of image representation, the Visual Dependency Representation, that captures the important spatial relationships between objects in an image. We use this image representation to improve the accuracy of automatic image description and semantically-similar image search.

The first contribution of this thesis is a collection of images annotated with descriptions of the actions, objects, and Visual Dependency Representations. The images were sourced from an existing data set, thus providing additional resources to the research community. We then present an approach to automatically predicting Visual Dependency Representations and show that extracting data from the image, or from the description of the image, improves the accuracy of our method. Finally, we examine where the Visual Dependency Representation is a useful way of representing an image for automatic image description and image search. We find significant improvements in both tasks compared to approaches based on the state of the art.

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.

(Desmond Elliott.)

Acknowledgements

I would like to thank Frank Keller and Victor Lavrenko for supervising my research. Their support, guidance, and patience were crucial factors in developing the work presented here. I would also like to thank Julia Hockenmaier and Jon Oberlander for examining this thesis and for their insightful comments on my work.

The Probabilistic Models and Machine Learning-for-NLP groups helped me to develop as a researcher, and the Informatics-funded trips to Firbush Point helped me remember there is more to life than being a researcher.

I would like to thank my family and friends for their support.

The European Research Council funded my research under award 203427 Synchronous Linguistic and Visual Processing.

CONTENTS

Chapter **1** Introduction **1**
Thesis Statement 4
Published Work 6

Chapter **2** Visual Dependency Representation and Data **7**
Introduction 7
Related Work on Structured Image Representations 9
Visual Dependency Representation 11
Data 20
Conclusions 31

Chapter **3** VDR Parsing **32**
Introduction 32
Related Work on Statistical Dependency Parsing 34
Graph-based Parsing Model 36
Input and Features 37
Quasi-synchronous Parsing Model 41
Experiments 44
Discussion 50
Conclusions 52

Chapter **4** Image Description **54**
Introduction 54
Related Work on Automatic Image Description 57
Image Description Models 60
Analysis of Automatic Evaluation Measures 66
Language Generation Experiments 74
Conclusions 83

Chapter **5** Image Retrieval **86**
Introduction 86
Related Work 87
Task and Baseline 90
Comparing Visual Dependency Representations 91
Data 93
Experiments 94
Conclusion 100

Chapter **6** Conclusions **102**
Future Work 104

Appendices **107**

A Image Annotation Guidelines **108**
Introduction 109
Nouns and Objects 109
Polygons and Labelling 113

B VDR Annotation Guidelines **114**
Introduction 115
Geometric Dependency Grammar 115
Process 115
dotty 120

1

> The drawing shows me at a glance what would be spread over ten pages in a book.
>
> Turgenev (1862), translated by Constance Garnett in 1917

If a picture is worth a thousand words, then what could we do with a computer that can automatically describe pictures? Large organisations such as newspapers and libraries could digitise and more easily access their vast archives of photographs produced before the era of digital photography. People with visual impairments could more inclusively experience digital resources such as the World Wide Web with automatically described images. And at a personal level, it would be possible to search through digital photographs using natural language sentences, such as "Find photographs of a person riding a bicycle".

Images are found alongside text for a variety of reasons, including: to help organise thoughts; to draw comparisons between ideas; to reiterate the purpose of the text; to present information in a compact manner; or to offer visual explanations of ideas (Marsh and White, 2003). People can easily produce descriptions of natural photographs but the cost of collecting human-written descriptions of large image collections has resulted in a concerted effort to automate the process of generating image keywords (Duygulu et al., 2002; Lavrenko et al., 2003; Guillaumin and Mensink, 2009), captions (Feng and Lapata, 2010) and literal descriptions (Farhadi et al., 2010; Yang et al., 2011; Hodosh et al., 2013). The distinction between the output of these three strands of research is that keywords are typically nouns: *dog*, *beach*, *person*, *sunset*. Captions typically support the point of a document and are not literal descriptions of an image. For example, a news article on the topic of the drinking behaviours of teenagers may be accompanied by an image captioned as "There is a tendency for young people to not seriously consider the implications of binge drinking". And literal descriptions typically explain what can clearly be seen in an image and are therefore not usually written by a person sharing a

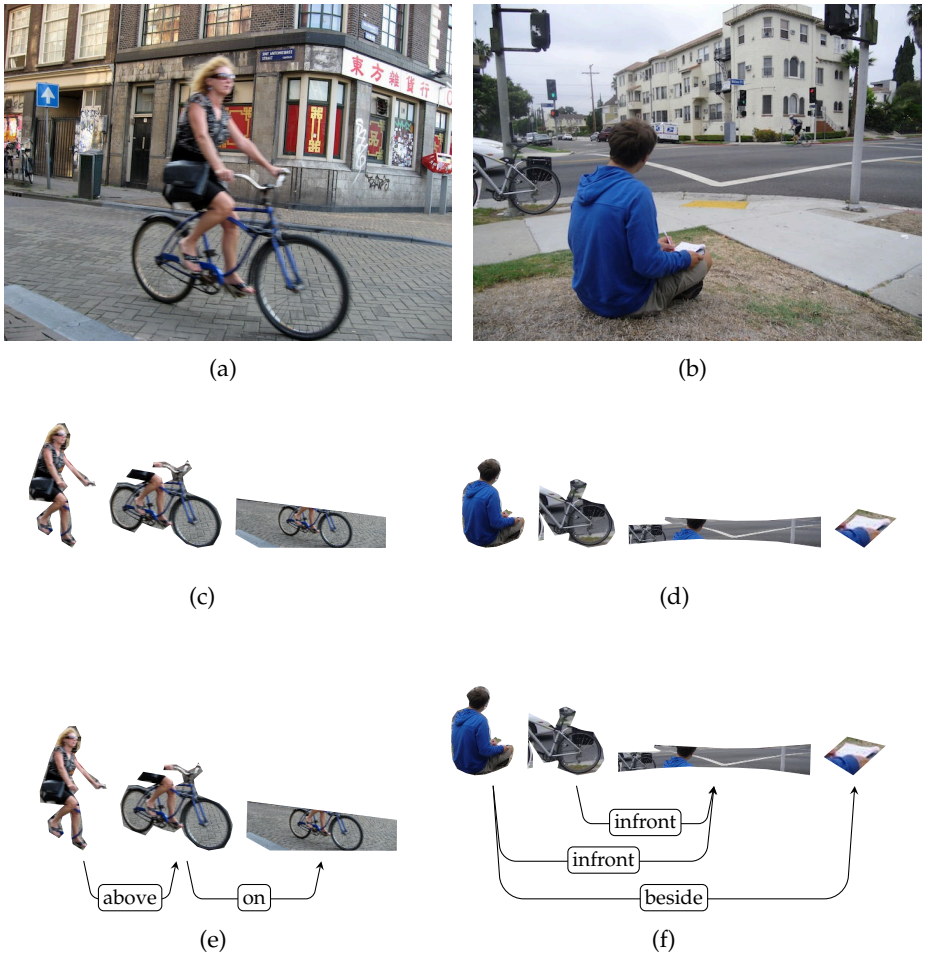


Figure 1.1 *An overview of how the Visual Dependency Representation helps distinguish between co-occurring objects and objects that co-occur to depict an action. (a) and (b) show a pair of images that represent different actions: riding a bike versus reading a book. (c) and (d) show hypothetical bag-of-regions representations of these images extracted by a perfect object detector. In the bag-of-terms representation there is no distinction between how the objects co-occur, so extrinsic applications are likely to predict the same underlying meaning (a person riding a bike). (e) and (f) show how the Visual Dependency Representation distinguishes the relationships between the objects. In particular, there is no relationship expressed between the boy and the bicycle in the second image, making it less likely for extrinsic applications to infer the wrong meaning of an image.*

photograph. However, these types of descriptions are extremely useful for finding images because they often describe the *who*, *what*, *where*, *when*, and *why* (Shatford, 1986). In this thesis, we focus on the task of generating literal descriptions that explain what is happening in an image.

One of the central problems of accessing large image collections is the *semantic gap* between how people think about images and how computers represent them (Smeulders and Worring, 2000). Computers natively represent images as a matrix of pixels, and it is rarely possible to compare images based on these matrices. This is because even slight changes in the illumination of an image results in a different set of pixel values, making a direct comparison difficult. The computer vision community has a substantial volume of literature devoted to the extraction of discriminative visual features that can be used to represent images, such as histograms of colour, shape, and texture, the Scale-Invariant Feature Transform (Lowe, 1999), and the Histogram of Oriented Gradients (Dalal and Triggs, 2005). One approach to representing an image is to create an unstructured bag-of-terms vector over the extracted visual features. Figure 1.1 (c) and (d) show an example of a bag-of-regions representation of an image, where the visual feature extraction method is a perfect object detector[1]. The bag-of-regions representation is easy to construct and manipulate, but it *discards* potentially useful information about how the visual features occur together in an image.

The work presented in this thesis lies firmly within the emerging field of connecting language with vision. This is a broad field encompassing work on multimodal distributional semantics (Silberer and Lapata, 2012; Silberer et al., 2013); image captioning (Feng and Lapata, 2010); literal image description (Farhadi et al., 2010) *inter alia*; multimodal image ranking (Hodosh et al., 2013); and video description (Regneri et al., 2013; Krishnamoorthy et al., 2013). We make contributions to the literal image description problem by proposing a new structured representation - the Visual Dependency Representation - that encodes information about the spatial relationships between different regions of an image. The Visual Dependency Representation is based on natural language dependency

[1]The perfect automatic object detector does not yet exist, but this is an illustrative example.

syntax (Tesnière, 1953), and makes it possible to distinguish between image regions that co-occur, and regions that contribute to explaining the underlying meaning of the image. The rationale for encoding the spatial relationships between regions is partly due to evidence in the human vision literature that people are more able to identify cued objects in an image where the spatial relationships between objects, or parts of objects, are consistent with their expectations (Biederman, 1972; Bar and Ullman, 1996). Figure 1.1 (e) and (f) show an example of the Visual Dependency Representation of an image, where the relationships between the person and the object they are interacting with are encoded in the image representation. In the work presented in this thesis, we work within the restricted domain of images depicting actions. The intuition is that over a collection of images, there should be predictable spatial relationships between the parts of an image that contribute to depicting the same action.

1.1 THESIS STATEMENT

> It is *useful* to use the structured Visual Dependency Representation of images for tasks that involve understanding the action depicted in an image.

We define *useful* in terms of improvements in the performance of the extrinsic tasks of automatic image description and query-by-example image retrieval. This hypothesis is directly tested on those tasks in Chapters 4 and 5 respectively. We also test a second hypothesis:

> It is possible to automatically predict Visual Dependency Representations from region-labelled images, such that the predicted representations are *useful* for the extrinsic tasks.

In Chapter 3 we present an approach for automatically predicting the Visual Dependency Representation of an image. The approach is based on a graph-based statistical dependency parser that can exploit data from image region annotations and parallel descriptions. The improvements in the extrinsic tasks of language generation and image retrieval are also determined when automatically predicted Visual Dependency Representations are used instead of gold-standard data.

In testing these hypotheses, we work from the assumption that we have a perfect visual feature detector. More specifically, the visual feature detector is of the form of an automatic object detector that always correctly localises and labels the objects in an image. This simplifying assumption allows us to explore the potential utility of the proposed representation in the absence of the noise introduced by automatic computer vision methods. In Chapter 6 we outline some thoughts on how we could incorporate noisy detections in the model.

The chapters of the thesis are organised as follows:

- In Chapter 2, the Visual Dependency Representation is described and a data set of photographs annotated with this representation is presented. This chapter includes a review of structured image representations and existing available data sets, how our data set was collected, and a quantitative analysis of its properties.
- In Chapter 3, we describe one approach for inducing the Visual Dependency Representation of an image. We use a quasi-synchronous dependency parser which operates over annotated image regions and dependency representations of descriptions. A collection of experiments are presented on combining different sets of features in the parsing model and an error analysis explains the problems encountered.
- In Chapter 4, we show how the Visual Dependency Representation can be used to improve image description generation. This chapter includes a review of recent approaches for automatic image description, and a correlation analysis of human judgements against possible automatic evaluation measures for image generation. We find that the Visual Dependency Representation improves the content-selection component of our language generation model because the representation encodes the relationships between objects that contribute to the depicted action.
- In Chapter 5, we show how Visual Dependency Representations can improve the accuracy of a query-by-example image retrieval model compared to a bag-of-terms baseline. We show how to compare

images that are represented as Visual Dependency Representations, and find improvements in retrieval accuracy at both the top of the ranked list and throughout the entire list. A post-hoc analysis shows that most of the gains are found for transitive verbs, and that there is a mixed result for light verbs.

- Finally, in Chapter 6 we present some concluding remarks and outline some future work on incorporating automatic visual feature extractors, and the semi-supervised learning of Visual Dependency Representations.

1.2 PUBLISHED WORK

Chapter 2 was presented as:

> D. Elliott and F. Keller. 2011. A Treebank of Visual and Linguistic Data. In Proceedings of the Workshop on Integrating Language and Vision at Neural Information Processing Systems 2011. Granada, Spain.

Chapter 4 was presented as:

> D. Elliott and F. Keller. 2013. Image Description using Visual Dependency Representation. In Proceedings of the 2013 Conference of Empirical Methods in Natural Language Processing. Seattle, Washington, U.S.A.

> D. Elliott and F. Keller. 2014. Comparing Automatic Evaluation Measures for Image Description. In Proceedings of the 52nd Annual Meeting of the Association of Computational Linguistics. Baltimore, Maryland, U.S.A.

Chapter 5 was presented as:

> D. Elliott, V. Lavrenko, and F. Keller. 2014. Query-by-example Image Retrieval using Visual Dependency Representations. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. Dublin, Ireland.

2

One approach to representing an image is the bag-of-terms vector, which can be formed by extracting and clustering visual features from an image, or from predicted or user-generated labels for an image. The bag-of-terms representation of an image has proven successful for automatic image description and retrieval; however, this unstructured representation discards potentially important information about the absolute and relative locations of terms, and how they relate to each other. The central argument of this thesis is that capturing and encoding the structural relationships between regions of images will improve the performance of extrinsic tasks.

In this chapter, we review approaches to extracting visual features from images, and argue that removing information about the structure of an image from its representation can lead to problems. We present a new structured representation of images, the Visual Dependency Representation, and show how it can be used to overcome the problems suffered by unstructured representations. In particular, we focus on representing the relationships between regions in images depicting actions, such as riding a bike or reading a book. We describe how Visual Dependency Representations can be created by human annotators and introduce a new data set of annotated images paired with descriptions and Visual Dependency Representations.

2.1 INTRODUCTION

The bag-of-terms vector is commonly used in natural language processing and information retrieval to create an unstructured representation of text (Manning et al., 2008). A bag-of-terms representation of a document is an unordered vector that encodes presence or absence of terms. There are many possible pre-processing steps to creating a bag-of-terms vector, such as removing stop-words from the input, and lemmatising the text. The elements of this vector can either be binary or weighted according

to a variety of weight schemes, such as absolute frequency, or *tf-idf*. This representation is easy to construct, easy to manipulate, and scales well to large-scale data sets.

One approach to representing the content of an image is a bag-of-terms vector (Datta et al., 2008). A bag-of-terms representation of an image is typically formed by extracting and clustering visual features from the image. The visual features can be chosen from colour, position, and size features extracted from segmented image regions (Shi and Malik, 2000), points-of-interest features (Lowe, 1999), histograms of oriented gradients (Dalal and Triggs, 2005), *inter alia*. An alternative to using clustered visual features directly is to use surrounding text (if available) or to create a term vector automatically from an image tagger (Lavrenko et al., 2003; Guillaumin and Mensink, 2009), or an object detector (Felzenszwalb et al., 2010).

In this thesis, we argue that the unstructured bag-of-terms vector can discard important information about how different parts of an image relate to each other. This argument was previously made by with regards to how creating higher-order representations over automatically extracted visual features (Lazebnik et al., 2006); here we work with object annotations as the atomic unit of image feature. Furthermore, we will explore this argument within the domain of images representing actions instead of indoor scenes. Figure 2.1 shows a pair of images that demonstrate the limitation of an unstructured image representations. In this example, both images are in a similar outdoor context and there is a person and a bicycle in both images, but the person in the second image is reading a book instead of riding the bicycle.

In this chapter we develop the structured representation of images that exploits the spatial relationships between actors and objects. The spatial relationships between the entities involved in the action have been found to have a significant effect on how humans process visual information. In a visual search task, Biederman (1972) found that people were less accurate at identifying cued objects when the photograph had been sliced into six segments and jumbled, compared to viewing the photograph in its original form, thus preserving the context. In a visual recognition task,

Figure 2.1 *An example of the type of problem encountered by a bag-of-words representation of an image. The bag-of-words approach to representing an image makes it difficult to distinguish between these images, each of which depict a person and a bike. This thesis argues that the spatial relationships between regions is crucial for understanding what is happening in an image.*

Bar and Ullman (1996) found that when objects were placed in incorrect spatial relationships, people were significantly less likely to correctly identify the underlying object. These observations suggest it is not that people and objects occur together, rather that they occur together in particular spatial configurations, that determine whether an action can be observed in an image.

2.2 RELATED WORK ON STRUCTURED IMAGE REPRESENTATIONS

There has been relatively little work on *explicit* structured representations of images. In this section we briefly review the key works in this area.

Structured Object Queries are object–relation–object tuples that capture which objects are present in a scene and the spatial relationship between the objects (Lan et al., 2012). They have been used for image retrieval with a latent rank Support Vector Machine (Yu and Joachims, 2009) on the SUN 09 data set (Choi et al., 2010). This representation was found to significantly increase the Mean Average Precision of retrieved images compared to a bag-of-terms baseline. The approach works on the output of pre-trained parts-based object detectors (Felzenszwalb et al., 2010) and scene-type classifiers (Oliva and Torralba, 2001). Only five types of structured object queries were studied because "it is impossible to consider all possible combinations of objects and relations as queries".

Spatial Pyramid Matching is an unsupervised approach to capturing the structure of an image (Lazebnik et al., 2006), based on pyramid matching

Figure 2.2 *Sample photographs from the Oxford 5K and University of Kentucky data set. These data sets contain images of the same object from different perspectives, which is a very different notion of structure than what we will try to capture in the Visual Dependency Representation.*

(Grauman and Darrell, 2005). Pyramid matching extracts features from an image at different levels of granularity and represents an image as the weighted sum of the features over those levels of granularity, regardless of where in the images the features were extracted. The intuition is that we can capture features that occur at the level of the entire image, down to features that occur in smaller patches. Spatial pyramid matching extends idea this by requiring pyramid matches to occur in the same region of the image. The features extracted using spatial pyramid matching can be used to train a Support Vector Machine to predict scene types (Fei-Fei and Perona, 2005), or to perform object recognition (Fei-Fei et al., 2004). It is also possible to directly use the output of the matching kernel as a means of ranking the similarity of images in a data set.

Geometry-Preserving Visual Phrases incorporate local and long-range interactions between visual words in a bag-of-terms representation (Zhang et al., 2011). The approach works on low-level image features and is encoded directly into the bag-of-visual words representation as *phrases* of visual words that co-occur together in certain contexts. It was evaluated on image retrieval on the Oxford 5K (Philbin et al., 2007) and University of Kentucky Dataset (Nister and Stew, 2006). Figure 2.2 shows examples of the types of images in data used for these experiments.

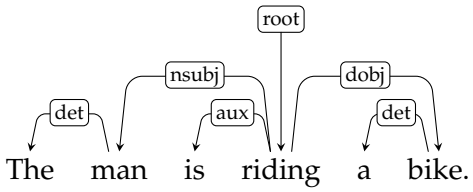
Visual Phrases (Sadeghi and Farhadi, 2011) represent the relationships between objects using a deformable parts object detector (Felzenszwalb et al., 2010) trained on bounding boxes that encapsulate the actor and the object involved in the action. The data was manually annotated with 17

possible interactions between eight classes of objects in the PASCAL VOC 2008 data set. The interactions annotated in the data are: person riding horse; person sitting on sofa; person sitting on chair; person lying on sofa; person lying on beach; person riding bicycle; horse and rider jumping; person next to horse; person next to bicycle; bicycle next to car; person jumping; person next to car; dog lying on sofa; dog running; dog jumping; person running; and person drinking from a bottle. It can be seen that some of these not necessarily events: a bicycle next to a car, or a person next to a horse.

It is only the Structured Object Queries that approach the type of structured representation of an image that we propose in this chapter. The Visual Dependency Representation, and in particular our approach to predicting it (see Chapter 3) does not not suffer from the problem of trying to enumerate all possible combinations of objects and relations. The Spatial Pyramid Matching and Geometry-Preserving Visual Phrases are actually bag-of-terms representations where some of the terms encode higher-order relationships between visual features. Furthermore, these representations work on automatically extract visual features, which makes it very difficult to compare it against representation based on image labels or object regions. The Geometry-Preserving Visual Phrases are evaluated on data sets of images of exactly the same target objects (see Figure 2.2), whereas the images we will use in our experiments represent the same event but in completely different environment (see Chapter 2.4 for more details).

2.3 VISUAL DEPENDENCY REPRESENTATION

In this section we introduce the Visual Dependency Representation (VDR), a novel structured representation of an image that models the relationships between regions of an image. This representation is inspired by natural language dependency syntax, which was originally formulated to express the relationships between words in a sentence in a dependency tree (Tesnière, 1953). An example of a dependency tree for a simple sentence can be seen below.



Formally, a dependency tree is a directed acyclic graph whose nodes represent the words in a sentence. A directed arc from a *head* to an *argument* is labelled with the syntactic relationship between those words, where the possible relationships between pairs of words are defined in a dependency grammar. In this example we can see that *riding* is the main event of the sentence, signified by it being attached to the root of the sentence. The *man* token is the subject of the verb token *riding*, denoted by the dependency arc between riding and man, and *bike* token is the object of *riding*, denoted by the arc between riding and bike. The labels on the arcs are the syntactic relations between the tokens, where the set of syntactic relations is defined to the by grammar of the formalism. In this example, we can see four different types of relation: *nsubj*, *aux*, *det*, and *dobj*.

Dependency representations of sentences can be automatically predicted by statistical dependency parsers, which we briefly review in Chapter 3.2. These representations have proven useful for tasks such as machine translation (Quirk and Menezes, 2005) and question-answering (Wang and Smith, 2007). In this thesis, we will demonstrate how dependency syntax can be applied to a completely different domain, namely that of deriving invariant representations of the relationships between regions of an image. We note here that there are other formalisms for representing the structure of a sentence, such as phrase-structure grammar (Chomsky, 1957) or Combinatory Categorical Grammar (Steedman, 1996), but neither of these seemed suited to capturing the direct relationships between regions of an image.

2.3.1 *Visual Dependency Grammar*

The Visual Dependency Representation of an image is constructed by creating a directed acyclic graph over the set of regions in an image using the spatial relationships in the Visual Dependency Grammar. The remainder

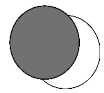
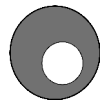
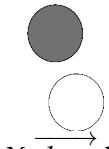
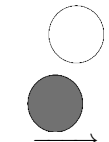
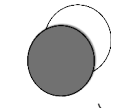
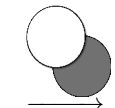
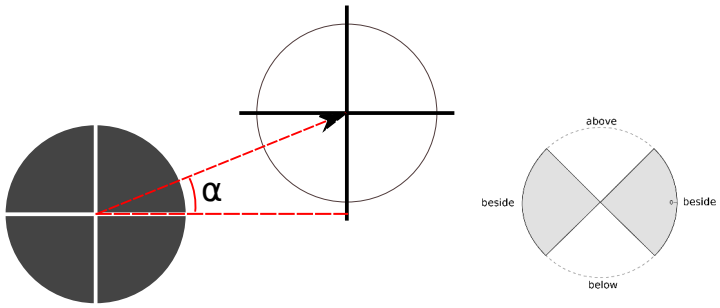
 X $\overrightarrow{on}$ Y	More than 50% of the pixels of region X overlap with region Y.
 X $\overrightarrow{surrounds}$ Y	The entirety of region Y overlaps with region X.
 X $\overrightarrow{beside}$ Y	The angle between the centroid of X and the centroid of Y lies between 315° and 45° or 135° and 225°.
 X $\overrightarrow{opposite}$ Y	Similar to *beside*, but used when there X and Y are at opposite sides of the image.
 X $\overrightarrow{above}$ Y	The angle between X and Y lies between 225° and 315°.
 X $\overrightarrow{below}$ Y	The angle between X and Y lies between 45° and 135°.
 X $\overrightarrow{infront}$ Y	The Z-plane relationship between the regions is dominant.
 X $\overrightarrow{behind}$ Y	Identical to *infront* except X is behind Y in the Z-plane.

Table 2.1 *Visual Dependency Grammar defines eight relations between pairs of annotated regions. To simplify explanation, all regions are circles, where* X *is the grey region and* Y *is the white region. All relations are considered with respect to the centroid of a region and the angle between those centroids. We follow the definition of the unit circle, in which* 0° *lies to the right and a turn around the circle is counter-clockwise.*



(a) (b)

Figure 2.3 *A visual explanation of how the angle between a pair of regions should be calculated. (a) How to determine the spatial relationship; (b) A visual explanation of which angles relate to which types of relationships. In this figure, we explain the spatial relationship between the dark region and the light region. The angle α is used to inform this decision. It is determined by drawing a straight line between the centroids of the regions and calculating the size of the angle with respect to the originating region, in this case the dark region. In this example, α maps onto the X $\overrightarrow{beside}$ Y relationship.*

of this chapter is devoted to explaining the grammar, how it is applied to images, and the creation of a data set with Visual Dependency Representations for images.

In analogy to dependency grammar for natural language syntax, we define *Visual Dependency Grammar* to describe the spatial relations between pairs of image regions. The motivation for encoding the spatial relationships between regions in an image is that we expect it to be useful for discriminating between object co-occurrence and interactions between objects that form an action.

The Visual Dependency Grammar is defined by eight spatial relationships between regions, explained in detail in Table 2.1. The spatial relations were developed in partnership with human annotators during a preliminary stage of developing the grammar. The relationships fall into two broad categories: regions that are in a proximity-based relationship - *beside*, *opposite*, *above*, *below* - or regions that overlap with each other *on*, *surrounds*, *infront*, and *behind*. The spatial relationships are defined in terms of three geometric properties: pixel overlap, the angle between regions, and the distance between regions. There are three calculations that need to be performed by a human or a computer when collecting evidence to label the relationship between a pair of regions. We note that

when humans label the relationships between objects, their decisions are based on *fuzzy* notions of angles and overlapping. A computer algorithm will make use of the precision definitions laid out here.

Angle between regions. The angle between regions is used to apply the *above*, *below*, and *besides* relationships. Figure 2.3 provides an explanation of which angle is should be calculated when making this estimation.

Overlapping regions. We follow the PASCAL VOC definition of overlap in the object detection task (Everingham et al., 2011). Namely, for regions X and Y, a pair of regions overlap when the ratio of the intersection of the regions to the union of the regions exceeds 50%:

$$overlaps(X, Y) = \frac{|X \cap Y|}{|X \cup Y} > 50\% \tag{2.1}$$

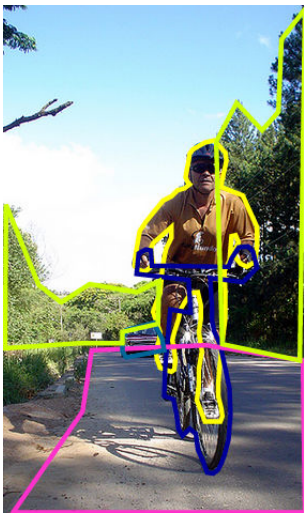
Distance. If the Angle between regions calculation results in a *beside* label, then there needs to be a distinction between what type of besides relationship exists. Regions that are more than half the image apart are deemed to be *opposite*, otherwise the regions are *beside* each other.

In comparison to Regional Connection Calculus-8 (Randell et al., 1992), which defines eight possible spatial relations between regions, the Visual Dependency Grammar forgoes the tangential relationships and the inverse-type relationships and includes two Z-dimension relationships. Now that the Visual Dependency Grammar and the rules for applying it have been defined, we now outline the five-step process to creating the Visual Dependency Representation of an image:

1. **Image description.** A description of the image is obtained either from the surrounding text or from a human. See Section 2.4.3 for more details on how we collected image descriptions.

2. **Region annotation.** Objects referred to in the description are annotated by a human. See Section 2.4.5 for more details on the regions were annotated.

3. **Initialise VDR.** A dummy region, referred to as the ROOT, is the



(a)

A man is riding a bike down the road.
A car and trees are in the background.

(b)

Figure 2.4 *A region-annotated image of a man riding a bike (a), and a human-written description of the image (b). The image and description are used as a running example to explain how Visual Dependency Representations can be created by either humans or induced by an algorithm. The regions annotated in the image are:* BIKE, CAR, MAN, ROAD, TREES*; the human-written description was obtained from Amazon Mechanical Turk.*

starting point for all Visual Dependency Representations. The ROOT region is assumed to refer to the image as a whole. This is analogous to the root node in dependency syntax for natural language.

4. **Identify central actor.** The image region that defines the subject of the image is attached to the ROOT node of the graph. The subject of the image can almost always be identified as the noun phrase that is the subject of the description of the image.

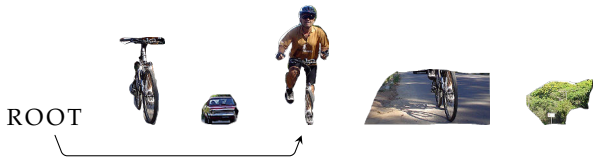
5. **Attach remaining regions.** The remaining annotated regions are attached to the graph based on the implicit or explicit relationship to other annotated regions, as defined in the description. Each arc introduced is labelled with one of the spatial relations defined in the grammar, or with no label if the region is not described in relation to anything else in the image.

The detailed set of instructions given to the human annotators who created the gold-standard data region annotations and Visual Dependency Representations can be found in Appendices A and B.

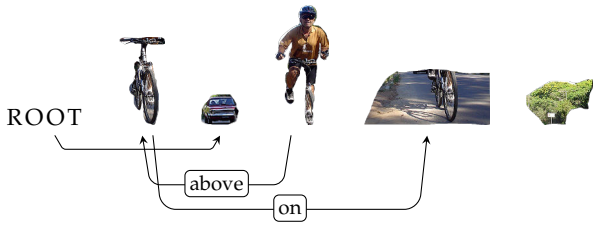
Worked Example

As an example of the output of this annotation process, consider the region-annotated image in Figure 2.4 (a), and its description in 2.4 (b). The starting point for the VDR of this image–description pair is the annotated image regions and the ROOT node:

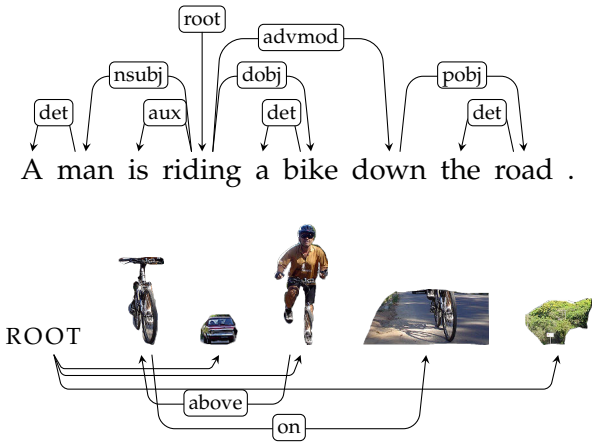
The MAN is the central actor in the image, as he is carrying out the depicted action (riding a bike). The region corresponding to MAN is therefore attached to ROOT without a spatial relation label on the arc.



The BIKE region is then attached to the MAN region using the $\overrightarrow{above}$ relation and BIKE is attached to the ROAD with the $\overrightarrow{on}$ relation. These attachments are made because the description explicitly describes the relationship between the man and the bike and the bike and the road.



In the second sentence of the description, CAR and TREES are mentioned without a relationship to anything else in the image, so they are attached to the ROOT node. This completes the process for creating the VDR of this image–description pair. In the example below we have included the syntactic dependency parse to show that the structures are not isomorphic.



There is some ambiguity regarding the dependencies to be expressed in a visual dependency tree. For example MAN *above* BIKE and BIKE *below* MAN are equivalent ways of expressing the same relationship in Figure 2.4.[1] This ambiguity is addressed by constructing visual dependency trees in the context of an image description. The image description always contains an region that is central to the image, a person performing an action in our data set; this central region is typically realized as the syntactic subject of the image description.

Multiple Actors

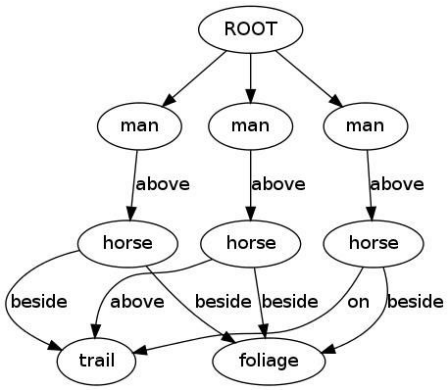
In images where more than one person is performing an action, the Visual Dependency Representation of the image can form a graph. This is especially prevalent in images where multiple subjects are performing an action on the same plane, as shown in Figure 2.5. We can see that the VDR creation process results in a structure where at least one node has multiple parents. In Figure 2.5, the TRAIL and FOLIAGE nodes exhibit this behaviour. Recall that the Visual Dependency Representation is based on dependency syntax theory for natural language, and this position will lead us to using a statistical dependency parser as the computational machinery for predicting image structures. It would be incompatible with our computational approach to have non-tree structure input data, and so we will not use data that exhibits these types of structures in the remainder of the thesis. This decision resulted in the removal of 14.4% of

[1]The linguistic analogue is the active and a passive form of the same sentence, for example, *the man is riding the bike* vs. *the bike is being ridden by the man*.

(a) An image with multiple subjects.

Three people are riding horses down a trail. They are surrounded by luscious foliage on all sides.

(b) A human-written description.



(c) A VDR for the annotated image-description pair.

Figure 2.5 *An example of a Visual Dependency Representation where the image and description concern multiple subjects performing an action on the same surface. In this example, the human-written description (b) expresses a relationship between the horse riders and the trail. This relationship can only be encoded in a VDR by producing a directed acyclic graph (c).*

	Images	Descriptions
VLT	2,424 (341)	7,272 (1,023)
PASCAL Sentences	1,000	5,000
Flickr8K	8,108	40,540
SBU 1M Captioned Photos	1,000,000	1,000,000

Table 2.2 *A comparison of data sets used in the NLP community for image description and retrieval experiments. The data set presented in this thesis is labelled VLT (Visual and Linguistic Treebank). The numbers in parenthesis indicate the proportion of our data set that has been completely annotated. The main obstacle to annotating the entire VLT data set is the training and payment of human annotators.*

the annotated data in the Visual and Linguistic Treebank.

2.4 DATA

There are several image collections available for different types of computer vision research. The specific task of object detection is well-served by the Caltech-101 and -256 data sets, which contains thousands of images of objects in 101/256 distinct categories (Fei-Fei et al., 2004); the PASCAL Visual Objects Classes data set (Everingham et al., 2010), which contains thousands of images annotated with bounding boxes of twenty object categories, and ImageNet, which contains millions of images, a subset of which are annotated with bounding boxes for thousands of categories (Deng et al., 2009).

2.4.1 *Related Data Sets*

There are a number of existing data sets pair images with image descriptions. The SBU Captioned Photo Dataset (Ordonez et al., 2011) consists of images with associated captions retrieved from Flickr, and the UIUC Pascal Sentence Dataset, Flickr8K Dataset (Rashtchian et al., 2010; Hodosh et al., 2013), and the IAPR Benchmark Dataset (Grubinger et al., 2006) pair images with descriptions generated by human annotators.

The PASCAL Sentences data set contains 1,000 images randomly sampled from the PASCAL Visual Objects Classes 2010 data set before the introduction of the Action Recognition Taster task (Rashtchian et al., 2010). The

images were sampled equally from the 20 object detection classes in the Object Detection task, and five descriptions of each image were collected from untrained annotators on Amazon Mechanical Turk.

The Flickr8K data set contains 8,108 images retrieved from the popular photo-sharing website Flickr (Rashtchian et al., 2010; Hodosh et al., 2013). The images were taken from six photo sharing groups on Flickr. 15,000 images were downloaded and manually inspected to confirm they depicted some kind of event that "would require a full sentence description, unlike the PASCAL data". Each image is associated with five human-written descriptions retrieved from Mechanical Turk.

The SBU 1M Captioned Photos data set contains 1,000,000 images, each of which is associated with the co-occurring Flickr image caption (Ordonez et al., 2011). The data set was created by retrieving a large number of images from Flickr using pairs of query terms taken from object labels, object attributes, actions, stuff[2], and scene types. The retrieved images were then filtered to include only those of a "satisfactory" length, contained at least two words on a "term list", and at least one preposition.

We concluded that none of these data sets were suitable for our research goals. There was no guarantee that the PASCAL Sentences Dataset would contain images depicting actions; there was no information on the range and frequency of the action types included in the Flickr8K Dataset; and the collection process for the SBU 1M Captioned Photos Dataset resulted in captions, not literal descriptions. The PASCAL Visual Objects Classes has been annotated with action labels. We use this subset of the PASCAL data set as the basis for the images in our experiments. In the remainder of this chapter we describe how we enriched this data set with human-written descriptions, object annotations, and Visual Dependency Representations.

2.4.2 *Visual and Linguistic Treebank*

Given the aforementioned issues with these data sets, we built the Visual and Linguistic Treebank that contains (a) a known number of depicted

[2]Stuff is a term that has gained some traction in the computer vision community. It has been used to refer to entities which are not objects, such as sky, sand, trees, etc. (Kulkarni et al., 2011)

actions and (b) could be characterised as being linguistically diverse. The images are derived from the PASCAL VOC Action Recognition task, a closed-domain data set containing images of people performing ten types of actions: jumping, walking, running, phoning, playing an instrument, reading, riding a bike, riding a horse, taking a photo; and using a computer. Table 2.2 presents a comparison of the VLT data set and the three other most commonly adopted data sets for NLP-related computer vision research. The VLT data set was annotated in a three-step process:

1. Three descriptions of each image were collected from untrained workers on Amazon Mechanical Turk;
2. Each image was annotated with the objects referred to in the collected descriptions; and
3. A Visual Dependency Representation was created for each description of an image.

Note that Steps (2) and (3) are dependent on the image description, as both the region annotations and the relations between them are derived from the description of the image.

2.4.3 *Collecting Image Descriptions*

We collected three descriptions of each image in our data set from untrained annotators on Amazon Mechanical Turk. The annotators were asked to describe an image in two sentences. The first sentence describes the action in the image, the person performing the action and the region involved in the action; the second sentence describes any other regions in the image not directly involved in the action. An example description is given in Figure 2.4b.

Descriptions were collected for all 2,424 images in the *trainval* section of the PASCAL Action Recognition data set, resulting in a total of 7,272 image descriptions. The annotators, drawn from those registered in the US with a minimum HIT acceptance rate of 95%[3], described an average of

[3]A Human Intelligence Task is a single task that a worker on Mechanical Turk can perform in return for payment. A 95% acceptance rate is considered a reliable indicator of worker performance.

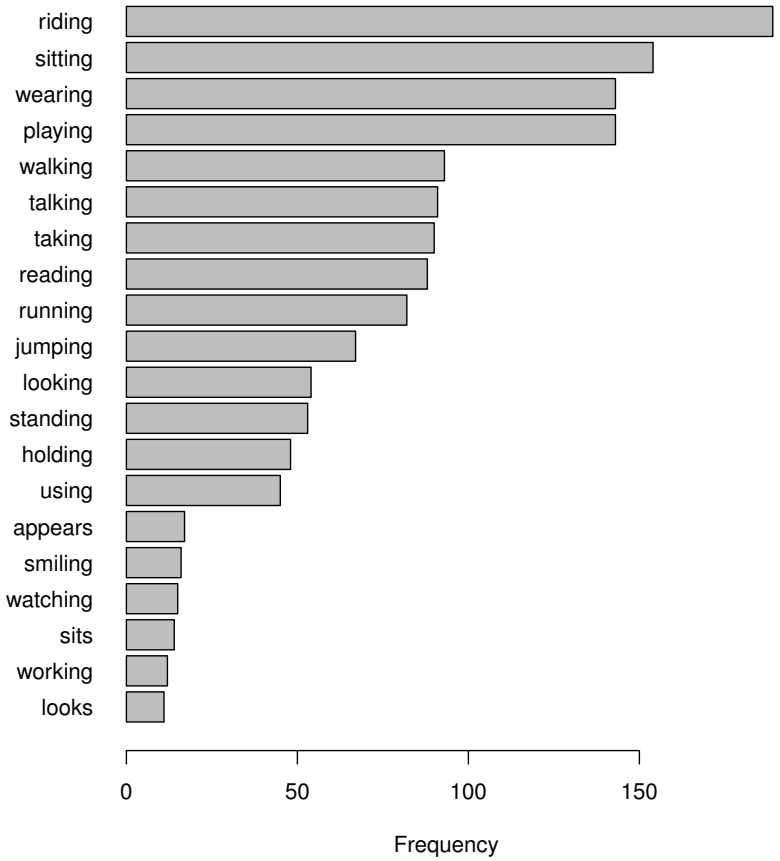


Figure 2.6 *Top 20 verbs used in image descriptions. All of the action categories in the underling images are represented in the top 20 verbs. Verbs were identified by Part-of-Speech tagging the descriptions and using tokens with a verb-type tag.*

145 ± 93 images. The annotators were encouraged to describe fewer than 300 images each to ensure a linguistically diverse data set. Annotators were paid \$0.04 per image and it took on average 67 ± 123 seconds to describe a single image. The average length of a description was 19.9 ± 6.5 words in a range of 8–50 words.

The descriptions were post-processed to include syntactic information. We part-of-speech tagged the data using the Stanford POS Tagger v.3.1.0 using the pre-trained *english-bidirectional-distsim* model, and dependency parsed the descriptions using Malt Parser v.1.7.2 using the pre-trained *engmalt.poly-1.7* model.

Figures 2.6 shows the 20 most frequently occurring verbs in the image descriptions. Verbs were identified using the POS-tagged descriptions

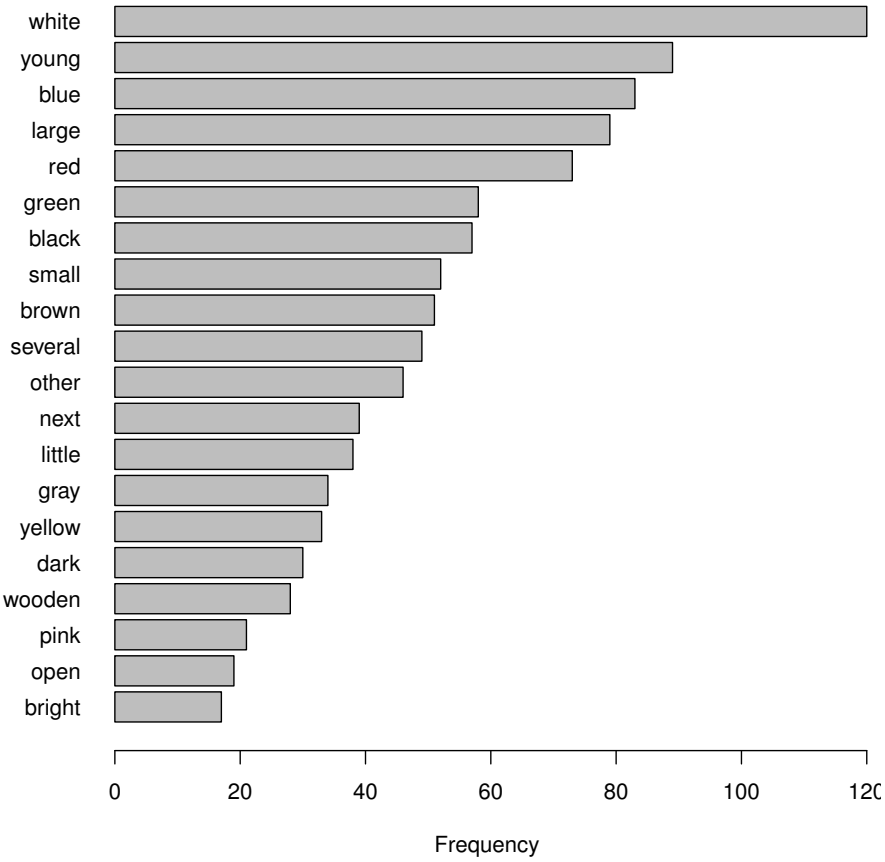


Figure 2.7 *Top 20 adjectives used the image descriptions. More than 50% of the adjectives refer to colour, and three refer to size. Adjectives were identified by Part-of-Speech tagging the descriptions and using tokens with an adjective-type tag.*

with the following tags: VBZ, VBG, VBP, VBN, VBS, and VB. We removed all forms of the verb *to be* as a stop-word. The twenty most frequently occurring verbs contain all of the verbs in our action classes, while the remainder appear to refer to actions people can perform while undertaking the actions depicted in the images. For example, a person can *sit* on a chair and *read* a book, or a person can *take* a walk in a park.

The twenty most frequently occurring adjectives, shown in Figure 2.7, were identified using the JJ, JJR, and JJS part-of-speech tags. The twenty most frequently occurring adjectives are dominated by colours. Eleven adjectives refer to colour, three refer to size, two refer to count (other, next), and only one refers to a physical attribute of the object (wooden). We chose not to use object attributes in extrinsic evaluations in Chapters

4 and 5 because the majority of the top adjectives were colour or size.

2.4.4 *Categorising Action Types*

We can categorise the actions in the our data set according to the linguistic type of the underlying verbs used to describe the images. One way of categorising the type of the action is whether the verb is transitive or intransitive (Dixon, 2005). This approach is based on the idea that the predicate of a sentence can be classified based on the predicate-argument structure. In the following example, the predicate "riding" has the subject "A man" and the object "a bike". It is classed as a transitive verb because the predicate takes one object, namely the bike.

(1) A man is riding a bike.

An intranstive verb does not require an object, however, many transitive verbs take optional objects:

(2) A man is running (a race).

This leads to the following split of verbs in our data set:

Transitive: *ride* a bike, *ride* a horse, *read* a book, *use* a computer, *talk* on the phone, *play* instrument, *take* a photo.

Intransitive: run, jump, walk.

We will see that splitting the actions into types will be instructive to understanding the strengths and limitations of Visual Dependency Representation for the downstream task of example-based image retrieval in Chapter 5.

2.4.5 *Collecting Region Annotations*

We trained two Ph.D students to annotate boundaries around regions of objects in images using the LabelMe annotation tool (Russell et al., 2008). The regions to be annotated were limited to those mentioned in the descriptions paired with the image. Region annotation was performed on

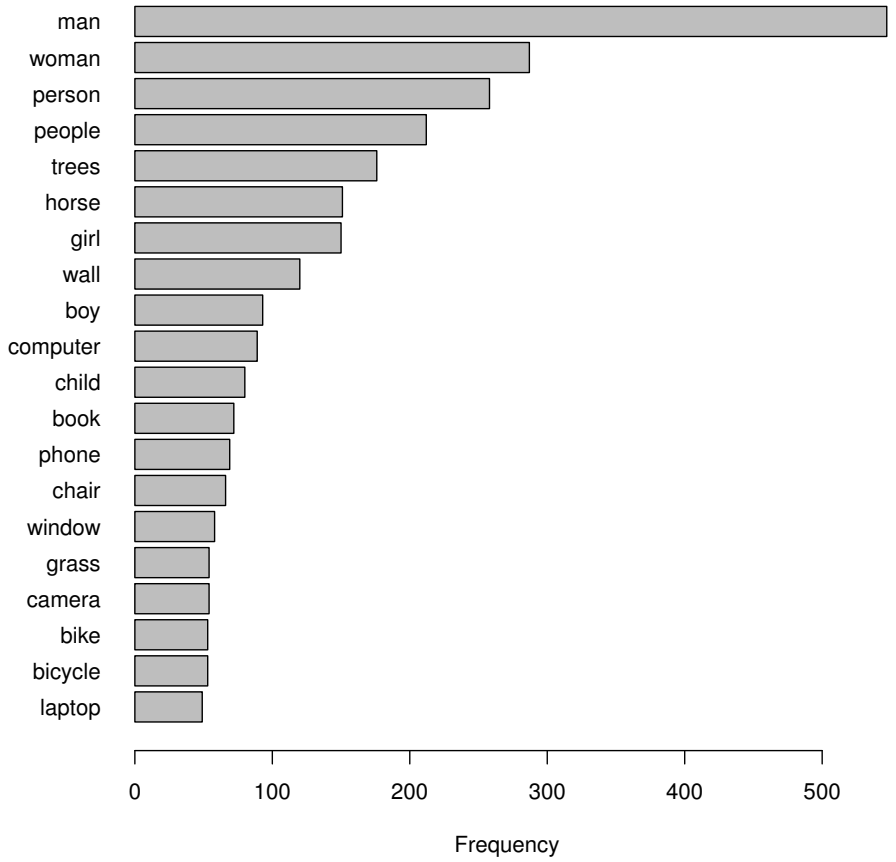


Figure 2.8 *Top 20 annotated regions in our data set.*

a subset of 341 images and resulted in a total of 5,034 annotated regions with a mean of 4.19 ± 1.94 annotations per image. Figure 2.9 shows the distribution of image region annotations in the data set; images with only one region annotation are likely to be a person performing an intransitive action. The limiting factor in fully annotating the data set was the cost of training and paying the students for their annotation work. The annotators were instructed to draw boundaries around every object that definitely existed in the image and to separate plural objects into distinct components, and to *hallucinate* the extent of objects when the objects occulded each other. This annotation decision is important to make use of the overlapping relationships $\overrightarrow{\text{on}}$, etc. Figure 2.10 shows an example of hallucinating the extent of an object in an image. Further details on how the region annotations were performed can be found in Appendix A.

A total of 496 distinct labels were used to label regions. Figure 2.8 shows

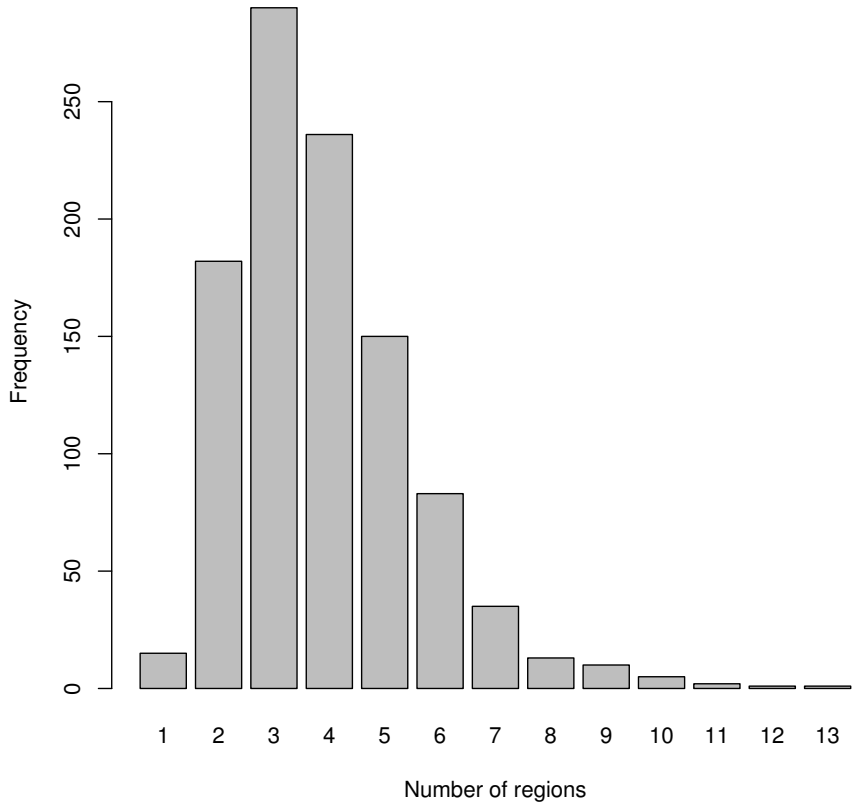


Figure 2.9 *Distribution of image region annotations in the data set.*

the distribution of the top 20 region annotations in the data; people-type regions are the most commonly annotated regions. Given the prevalence of labels referring to the same types of regions, we defined 26 sets of equivalent labels to reduce label sparsity (e.g., BIKE was considered equivalent to BICYCLE). This was done by ranking the annotation labels in order of frequency in the data set and inspecting the labels for *semantic* equivalence. In this case, labels were determined to be semantically equivalent if they undoubtedly referred to the same type of objects.

The normalization process reduced the size of the region label vocabulary from 496 labels to 362 labels. Inter-annotator agreement was 74.3% for region annotations; this was measured by computing polygon overlap over the annotated regions using the PASCAL VOC defition of overlap defined in Equation 2.1. In addition to the original polygon labels, we used one level of WordNet hypernyms to compute polygon overlap, such

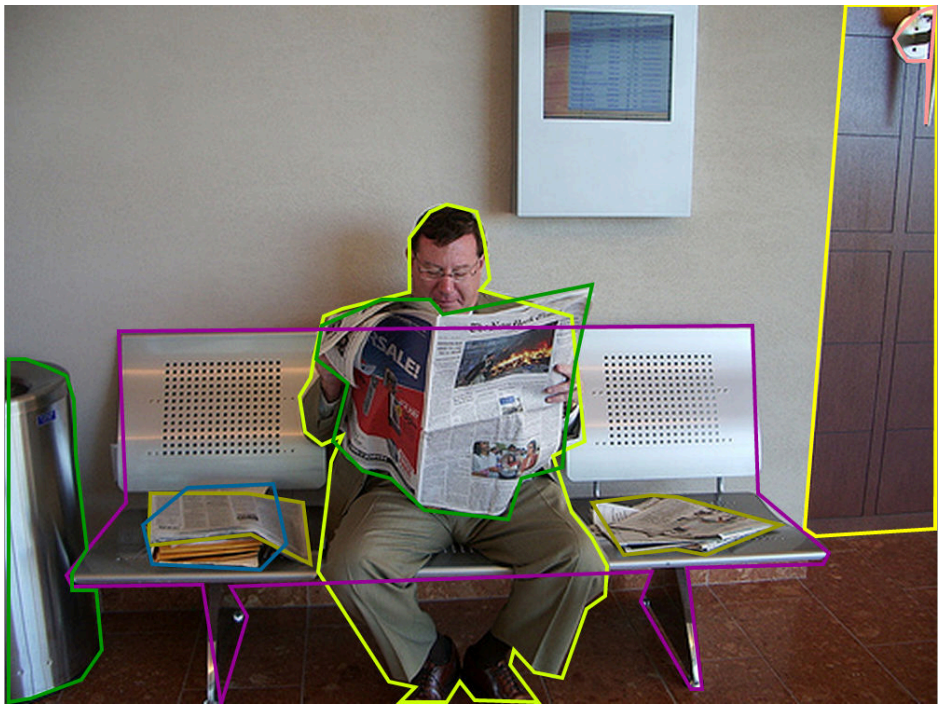


Figure 2.10 *An example where the annotator needed to hallucinate the extent of the bench occluded by the man.*

as man $\rightarrow$ {man, person}; furthermore, we also assumed that when region annotations overlaped by more than 95% that there was an overlap. This was necessary in situations where the annotators used the labels "road" and "street", which are distantly related in WordNet.

The size of the region label vocabulary is substantially larger than the number of classes typically used to design and evaluate object detection algorithms. In fact, the standard object detection evaluation programme is run on only 20 object classes, each of which contains thousands of labelled training examples. The performance of the best model at PASCAL 2012 has a mean average precision of 40.9%, which makes it too noisy a component in the entire pipeline of generating structured representations of images. For this reason, we decided to work with gold-standard object annotations to let us study the potential utility of the representation, instead of having to constantly compensate for an extremely noisy input component. Chapter 6 outlines an approach to incorporating automatic object detector output into the experiments presented in Chapters 3, 4, and 5.

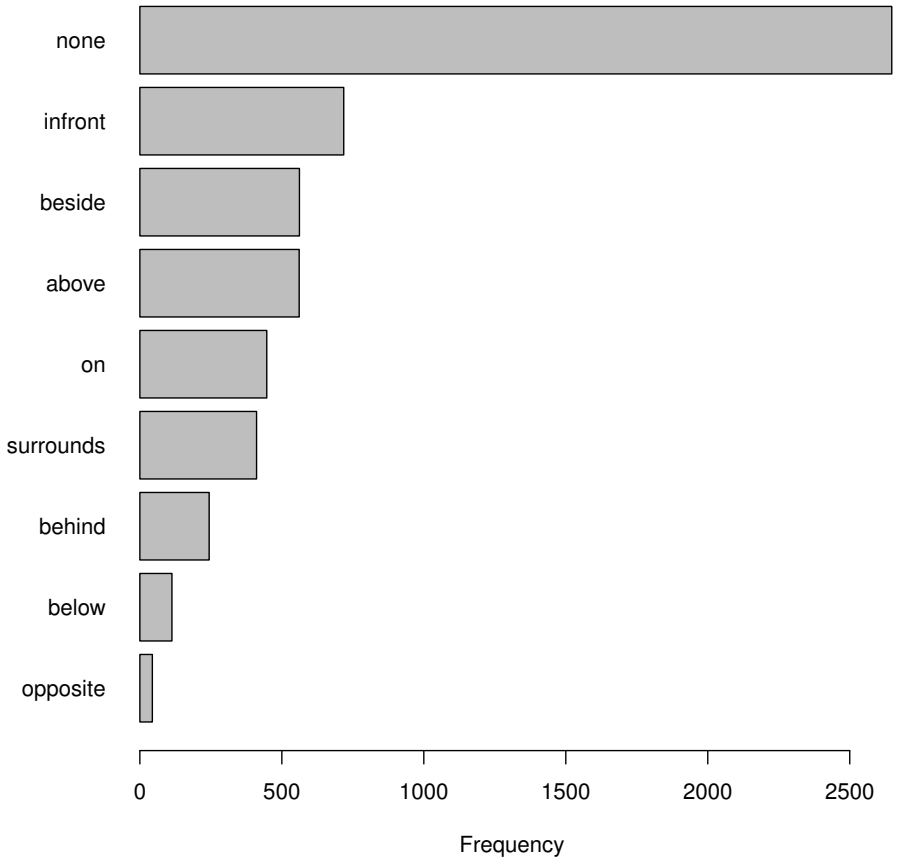


Figure 2.11 *Distribution of the spatial relations.*

Action Labels

The original PASCAL action recognition dataset contains ground truth action class annotations for each image. These annotations are in the form of labelled bounding boxes around the person performing the action in the image.

2.4.6 *Collecting Visual Dependency Representations*

We trained two Ph.D students to construct gold-standard Visual Dependency Representations for each image–description pair. The process for creating a visual dependency representation of an image is described in Section 2.3. There were three descriptions for each of the 341 region-annotated images, resulting in a set of 1,023 visual dependency represen-

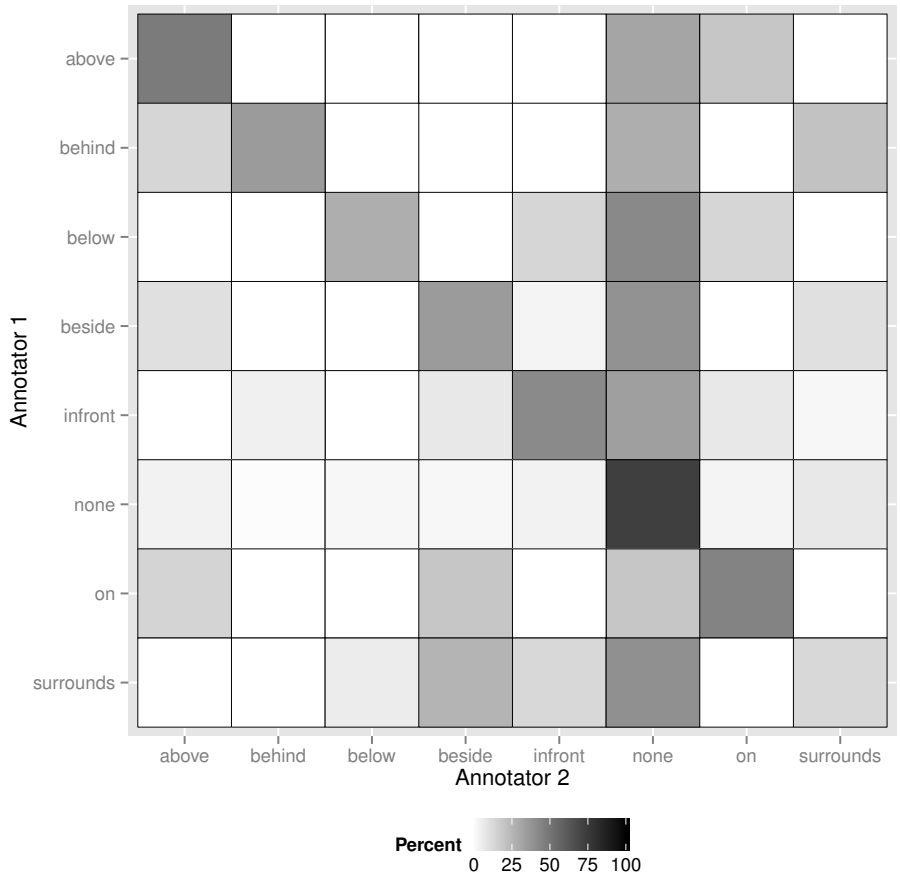


Figure 2.12 *Inter-annotator confusion matrix for VDR annotation. The largest source of disagreement between annotators was on whether to attach objects to the root node of a structure.*

tations. The annotated data set comprised a total of 5,748 spatial relations, corresponding to a mean of 4.79 ± 3.51 relations per image.

Figure 2.11 shows the distribution of spatial relation labels in the data set. It can be seen that the majority of regions are attached to the ROOT node, i.e., they have the relation label *none*. This property of the data set has a predictable effect on the performance of the models described in the next chapter.

Inter-annotator agreement on a subset of the data was measured at 84% agreement for labelled dependency accuracy and 95.1% for unlabelled dependency accuracy. This suggests the task of generating visual dependency representations can be performed reliably by human annotators. These figures also provide an upper bound on the performance of com-

putational prediction models presented in Chapter 3. Figure 2.12 shows a confusion matrix for the differences between annotators; it can be seen the largest source of disagreement is on whether to ROOT attach objects. There is also some disagreement in ABOVE / ON and BEHIND / SURROUNDS relationships.

2.5 CONCLUSIONS

This chapter introduced a novel structured representation of images, the Visual Dependency Representation. Visual Dependency Representations encode structure in an image by capturing the spatial relationships between the regions of an image. This representation is motivated by the observation that unstructured representations, such as a bag-of-words vector, have no means to distinguish images where objects co-occur from images where objects co-occur to represent an action.

We described the process for creating Visual Dependency Relationships and presented a new data set of images, paired with descriptions, object annotations, and the proposed Visual Dependency Representations. The distribution of adjectives, verbs, and object types was considered, and used to help justify the decision to not use automatic visual extraction techniques in the experiments presented later in the thesis.

The Visual and Linguistic Treebank data set provides the basis on which the remainder of the thesis is presented. In Chapter 3, we present an approach to automatically predicting Visual Dependency Representations from region-annotated images; in Chapter 4, we use Visual Dependency Representations of images as part of the image description task; and in Chapter 5, we use it to represent images in a query-by-example image retrieval task.

3

In this chapter we introduce the task of Visual Dependency Representation parsing. VDR Parsing is the automatic prediction of the Visual Dependency Representation of an image from its labelled region annotations and (optional) description. A discriminative graph-based dependency parser forms the basis of the approach. We present three variants of the parser, each of which progressively exploits more of the available data. The best-performance is found when the parsing model can exploit both the image and the descriptions when attempting to predict the Visual Dependency Representation. However, significant improvements are found when using either modality in isolation of the other. The VDR Parser presented in this chapter forms the foundation of automatic image structure prediction that is used in the subsequent chapters on image description and image retrieval.

3.1 INTRODUCTION

In Chapter 2 we presented an approach to modelling the structure of an image using the Visual Dependency Representation (VDR). This structured image representation captures the spatial relationships between regions of an image that contribute to explaining the portrayed event. In Chapter 2.3 we described how humans can create gold-standard representations given a collection of region annotations and a corresponding image description. We now turn our attention to the task of how to automatically predict the VDR of an image, which we frame as a parsing task and learn a parsing model from region-annotated images, optionally aligned with corresponding descriptions. Figure 3.1 presents an overview of how we learn the VDR parsing model. The MSTParser of McDonald et al. (2005a) is used as the basis of our approach; MSTParser is a discriminative graph-based parser that constructs a fully-connected graph over features extracted from the input. In natural language dependency parsing, the linear ordering of the input is typically used to help constrain

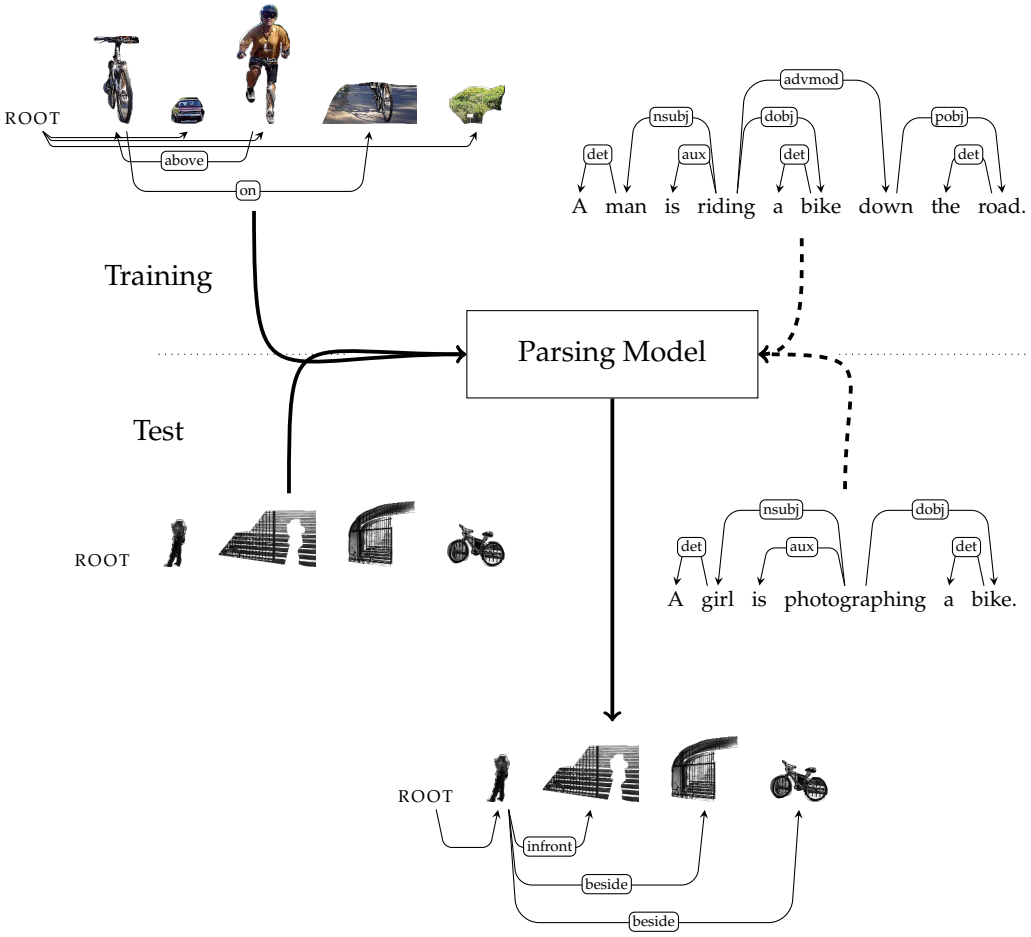


Figure 3.1 *An overview of the image parsing process. The parsing model is trained over human-annotated Visual Dependency Representations of images; at test time, the VDR of an unseen image is predicted over a collection of region annotations. We also experiment with a quasi-synchronous parsing model that can optionally exploit the syntactic structure of the descriptions associated with the images when training the model and then when predicting the structure of an image.*

and guide the feature exctraction process. There is no such linear ordering in image data, so we order the image regions by alphabetical region label and require that none of the features consider the linear ordering of the input data. The fully-connected graph is then pruned to a maximum spanning tree to produce the final dependency representation. The feature-based nature of the model makes it well-suited to this task because we can easily combine different sources of evidence from the input. The parsing model is used to predict the structure of a region-annotated image, optionally aligned to a corresponding image description.

We present VDR parsing results obtained using three feature sets of increasing complexity. The first set contains only features extracted from

the Visual Dependency Representations, the second set additionally uses image features, and the third set adds features extracted from the image descriptions. In order to use the image descriptions, we expand the parsing model to a quasi-synchronous dependency parser (Smith and Eisner, 2006, 2009) to exploit the structural correspondences between visual dependency representations and linguistic dependency representations computed over image descriptions.

In 10-fold cross validation experiments, we find significant improvements in Visual Dependency Representation prediction accuracy compared to a baseline that assumes there is no structure in an image. Parsing accuracy increases when we include features from the annotated regions, or using quasi-synchronous features extracted from the descriptions. The maximum performance we report is when the model exploits features from both the image and the description at the same time.

3.2 RELATED WORK ON STATISTICAL DEPENDENCY PARSING

Dependency grammar is a language structure formalism based on the notion that the tokens in a sentence are related by directed labelled dependency relations (Tesnière, 1953). The dependency structure of a simple sentence can be seen in the example in Figure 3.2, reproduced from Chapter 2.3. A dependency relation is formed between a *head* and its *argument*. As an example, consider the dependency relation between the head *riding* and the argument *man*: *man* is the nominal subject of the verb. This approach to representing the structure of a sentence differs from alternatives such as phrase-structure grammar in the sense that there is no attempt to form noun phrases or verb phrases as a hierarchical representation of the sentence structure.

Dependency parsing is the task of automatically assigning a dependency structure of a sequence of tokens in a sentence. There is a substantial body of work on automatic dependency parsing, ranging from unsupervised generative models (Klein and Manning, 2004) to supervised graph-based (McDonald et al., 2005b) and transition-based models (Nivre et al., 2004). A detailed treatment of dependency parsing is out-of-scope in this thesis, however, an excellent overview of supervised dependency parsing mod-

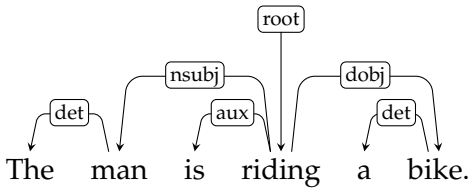
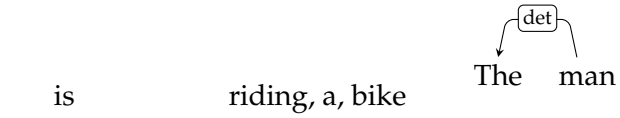


Figure 3.2 *An example of a dependency parsed sentence.*



(a) Stack (b) Buffer (c) Relations

Figure 3.3 *A transition-based parser in the middle of producing a dependency parse for the sentence in Figure 3.2.*

els can be found in Kübler et al. (2009). In the remainder of this section, we contrast and compare graph-based and transition-based parsing models, and how they could be applied to image parsing.

A transition-based dependency parser builds a dependency structure over a sentence by performing a series of *shift-reduce* operations on a sequence of tokens (Nivre et al., 2004). The principle components of such a parser are: the *stack*, which contains tokens that have been set aside for processing; the *buffer*, which contains tokens that have yet to be processed; and the set of proposed dependency *relations* between tokens in the sequence. The act of producing a dependency parse is achieved by performing four possible operations to manipulate the stack and buffer to produce dependency relations between tokens in the sentence. A **Left-Arc** adds a dependency relation from token w_j at the front of the buffer to w_i on top of the stack. **Right-Arc** adds a new dependency relation from token w_i on the stack to w_j in the buffer. **Reduce** remove token w_i from the top of the stack, and **Shift** moves the next token in the buffer to the top of the stack.

Figure 3.3 shows an example of a transition-based parser in the middle of parsing the example shown in Figure 3.2. The next optimal step for the parser is to perform a **Left-Arc** operation to add a new dependency relation between "is" on the *stack* and "riding" in the *buffer*. Deciding which transition operation to perform while processing a sequence of tokens is determined by learning the parsing model parameters over the

training data (Nivre et al., 2004).

The incremental nature of a transition-based dependency parser makes it unsuitable for the task of predicting the Visual Dependency Representations of images. There is no sense of a sequential order of regions in an image, and so a dependency parser than does not consider all possible hypotheses when constructing a dependency parse of an image will fail to construct a good Visual Dependency Representation. The alternative to shift-reduce dependency parsing is graph-based parsing. A graph-based dependency parser creates a dependency structure by constructing a fully-connected graph between tokens in the input. The graph is then selectively pruned to produce the Maximum Spanning Tree that defines the dependency parse of the input (McDonald et al., 2005b). This approach is well-suited to the problem of predicting Visual Dependency Representations because it can readily propose relationships between any of the tokens in the input. We will use this parsing strategy as the basis of our approach to Visual Dependency Representation parsing, as outlined in the next section.

3.3 GRAPH-BASED PARSING MODEL

A modified version of the MSTParser graph-based dependency parser of McDonald et al. (2005b) is used to predict Visual Dependency Representations. This parsing model is well-suited to the task because generates a fully-connected weighted graph over the input and then prunes the graph to the maximum spanning tree.

The MSTParser predicts the dependency structure y of a natural language sentence $\mathbf{x}$ by finding the structure y that maximises the score $s(\mathbf{x},y)$ of the sum of the edges (i, j) in the tree y:

$$s(\mathbf{x}, y) = \sum_{(i,j)\in y} \mathbf{w} \cdot \mathbf{f}(i, j) \tag{3.1}$$

where $\mathbf{f}(i, j)$ is a high-dimensional feature vector that represents the edge between tokens i and j in $\mathbf{x}$, and $\mathbf{w}$ is the weight vector corresponding to the features in $\mathbf{f}$. The feature types in the feature vector $\mathbf{f}$ can be any

evidence gathered from the input itself, such as the tokens of the input, or from any relevant external source.

The process of determining the best scoring y is governed by the feature functions that represent a directed edge between token i and j in y. In McDonald et al. (2005a), the features are calculated at three levels of granularity:

1. **Unigram.** Part-of-speech tags and surface forms.
2. **Bigram.** Part-of-speech tags and surface forms of head-argument pairs.
3. **Surrounding.** Part-of-speech tags and surface forms or the previous and next tokens in the input.

There is a fundamental incompatibility between the original definition of the **Surrounding** type of feature and the VDR parsing task. These features capture statistics of the order of the input, and it does not make sense to talk about the order of the input in image. Given this observation, we developed a different set of feature functions that can be used to extract relevant evidence from the input. These new feature types will be discussed in greater detail in the next section.

In the remainder of this chapter, we will use the following notation: $\mathbf{x}_{vis}$ is the collection of annotated regions and y_{vis} is a visual dependency representation of the image; (i, j) is a directed arc from region i to region j in y_{vis}, $\mathbf{f}(i, j)$ is a high-dimensional feature representation of the arc (i, j), and $\mathbf{w}$ is a vector of feature weights to be learned by the model. The overall score of a visual dependency representation is then computed as:

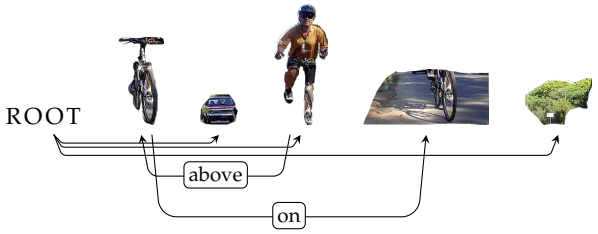
$$s(\mathbf{x}_{vis}, y_{vis}) = \sum_{(i,j) \in y_{vis}} \mathbf{w} \cdot \mathbf{f}(i, j) \tag{3.2}$$

3.4 INPUT AND FEATURES

The Visual Dependency Representation of an image pair is stored on disk in CoNLL-X format, an example of which can be seen in Figure 3.4. Each region is represented by its label, the centroid of the region, its parent

```
ID WORD  FEATS          HEAD REL
1  bike  187.42|343.76  4    above
2  car   135.22|330.01  0    -
3  man   197.61|228.17  0    -
4  road  171.23|394.78  1    on
5  trees 105.65|298.77  0    -
```

(a)



(b)

Figure 3.4 *An example of how the Visual Dependency Representation of an image is stored on disk in CoNLL-X format. The data representation is shown in (a), and the VDR is shown in (b). The FEATS field is used to store the centre of mass of the region in the original image, which allows features to be extracted from the image when training or testing the parser. Other useful information could be stored in this field, such as the certainty with which an object detection is made, or alternative labellings for the region. This figure is based on the example used in Figure 2.4c.*

region in the VDR, and the spatial relation between the parent and region. The FEATS field is used to store the coordinates of the centroid of each region, which is then used to enable fast spatial relationship calculations in the parser.

The feature functions used by the VDR parsing model vary based on the source from which the evidence is being collected. We define three groups of feature functions that are used in the model: VDR, IMG, and QDG. The final choice of which combinations of features to use was based on the performance of the parsing model on held-out development data.

VDR *feature set*

The features functions in the VDR feature set are defined over the region labels and edge labels in the visual dependency representation itself, as shown in Table 3.1. Because the VDR of an image does not have the same ordering effects as natural language, none of the features encode the linear

VDR feature set
head
arg
head arg

Table 3.1 VDR *feature set: head: label of the head region in the VDR; arg: label of the argument region. All three features are conjoined with the edge label between the head and the arg.*

order of the input (unlike the original McDonald et al. (2005a) parser, which operates over sentences, which are ordered sequences of words). The unigram features *head* and *arg* capture statistics about how regions appear as either heads or arguments in the VDR, and the bigram feature *head arg* captures which region labels are in head-argument relationships. We experimented with features that capture the number of arguments a head takes and the number of siblings an argument expects, but these decreased parsing accuracy and so are not present in the final model.

IMG *feature set*

We can add additional feature functions to the model by incorporating information from the image regions. We refer to the models that use these features with a +IMG suffix; the image features used are listed in Table 3.2. We extract five different types of features from the image regions: the position of the region in the image, the distance of the region from the centre, the size of the region, the distance between pairs of regions, and the spatial relationship between two regions standing in a head–argument relationship.

pos(·): The position of a region is defined as the quadrant in which the centre of mass of the corresponding polygon is located. Each image is split into four quadrants and the position of the region is generated from the quadrant it falls into.

dfc(·): The distance of the region from the centre (dfc) is found by calculating the Euclidean distance of the region (as defined by its centroid) from the centre of the image, normalised by the maximum possible euclidean distance from the centre to the edge of the image, and

IMG feature set
head spatial(head, arg)
arg spatial(head, arg)
head arg spatial(head, arg)
pos(head)
dfc(head)
dfc(head) spatial(head, arg)
size(head)
size(arg)
size(arg) spatial(head, arg)
dbr(head, arg)
dbr(head, arg) spatial(head, arg)

Table 3.2 IMG *feature set. See text for full details. size(·): the size of the region; pos(·): the quadrant the centroid of the region belongs to. dfc(·): the distance of the centroid of the region from the centre of the image; dbr(·, ·): the distance between the centroids of the regions; spatial(·, ·): the spatial relationship between the head and the argument. All features are conjoined with the label of the edge between the head and the argument.*

binned into 10% intervals.

dbr(·, ·): The distance between pairs of regions (dbr) is a generalised form of the (dfc) feature, where both end points of the line between the regions are regions themselves.

size(·): The size of the region (size) is calculated by counting all of the pixels inside the region and normalising it by the total number of pixels in the image, binned into 10% intervals.

spatial(·, ·): The spatial relationship between two regions (spatial) is calculated by taking the centroid of each region polygon and calculating the angle formed between the head and the argument polygon. The angle is then mapped onto the five bins: ON, ABOVE, BESIDE, BELOW, and NONE, using the definitions in Table 2.1 (which also use angles). These bins do not cover the full set of relations in the grammar, due to the challenge of accurately inferring 3D relations such as INFRONT or BEHIND from static images (Saxena et al., 2006).

It is possible to extend the parsing model by incorporating feature functions that extract evidence from the written description associated with the image. The key insight underlying this approach is that generating a visual dependency representation of an image is a form of tree-to-tree translation: we translate from a syntactic dependency tree to a visual dependency representation. Tree-to-tree translation can be formalized as parsing with a synchronous grammar (Wu, 1997; Chiang, 2005). An example is a synchronous context-free grammar (SCFG), a generalization of a context-free grammar. An SCFG generates pairs of syntactic trees (for example source language trees and target language trees). The paired trees are isomorphic, which means that for each rule-rewrite, only the order of the non-terminals of in the two trees can differ, not their identity.

SCFG and other synchronous grammar formalisms that generate isomorphic trees result in a tight coupling between source and the target trees. For the present task of pairing visual dependency representations and syntactic dependency trees, we require a more flexible approach, as the two types of representations can diverge significantly. We use the Quasi-synchronous Dependency Grammar (QDG) (Smith and Eisner, 2006, 2009) formalism, which allows for essentially arbitrary correspondences between two dependency representations, specified in terms of aligned tree configurations. The QDG model of Smith and Eisner (2006) can be used to induce a probabilistic grammar of a target language (in this case the VDR), given a set of source language parse trees, aligned on the word level. Quasi-synchronous grammars have found applications in parser adaptation (Smith and Eisner, 2009), paraphrasing (Das and Smith, 2009), and machine translation (Gimpel and Smith, 2009).

The scoring function in Equation (3.2) can be extended to operate over QDG representations instead of over standard dependency representations. The first sum in Equation 3.3 is identical to Equation 3.2. The second sum incorporates features that link arcs (i, j) in the Visual Dependency Representation with the syntactic dependency tree y_{text} of the image description $\mathbf{x}_{text}$, via an alignment configuration a, and $\mathbf{f_b}$ and $\mathbf{w_b}$ are the feature and weight vector for the QDG features.

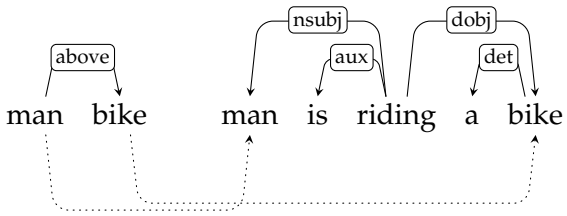
$$s(\mathbf{x}_{vis}, y_{vis}, \mathbf{x}_{text}, y_{text}, a) = \sum_{(i,j)\in y_{vis}} \mathbf{w_m} \cdot \mathbf{f_m}(i,j) + \sum_{(i,j)\in y_{vis}} \mathbf{w_b} \cdot \mathbf{f_b}(i,j,y_{text},a) \tag{3.3}$$

This model, referred to with a +QDG suffix, is trained over visual dependency representations and image descriptions. We follow Smith and Eisner (2009) and extend the McDonald et al. (2005b) parser with features that model the correspondences between two dependency representations, as described in the remainder of this section. The image descriptions are represented as syntactic dependency trees and aligned with the visual dependency representations on the word level. The alignments are between tokens in the description and region labels in the VDR. The alignments for the Quasi-synchronous model are word alignments (or, more precisely, region label-to-word alignments), which we computed using a simple lexical matching algorithm, which performs a string comparison between the labels of the regions in a VDR and the tokens in the description. Region labels were normalized (see Section 2.4.5) and words in the description were stemmed. Furthermore, the matching process was augmented with a WordNet hypernym lookup to increase the likelihood of matches. This process involved taking the region labels and description tokens, finding all of their synonyms in the respective WordNet synsets, and then for each synonym, taking its hypernym and determining whether the region label and the descripionn token share any lexically identical hypernyms. This process was useful for matching regions like *girl* and *woman* through the hypernym *person*. Essentially these labels refer to the same type of object but it has been expressed in a different way by different workers on Mechanical Turk.

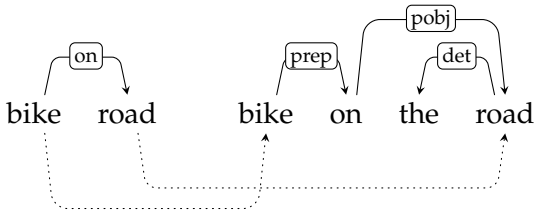
The features linking visual and syntactic dependency representations are listed in Table 3.3. The quasi-synchronous features express additional information about the relationships between a pair of regions by capturing the syntactic configuration of the words in the aligned image description. The alignment configurations that connect pairs of aligned nodes in the visual and syntactic representations are always parent-child on the visual side, but we allow the following configurations on the description side:

QDG feature set
head config(head, arg)
arg config(head, arg)
head arg config(head, arg)
head arg verb(head, arg)
head arg config(head, arg) verb(head, arg)

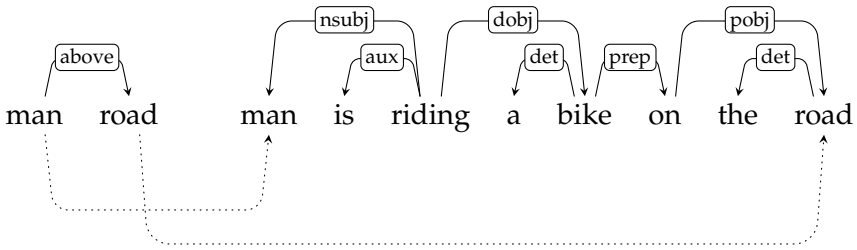
Table 3.3 QDG *feature set. config(·, ·): the syntactic configuration of the two words in the syntactic dependency tree (see text). verb(·, ·): the verb on the path between the two words in the syntactic dependency tree.*



(a) Siblings configuration



(b) Ancestor-descendant configuration



(c) C-command configuration

Figure 3.5 *Example of the siblings,ancestor-descendant, and c-command alignment configurations in our data, taken from Figure 2.4. The image labels are shown on the left and the image description is shown on the right. In the siblings example (a), it can be seen that* MAN *is the parent of* BIKE *with the spatial relation above when the words are in a siblings relationship in the description.*

parent-child, child-parent, siblings, ancestor-descendant, and c-command. Figure 3.5 shows examples of these configurations in our data. In (a), the feature function templates extract features that capture the spatial relationship between the *man* and *bike* when the *man* is the subject and the *bike* is the object of the verb *to ride*; (b) captures the prepositional relationship between the *bike* and the *road*. (c) captures the relationship between the c-command relationship between the *man* and the *road* in the description. The c-command relationship has been described as capturing the "relative position of constituents in a parse tree" (Radford, 2004). In our case, it is useful for capturing deeper syntactic structure in the description that is often formed through prepositional phrases.

Finally, we can add the image features from VDR+IMG to this model to obtain fully multi-modal VDR Parser, which we refer to as VDR+IMG+QDG. This parser is trained over the region-annotated images with Visual Dependency Representations and aligned syntactic dependency trees over image descriptions.

3.6 EXPERIMENTS

We now present results for the VDR parsing models introduced in this chapter. The VDR and VDR+IMG models are trained on data annotated with region boundaries, region labels, and Visual Dependency Representations. The VDR+QDG and VDR+IMG+QDG models have access to image descriptions, which are annotated with syntactic dependency trees and word-aligned with the visual dependency representations.

The task for all models is to take an image with gold-standard labelled region boundaries and predict the correct Visual Dependency Representation. The quasi-synchronous models VDR+QDG and VDR+IMG+QDG have access to image descriptions at test time.

3.6.1 *Data and Evaluation Measures*

The VDR parsing models are evaluated on the 1,023 Visual Dependency Representations in the Visual and Linguistic Treebank, described in Chapter 2. The parsing experiments in this chapter, and the subsequent experi-

ments on language generation (Chapter 4) and image retrieval (Chapter 5), were run over the same 10 randomly generated splits of the data set into 80% training, 10% development, and 10% test data. Recall from Chapter 2.4 that each image is associated with three descriptions and each image–description pair has its own Visual Dependency Representation. The data was split into training/development/testing splits according to each image, not according to each Visual Dependency Representation, to avoid the models being tested on a fraction of the training data if it were split without this restriction.

The performance of each model is measured using *labelled* (Equation 3.4) and *unlabelled* (Equation 3.5) directed attachment accuracy, with statistically significant differences calculated using a dependent t-test. The unlabelled accuracy indicates how well a model can predict which objects should be in an interacting relationship with each other. It is a less conservative measure than labelled accuracy, which requires both the arcs and the arc edges to be correct. We will see in the chapters on language generation and image retrieval that either form of Visual Dependency Representation can be useful.

$$\text{Labelled accuracy} = \frac{\text{\# regions with the correct parent and arc label}}{\text{total number of regions}} \tag{3.4}$$

$$\text{Unlabelled accuracy} = \frac{\text{\# regions with the correct parent}}{\text{total number of regions}} \tag{3.5}$$

We also distinguish *root attachment accuracy*, the proportion of regions that correctly attach to the root node, and *non-root attachment accuracy*, the proportion of regions that correctly attach to any other node. This distinction is useful for determining how well the models relate image regions to each other. Recall from Chapter 2.4 that more than 50% of the data contains root attachments, which makes a naive most-frequently-observed baseline easy to confuse with useful VDR parsing.

	Unlabelled Accuracy			
	Mean	Root	Non-root	Features
FLAT	49.0 ± 3.0	100 ± 0.0	0.0 ± 0.0	N/A
VDR	$61.9 \pm 4.5^{\star}$	88.3 ± 3.8	$36.6 \pm 5.8^{\star}$	3,800
VDR+IMG	$64.2 \pm 4.7^{\dagger}$	89.4 ± 2.8	$40.3 \pm 6.0^{\dagger}$	7,500
VDR+QDG	$65.0 \pm 4.5^{\dagger}$	87.8 ± 3.5	$43.2 \pm 5.6^{\dagger}$	9,500
VDR+IMG+QDG	$66.1 \pm 4.3^{\dagger}$	90.2 ± 2.1	$43.1 \pm 5.5^{\dagger}$	13,500

Table 3.4 *Unlabelled directed image parsing results. The metrics are root and non-root attachment accuracy and their mean. Incorporating features from the image (*VDR+IMG*) or an aligned description (*VDR+QDG*) improves accuracy, however, the best results are achieved when combining features from both modalities (*VDR+IMG+QDG*). *: significantly different compared to* FLAT *at* $p < 0.05$*; †: siginficantly different compared to* VDR *at* $p < 0.05$.

3.6.2 *Baselines*

Predicting the VDR of an image is a new task and there are no obvious baselines against which we can compare the performance of the models presented in this chapter. We propose FLAT as a baseline model that attaches every region label to the root node of the image; it does not need to be trained and simply operates over the set of labelled regions. It is based on the assumption that a bag-of-regions representation is sufficient and no information extracted from the image or the description is useful in understanding the relationships between regions.

3.6.3 *Results*

Tables 3.4 and 3.5 summarise the unlabelled and labelled parsing performance of the different models on the VDR parsing task. We start by noting that the performance of the FLAT baseline can be explained by the nature of the second sentence of the image descriptions: it is often a list of salient but unrelated regions attached to the root node of the representation. This model is equivalent to applying the most frequently occurring relationship in the data set – none, as described in Figure 2.11 in Chapter 2.

On the unlabelled directed accuracy measure, the VDR Parser with only

	Labelled Accuracy			
	Mean	Root	Non-root	Features
FLAT	49.0 ± 3.0	100 ± 0.0	0.0 ± 0.0	N/A
VDR	$54.2 \pm 4.6^{\star}$	88.3 ± 3.8	$21.5 \pm 4.4^{\star}$	3,800
VDR+IMG	$55.2 \pm 4.8^{\star}$	89.4 ± 2.8	$22.6 \pm 4.6^{\star}$	7,500
VDR+QDG	$55.0 \pm 4.1^{\star}$	87.8 ± 3.5	$23.5 \pm 3.4^{\star}$	9,500
VDR+IMG+QDG	$56.0 \pm 4.0^{\star}$	90.2 ± 2.1	$23.3 \pm 3.6^{\star}$	13,500

Table 3.5 *Labelled directed image parsing results. The metrics are root and non-root attachment accuracy and their mean. Incorporating features from the image (*VDR+IMG*) or an aligned description (*VDR+QDG*) improves accuracy, however, the best results are achieved when combining features from both modalities (*VDR+IMG+QDG*). *: significantly different compared to* FLAT *at* $p < 0.05$.

the VDR feature set is significantly better than assuming no structure in the image (FLAT). When we extract features directly from the annotated image regions (VDR+IMG), we observe another significant improvement in parsing accuracy compared to using only the VDR feature set. And if we adopt the Quasi-synchronous parsing model (VDR+QDG), we observe an independently significant improvement over only using features from the structured representations. The best performing parser is found when we extract features from the image and from the aligned description (VDR+IMG+QDG).

If we compare the parsing models on labelled dependency accuracy, we see the same pattern of performance improvements, albeit with less absolute improvements in accuracy. This drop in accuracy is expected because it is harder to get both the dependency attachments and the attachment label correct than only get the dependency attachments.

The number of features extracted by the feature functions defined in Chapter 3.4 can be seen in Table 3.5. We note here that the complexity of the parsing algorithm does not change as more features are added to the model. However, the models that exploit evidence from the image regions take the longest time to train because it is computationally expensive to extract features from images.

We plotted a confusion matrix of labelled parsing predictions to better understand the types of mistakes the VDR Parser makes when labelling

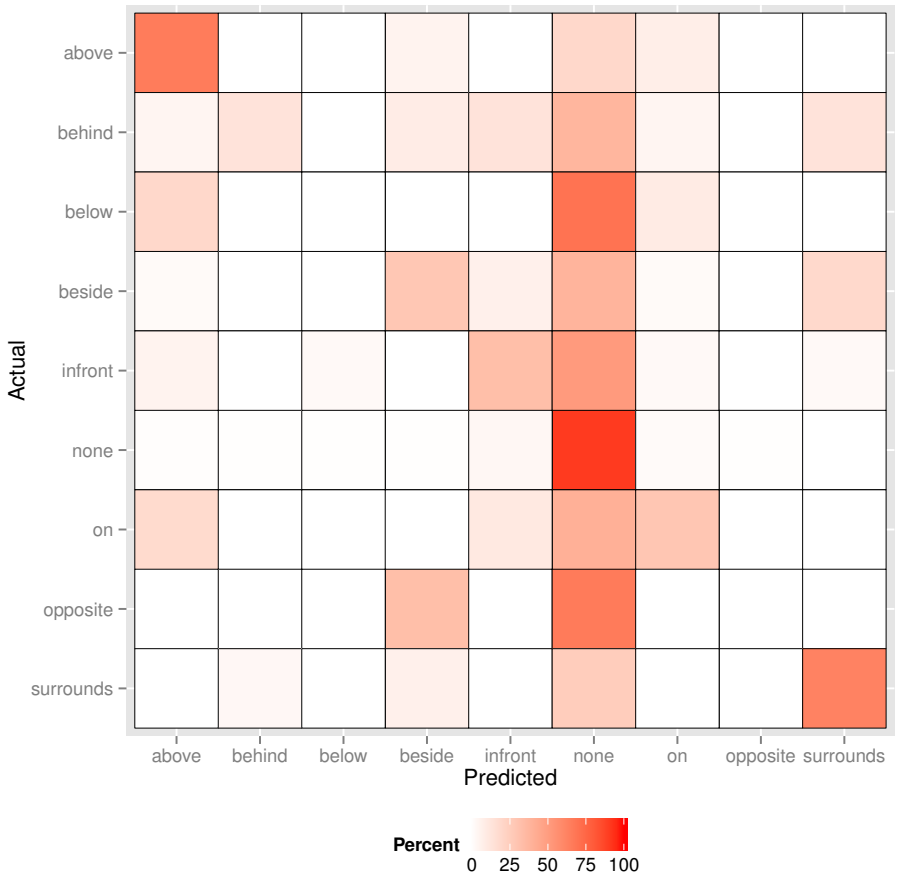


Figure 3.6 *Confusion matrix of the* VDR+IMG+QDG *model. It can be seen that* ROOT *attachment is over-predicted and there is some confusion between the* BESIDE *and* OPPOSITE*, and* ABOVE *and* ON.

the arc between a pair of regions, as shown in Figure 3.6. The confusion matrix was constructed from the output of the best performing parser, VDR+IMG+QDG. It is clear to see that root attachment (attaching an arc with no label to the ROOT node) is over-predicted by the parser, which can be explained by recalling that 50% of the attachments in the data set are root attachments. We also observe that there is some confusion between *above* and *on*, which is understandable given the fairly subtle difference between these two relations. The main difference between the human inter-annotator disagreements and the parser disagreements is that the VDR parser is much more likely to over-predict ROOT attachment. The interested reader can inspect the differences in more detail by comparing Figure 3.6 to the confusion matrix for inter-annotator agreement in Figure 2.12.

	Unlabelled Accuracy			
	Mean	Root	Non-root	Features
BOTH	66.0 ± 4.3	90.3 ± 2.1	42.9 ± 5.0	9,500
FIRST	65.8 ± 3.8	89.7 ± 2.3	43.0 ± 4.4	8,500
SECOND	$63.2 \pm 4.4^{\star}$	87.8 ± 4.4	$39.8 \pm 3.1^{\star}$	6,500
	Labelled Accuracy			
	Mean	Root	Non-root	Features
BOTH	55.9 ± 3.8	90.3 ± 2.1	23.1 ± 3.4	9,500
FIRST	56.0 ± 4.0	89.7 ± 2.3	23.7 ± 3.3	8,500
SECOND	54.1 ± 3.9	87.8 ± 4.4	21.8 ± 3.1	6,500

Table 3.6 *The Effect of Alignment Coverage on parsing accuracy in the* QDG+IMG+VDR *model. There is no significant difference between using alignments from only the* FIRST *sentence or* BOTH SENTENCES*, but using alignments from only the* SECOND *sentence significantly decreases parsing accuracy. *: significantly different compared to using both sentences at* $p < 0.05$.

3.6.4 *Alignment Coverage in the Quasi-synchronous Models*

The image descriptions in our data set consist of two sentences, where the first describes the action depicted, and the second describes regions unrelated to the action. It is conceivable that the first and second sentences differ in structural complexity and thus give rise to differences in parsing accuracy in the +QDG models. We investigate this by using alignment features from only the first sentence, only the second sentence, and from both sentences of the image description. We trained a non-projective VDR+IMG+QDG model to study the role of alignment feature coverage. Table 3.6 shows the effect of reducing the alignment coverage in the training data. It can be seen that there are no significant differences between using alignments in BOTH or only the FIRST sentence. We do observe significant decreases in performance when using alignments from only the SECOND sentence. This can be explained by remembering that we asked workers on Mechanical Turk to describe the objects involved in the action in the first sentence; the objects referred to in the second sentence are therefore background or contextual objects.

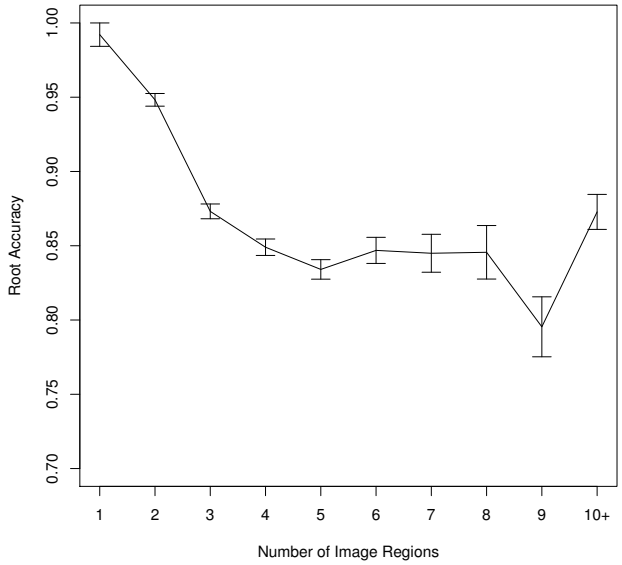
3.6.5 *The Number of Annotated Image Regions*

In VDR parsing, the number of labelled image regions can be thought of as analogous to sentence length, which has been shown to affect the accuracy of dependency parsers (Mcdonald and Nivre, 2011). We therefore explored how the number of labelled image regions affects the accuracy of the QDG-VDR model. Figure 3.7 shows root and non-root accuracy after binning the test images by number of image regions. It can be seen that root attachment accuracy plateaus after five region labels per image (though the variance increases, as later bins contain fewer data points). Non-root attachments, however, become increasingly difficult as the number of region labels increases. This mirrors the effect of sentence length found in language parsing.

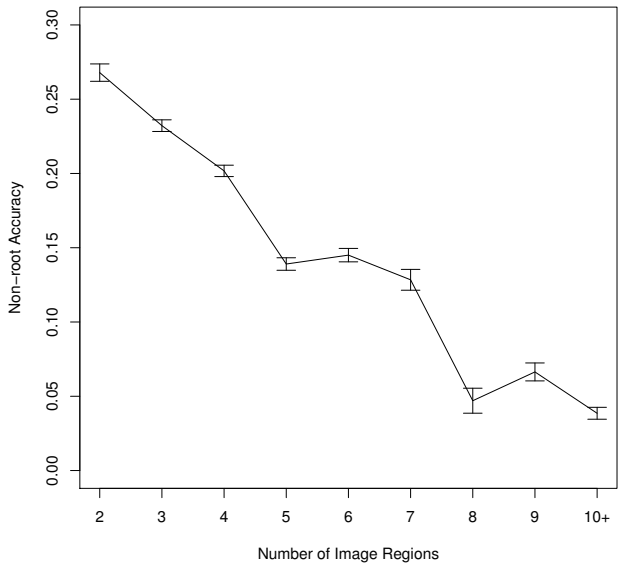
3.7 DISCUSSION

In a series of experiments, we found that a VDR Parser trained over only the region labels and the relationships between the regions (VDR) was better at predicting Visual Dependency Representations than assuming the image had no structure at all (the FLAT model). The implication of this finding is that if we could find a perfect automatic image tagger, then the VDR Parser would be sufficient to predict the relationships between the detected objects. It follows that if we could find a perfect object detector, then we could predict significantly better image structures than only using automatically predicted labels, as shown in the models with a +IMG suffix. We can extract simple image features, such as the position, size, and distance between regions in the image can be extracted from annotated regions. These additional features significantly improve the accuracy of the VDR Parser over using only the labels and relationships between the labels.

However, a perfect image tagger or object detector does not exist and it remains an open problem about how to integrate such a noisy automatic input into the VDR parsing process. We will discuss one method of integrating the output of an automatic object detector in Chapter 6 of this thesis.



(a) Root attachment accuracy



(b) Non-root attachment accuracy

Figure 3.7 *Labeled parsing accuracy of* VDR+IMG+QDG *by number of image regions. Root attachment accuracy (a) stabilises from five image regions; whereas non-root accuracy (b) decreases as the number of regions increases.*

The use of verbs and syntactic configurations in aligned image descriptions provides an orthogonal and equally significant improvement in parsing accuracy to the models with a +QDG suffix. This finding is promising, especially considering the growing prevalence of captioned images on the internet as potential source data.

The relative differences in labelled and unlabelled parsing accuracy improvements in Tables 3.4 and 3.5 shows that predicting both the relationship and the type of the relationship is a difficult problem. One direction for improving parsing accuracy would be to implement a new decoder for the labelled parsing features in the model. The current decoder, inherited from the MSTParser implementation for natural language dependency parsing, only constructs and decodes features for labelled parsing using unigram evidence. It may be possible to extend this to a more complex decoder, which would require a significant amount of engineering work. In whichever way, we will see in Chapters 4 and 5 that we observe significant improvements in image description and image retrieval with the current decoder implementation. Another option for improving parsing accuracy is to obtain more training data, which would obviously result in a more robustly trained parsing model.

3.8 CONCLUSIONS

In this chapter we showed how to automatically predict the Visual Dependency Representation of an image using a statistical dependency parser. The basis of our approach was a state of the art graph-based parsing model, which made it possible to extract evidence from both the visual data, in the form of the image regions, and the corresponding linguistic data, in the form of the descriptions.

We presented a series of experiments on the effect of extracting and combining different sources of evidence into the VDR Parser. Orthogonal improvements were found from using either visual or linguistic features. The improvements due to visual features were due to capturing information about the locations of regions in the image and the 2D spatial relationships between the regions in the image data. The linguistic features captured information about the types of verbs that appeared between

words in aligned image descriptions.

The main experimental results were supplemented by a series of ablation experiments to understand the effect of different sources of evidence in the parsing model. We found that parsing accuracy was significantly affected by projective dependency parsing, the number of alignments used in the quasi-synchronous models, and the relationship between the number of annotated image regions and the accuracy of the predicted structure.

In Chapter 4 we will show how automatically predicted Visual Dependency Representations can be used to generate significantly better descriptions of what is happening in an image. And in Chapter 5, we will show how the Visual Dependency Representation can be used to significantly improve example-based image retrieval models.

4

Describing the main event of an image involves identifying the depicted objects and predicting the relationships between those objects. Previous approaches have represented images as unstructured bags of regions, which makes it difficult to accurately predict meaningful relationships between regions. In this chapter, we show how Visual Dependency Representations can be used to improve automatic image description. We test this hypothesis using the data set of region-annotated images, associated with Visual Dependency Representations and gold-standard descriptions introduced in Chapter 2. We describe two template-based description generation models that operate over visual dependency representations. In an image description task, we find that these models outperform approaches that rely on object proximity or corpus information to generate descriptions on both automatic measures and on human judgements.

4.1 INTRODUCTION

Humans are readily able to produce a description of an image that correctly identifies the objects and actions depicted. Automating this process is useful for applications such as image retrieval, where users can go beyond keyword-search to describe their information needs; caption generation for improving the accessibility of existing image collections; story illustration; and in assistive technology for blind and partially sighted people. Automatic image description presents challenges on a number of levels: recognizing the objects in an image and their attributes are difficult computer vision problems; while determining how the objects interact, which relationships hold between them, and which events are depicted requires considerable background knowledge.

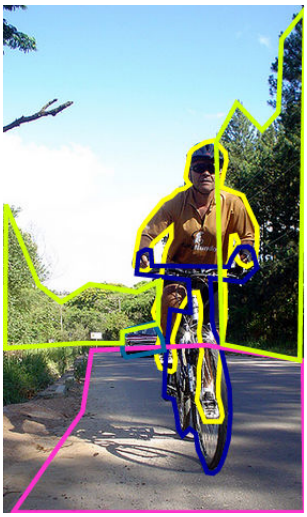
Previous approaches to automatic description generation have typically tackled the problem using an object recognition system in conjunction with a natural language generation component based on language models or templates (Kulkarni et al., 2011; Li et al., 2011). Some approaches have

utilised the visual attributes of objects (Farhadi et al., 2010), generated descriptions by retrieving the descriptions of similar images (Ordonez et al., 2011; Kuznetsova et al., 2012), relied on an external corpus to predict the relationships between objects (Yang et al., 2011), combined sentence fragments using a tree-substitution grammar (Mitchell et al., 2012), or frame the problem as ranking the descriptions that co-occur with text (Hodosh et al., 2013).

A common aspect of existing work is that an image is represented as a bag-of-terms. Bags-of-terms encode information about which objects co-occur in an image, but they are unable to express how the regions relate to each other, which makes it difficult to describe what is happening. As an example, consider Figure 4.1 (a), which depicts a man riding a bike. If the man was instead repairing the bike, then the bag-of-regions representation would be the same, even though the image would depict a different action and would have to be described differently. This type of co-occurrence of regions indicates the need for a more structured image representation; an image description system that has access to structured representations would be able to correctly infer the action that is taking place, such as the distinction between repairing or riding a bike, which would greatly improve the descriptions it is able to generate.

In this chapter, we use the Visual Dependency Representations introduced in Chapter 2 to encode the structure of images. This representation captures the spatial relations between regions of an image. An example can be found in Figure 4.1 (c), which depicts the VDR for Figure 4.1 (a). This VDR captures the notion that MAN is $\overrightarrow{\text{above}}$ the BIKE, and that the BIKE is $\overrightarrow{\text{on}}$ the ROAD. These relationships make it possible to infer that the man is riding a bike down the road, which corresponds to the first sentence of the human-generated image description in Figure 4.1 (b).

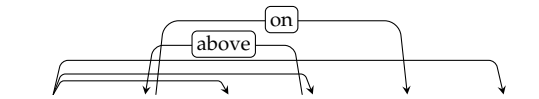
In order to test the hypothesis that structured image representations are useful for description generation, we present a series of template-based image description models. Two of these models are based on approaches in the literature that represent images as bags of regions. The other two models use Visual Dependency Representations, either on their own or in conjunction with gold-standard image descriptions at training time.



(a)

A man is riding a bike down the road.
A car and trees are in the background.

(b)



(c)

Figure 4.1 *(a) An image of a man riding a bicycle. (b) A human-written description of the image. (c) A Visual Dependency Representation of the image, given the annotated regions and the description. Reproduced from Chapter 2 for ease-of-access.*

Prior to presenting the results of the image description experiment, we estimate the correlation of five automatic evaluation measures with human judgements. The automatic evaluation measures are adopted from the Machine Translation and document summarisation communities. We compare unigram BLEU and Smoothed BLEU (Papineni et al., 2002), ROUGE-SU4 (Lin and Och, 2004), TER (Snover et al., 2006), and Meteor (Denkowski and Lavie, 2011), and find that Meteor has the best correlation with human judgements of semantic correctness, and that Smoothed BLEU and ROUGE-SU4 are moderately correlated with judgements. In the image description experiment we find that descriptions generated using the VDR-based models are significantly better than those generated using bag-of-terms models in automatic evaluations and in human judgements.

Finally, we also show that the benefit of the visual dependency representation is maintained when image descriptions are generated from automatically parsed VDRs. We use the image parser introduced in Chapter

	Transfer	Template	Ranking
Farhadi et al. (2010)	✓		
Yang et al. (2011)		✓	
Kulkarni et al. (2011)		✓	
Li et al. (2011)		✓	
Ordonez et al. (2011)	✓		
Mitchell et al. (2012)		✓	
Kuznetsova et al. (2012)		✓	
Hodosh et al. (2013)			✓

Table 4.1 *An overview of existing approaches to image description. The literature has been categorised into approaches that transfer a description from an existing corpus, approaches that use some form of template to guide the language generation process, and framing the task as ranking the descriptions that co-occur with images.*

3 to predict VDRs over a set of annotated object regions. This result reaffirms the potential utility of this representation as a means to describe events in images. Note that throughout this chapter we work with gold-standard region annotations; this makes it possible to explore the effect of structured image representations independently of automatic object detection.

4.2 RELATED WORK ON AUTOMATIC IMAGE DESCRIPTION

In this section, we summarise existing approaches to image description generation in the literature. Table 4.1 provides an overview of the approaches, which we have broadly categorised approaches as based on transferring the description from the training or test data (Farhadi et al., 2010; Ordonez et al., 2011; Kuznetsova et al., 2012; Hodosh et al., 2013), based on pre-defined templates - either explicit templates (Yang et al., 2011; Kulkarni et al., 2011), n-gram phrase combination (Li et al., 2011; Kulkarni et al., 2011), or tree substitution grammar (Mitchell et al., 2012).

Farhadi et al. (2010) generate descriptions of images by projecting images and corresponding descriptions into a shared "meaning space". The shared meaning space is represented by an (object, action, scene) triplet, where each variable in the triplet is drawn from a predefined set of values.

The most likely triplet is then calculated from an image, given a set of linearly weighted feature functions over data drawn from the image itself. A description of an unseen image is generated by transfering the description of an image that most closely matches the meaning representation tuple predicted for the image. This approach is evaluated using a readability analysis using two subjects on the UIUC PASCAL Sentences data set.

Yang et al. (2011) also generate descriptions of images from an abstract meaning representation. In their work, an image is represented as a (nouns, verb, scene-type, preposition) tuple. The nouns are determined by running state of the art object detectors of the image (Felzenszwalb et al., 2010); the verb is determined by predicting the most likely verb that relates pairs of detected objects in a dependency parsed representation of the Gigaword corpus (Napoles et al., 2012); the scene-type is determined using a gist scene-type detector (Oliva and Torralba, 2001); and the preposition is extracted from prepositional phrases in the training corpus. Sentences are then generated from a predicted meaning representation tuple by filling in slots in a sentence template. The evaluation is performed with ROUGE-1 and a relevance/readability analysis on Mechanical Turk using the UIUC PASCAL Sentences data set.

Kulkarni et al. (2011) generate descriptions of images using n-gram language models or sentence templates. Images are represented by running parts-based object detectors and stuff detectors of the image data. Attributes of these detections are then detected, and pairwise spatial prepositions are calculated between detected objects and stuff. A Conditional Random Field is used to predict the best possible labelling for an image, given the detections. The language generation approach is based on either combining n-grams or using "linguistically motivated" constraints on a template-based approach. Sentences are generated by taking the labelling predicated for an image and generating the description. This approach is evaluated with unigram BLEU and clarity / grammaticality judgements from two subjects using the UIUC PASCAL Sentences data set.

Li et al. (2011) generate descriptions of images by selecting and then fusing phrases of text from a very large text corpus, given the set of detections for an unseen image. The phrase selection stage deliberately over

selects phrases from the corpus and the phrase fusion stage attempts to optimally combine these phrases to create a relevant and grammatically correct description. Image features are extracted using a parts-based object detector and stuff detectors. Each of the extracted image regions are also processed to extract visual attributes and pairwise spatial relationships are calculated between detections. These features are then combined into a meaning representation as a set of ((adjective, noun), preposition, (adjective, noun)) tuples for the detections in the image. Descriptions were generated by extracting three phrases for each tuple in the meaning representation from the Google Web 1T corpus: one with the first object in the tuple, one with the second object in the tuple, and one with the relation between the objects. These sets of phrases are then combined to create a maximally compatible description of an image. The approach is evaluated using unigram BLEU and creativity / fluency / relevance judgements with two judges using the UIUC PASCAL Sentences data set.

Ordonez et al. (2011) generate descriptions of images by transferring the description of an image from their large training data set that most closely matches the unseen image. Images are processed to extract objects, stuff, people, and scene types. These extracted features from the image are then compared against images in the 1 million captioned photos data set, and the caption for the most similar image is literally transferred to describe the new image. The approach is evaluated with unigram BLEU using the SBU Captioned Photo Dataset.

Mitchell et al. (2012) generate descriptions of images by combining phrase structures in a tree-substitution grammar framework. The visual feature extraction is based on that of Kulkarni et al. (2011). Syntactic information, such as common modifiers, determiners, and verbs are calculated from the captions in the SBU Captioned Photo Dataset. Evaluation is performed using Amazon Mechanical Turk on the SBU Captioned Photo Dataset.

Kuznetsova et al. (2012) generate descriptions of images by retrieving and combining phrases from a large corpus of images and corresponding descriptions. Object detectors are used to find potentially relevant noun phrases, the parallel collection of descriptions are used to find verb phrases, stuff detectors are used to find scene description phrases. The

	Regions	VDR	External Corpus	Parallel text
PROXIMITY	✓			
CORPUS	✓		✓	
STRUCTURE	✓	✓		
PARALLEL	✓	✓		✓

Table 4.2 *The data available at training time to the language generation models compared in this Chapter.*

selected phrases are then combined in an integer linear programming framework to select the best combination of phrases and realised by adhering to linguistic, discourse, and consistency constraints in language. The approach is evaluated on a subset of the SBU Captioned Photo Dataset for which the computer vision components are found to be relaible using unigram BLEU and human judgements.

Hodosh et al. (2013) present an alternative approach to the problem by arguing that the task should be framed as a cross-modal ranking problem. They address the cross-modal ranking problem by projecting image and text representations into a shared semantic space using kernel canonical correlation analysis (Bach and Jordan, 2002). The parameters of the projection weights are estimated on training data of images paired with descriptions. The estimated shared semantic space is used to generate a ranked list of either the best descriptions for an image, or the best images for a description. Images are represented as vectors of colour, texture, shape, and Spatial Pyramid Kernel SIFT features (Lazebnik et al., 2006), and the descriptions are represented as a trigram kernel with distributional similarity matching between tokens. The tasks evaluated are sentence retrieval given an image, and image retrieval given a sentence in the Flickr8K data set. This approach can address both the image search and language generation problems, but it does so without addressing the problems of "traditional" natural language generation.

4.3 IMAGE DESCRIPTION MODELS

We present four template-based models for generating image descriptions in this section. Table 4.2 presents an overview of the information available to each model at training time, ranging from only the annotated regions of

an image to using Visual Dependency Representation of an image aligned with the syntactic dependency representation of its description. At test time, all models have access to image regions and their labels, and use these to generate image descriptions. Two of the models also have access to VDRs at test time, allowing us to test the hypothesis that image structure is useful for generating good image descriptions.

The aim of each model is to determine what is happening in the image, which regions are important for describing it, and how these regions relate to each other. Recall that all the images in the data set from Chapter 2.4 depict actions, and that the gold-standard annotation was performed with this in mind. A good description therefore is one that relates the main actors depicted in the image to each other, typically through a verb; a mere enumeration of the regions in the image will not be sufficient. All of the models compared in this chapter attempt to generate a two-sentence description, as per the gold standard descriptions.

In the remainder of this section, we will use Figure 4.1 (a) as a running example to demonstrate the type of language each model is capable of generating. All models share the set of language generation templates in Table 4.3.

4.3.1 PROXIMITY

PROXIMITY is based on the assumption that people describe the relationships between regions that are near each other. It has access to only the annotated image regions and their labels.

Region–region relationships that are potentially relevant for the description are extracted by calculating the proximity of the annotated regions. Here, o_i is the subject region, o_j is the object region, and s_{ij} is the spatial relationship between the regions. Let $R = \{(o_i, s_{ij}, o_j), \ldots\}$ be the set of possible region–region relationships found by calculating the nearest neighbour of each region in Euclidean space between the centroids of the polygons that mark the region boundaries. The tuple with the subject closest to the centre of the image is used to describe what is happening in the image, and the remaining regions are used to describe the background.

T_1	DT o_i AUX REL DT o_j. T_5? *A girl is reading a book.*
T_2	There AUX also $\{\text{DT } o_i\}_{i=1}^{\lvert unrelated \rvert}$ in the image. *There is also a cat and a window in the image.*
T_3	DT o_i AUX REL DT o_j REL DT o_k. T_5? *The girl is reading a book on a chair.*
T_4	REL DT o_j. *reading a book.*
T_5	PRP AUX $\{\text{REL DT } o_i\}_{i=1}^{\lvert dependents \rvert}$. *She is beside a window.*

Table 4.3 *The language generation templates and an example of the type of sentence, or fragment, generated by each template.*

The first sentence of the description is realised with template T_1 from Table 4.3. o_i is the label of the subject region and o_j is the label of the object region. DT is a simple determiner chosen from {the, a}, depending on whether the region label is a plural noun; AUX is either {is, are}, depending on the number of the region label; and REL is a word to describe the relationship between the regions. For this model, REL is the spatial relationship between the centroids chosen from {above, below, beside}, depending on the angle formed between the region centroids, using the definitions in Table 2.1. The second sentence of the description is realised with template T_2 over the subjects o_i in R that were not used in the first sentence. An example of the language generated is:

(1) The man is beside the bike. There is also a road, a car, and trees in the image.

With the exception of visual attributes to describe size, colour, or texture, this model is based on the approach described by Kulkarni et al. (2011).

4.3.2 CORPUS

The biggest limitation of PROXIMITY is that regions that are near each other are not always in a relevant relationship for a description. For example, in Figure 4.1, the BIKE and the CAR regions are nearest neighbours but they are unlikely to be described as being in an relationship by a human annotator. The model CORPUS addresses this issue by using an

external text corpus to determine which pairs of regions are likely to be in a describable relationship. Furthermore, CORPUS can generate verbs instead of spatial relations between regions, leading to more human-like descriptions. CORPUS is based on Yang et al. (2011), except we do not use scene type (indoor, outdoor, etc.) as part of the model. At training time, the model has access to the annotated image regions and labels, and to the dependency-parsed version of the English Gigaword Corpus (Napoles et al., 2012). The corpus is used to extract subject–verb–object subtrees, which are then used to predict the best pairs of regions, as well as the verb that relates the regions. We use the morph toolkit to lemmatise the verbs in the corpus (Minnen et al., 2001).

The set of region–region relationships $R = \{(o_i, v_{ij}, o_j), \ldots\}$ is determined by searching for the most likely o_j^*, v^* given an o_i over a set of verbs $\mathcal{V}$ extracted from the corpus and the other regions in the image. This is shown in Equation 4.1.

$$o_j^*, v^* | o_i = \arg\max_{o_j, v} p(o_i) \cdot p(v|o_i) \cdot p(o_j|v, o_i) \tag{4.1}$$

We can easily estimate $p(o_i)$, $p(v|o_i)$, and $p(o_j|v, o_i)$ directly from the corpus. If we cannot find an o_j^*, v^* for a region, we back-off to the spatial relationship calculation as defined in PROXIMITY. When we have found the best pairs of regions, we select the most probable pair and generate the first sentence of the description using that pair an template T_1. The second sentence is realised with template T_2 over the subjects in R not used in generating the first sentence. An example of the language generated is:

(2) The man is riding the bike. There is also a car, a road, and trees in the image.

In comparison to PROXIMITY, this model will only describe pairs of regions that have observed relations in the external corpus. The corpus also provides a verb that relates the regions, which produces descriptions that are more in line with human-generated text. However, since noun co-occurrence in the corpus controls which regions can be mentioned in the description, this model will be prone to relating regions simply because their labels occur together frequently in the corpus.

4.3.3 STRUCTURE

The model STRUCTURE exploits the visual dependency representation of an image to generate language for only the relationships that hold between pairs of regions. It has access to the image regions, the region labels, and the visual dependency representation of an image.

Region–region relationships are generated during a depth-first traversal of the VDR using templates T_1, T_3, T_4, and T_5. The VDR of an image is traversed and language fragments are generated and then combined depending on the number of children of a node in the tree. If a node has only one child then we use T_1 to generate text for the head-child relationship. If a node has more than one child, we need to decide how to order the language generated by the model. We generate sentence fragments using T_4 for each child independently and combine them later. In STRUCTURE, the sentence fragments are sorted by the Euclidean distance of the children from the parent. In order to avoid problematic descriptions such as "The woman is above the horse is above the field is beside the house", we include a special case for when a node has more than one child. In these cases, the nearest region is realized in direct relation to the head using either T_3 (two children) or T_1 (more than two children), and the remaining regions form a separate sentence using T_5. This sorting and combing process would result in *"The woman is above the horse. She is above a field and beside the house"* for the case mentioned above.

An example of the type of description that can be generated during a traversal is:

(3) The man is above the bike above the road. There is also a car and trees in the image.

In comparison to PROXIMITY, this model can exploit a representation of an image that encodes the relationships between regions in an image (the VDR). However, it is limited to generating spatial relations, because it cannot predict verbs to relate regions.

4.3.4 PARALLEL

The model PARALLEL is an extension of STRUCTURE that uses the image descriptions available to predict verbs that relate regions in parent-child relationships in a VDR. At training time it has access to the annotated regions and labels, the visual dependency representations, and the gold-standard image descriptions. Recall from Section 2.4 that the descriptions were dependency-parsed using the parser of McDonald et al. (2005a) and alignments were calculated between the nodes in the VDRs and the words in the parsed image descriptions.

We estimate two distributions from the image descriptions using the alignments:

$$p(verb|o_{head}, o_{child}, rel_{head-child}) \tag{4.2}$$

$$p(verb|o_{head}, o_{child}) \tag{4.3}$$

The second distribution is used as a backoff when we do not observe the arc label between the regions in the training data. The generation process is similar to that used in STRUCTURE, with two exceptions: (1) it can generate verbs during the generation steps, and (2) when a node has multiple dependents, the sentence fragments are sorted by the probability of the verb associated with them. This sorting step governs which child is in a relationship with its parent. When the model generates text, it only generates a verb for the most probable sentence fragment. The remaining fragments revert back to spatial relationships to avoid generating language that places the subject region in multiple relationships with other regions. An example of the language generated is:

(4) The man is riding the bike on the road. There is also a car and trees in the image.

In comparison to CORPUS, this model generates descriptions in which the relations between the regions determined by the image itself and not by an external corpus. In comparison to PROXIMITY and STRUCTURE, this model generates descriptions that express meaningful relations between

the regions and not simple spatial relationships.

4.3.5 *VDR Parsing*

The STRUCTURE and PARALLEL models rely on visual dependency representations, but it is unrealistic to assume gold-standard representations will always be available because they are expensive to construct. We use the VDR+IMG VDR Parser introduced in Chapter 3 to automatically predict the VDR from region-annotated images, providing the input for the STRUCTURE-PARSED and PARALLEL-PARSED models at test time.

4.4 ANALYSIS OF AUTOMATIC EVALUATION MEASURES

The automatic image description task has been compared to translating an image into text (Li et al., 2011; Kulkarni et al., 2011), or summarising an image (Yang et al., 2011) and these observations have contributed to the adoption of the evaluation measures used in those communities to determine the quality of generated text.

In this section we estimate the correlation of human judgements with five automatic evaluation measures on two image description data sets. Our work extends previous studies of evaluation measures for image description (Hodosh et al., 2013), which focused on unigram-based measures and reported Cohen's κ agreement scores between human judgements and text-based evaluation measures, rather than correlations.

4.4.1 *Methodology*

We estimate Spearman's rank correlation co-efficient of five different automatic evaluation measures against human judgements for the automatic image description task. Spearman's rank is a non-parametric correlation co-efficient that restricts the ability of outlier data points to skew the co-efficient. Each of the automatic measures are calculated on the sentence level and correlated directly against the human judgements.

4.4.2 *Data*

We perform the correlation analysis on the Flickr8K data set of Hodosh et al. (2013). The results of a correlation analysis on our data set can be found in Chapter 4.5.

The test-portion of the Flickr8K data set contains 1,000 images paired with five reference descriptions. The images were retrieved from Flickr, the reference descriptions were collected from Mechanical Turk, and the human judgements were collected from expert annotators as follows: each image in the test data was paired with the highest scoring sentence(s) retrieved from all possible test sentences by any of the ranking models in Hodosh et al. (2013). Each image–description pairing in the test data was judged for semantic correctness by three expert human judges on a scale of 1–4. We calculate automatic measures for each image–retrieved sentence pair against the five reference descriptions for the original image.

4.4.3 *Automatic Evaluation Measures*

BLEU measures the effective overlap between a reference sentence X and a proposed translation sentence Y. It is defined as the geometric mean of the effective n-gram precision scores, multiplied by the brevity penalty factor. p_n measures the effective overlap by calculating the proportion of the maximum number of n-grams co-occurring between a candidate and a reference and the total number of n-grams in the candidate text. The brevity penalty BP penalises short translations. A formal definition:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$p_n = \frac{\sum_{c \in cand} \sum_{n-gram \in c} count_{clipped}(n-gram)}{\sum_{c \in cand} \sum_{n-gram \in c} count(n-gram)}$$

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leqslant r \end{cases}$$

Unigram BLEU without a brevity penalty was reported by Kulkarni et al. (2011); Li et al. (2011); Ordonez et al. (2011); Kuznetsova et al. (2012),

and to the best of our knowledge, the only work to use higher-order n-grams with BLEU is Elliott and Keller (2013). In this analysis we use the smoothed BLEU implementation of Clark et al. (2011) to perform a sentence-level analysis, setting $n = 1$ and no brevity penalty, or $n = 4$ with the brevity penalty. Note that a higher BLEU score is better.

ROUGE measures the longest common subsequence of tokens between a candidate Y and reference text X. There is also a variant that measures the co-occurrence of pairs of tokens in both the candidate and reference (a skip-bigram): ROUGE-SU*. The skip-bigram calculation is parameterised with d_{skip}, the maximum number of tokens between the words in the skip-bigram. Setting d_{skip} to 0 is equivalent to bigram overlap and setting d_{skip} to ∞ means tokens can be any distance apart. If α=|SKIP2(X,Y)| is the number of matching skip-bigrams between the reference and the candidate, then skip-bigram ROUGE is formally defined as:

$$R_{SKIP2} = \frac{\alpha}{\binom{\alpha}{2}}$$

Yang et al. (2011) and Hodosh et al. (2013) measure performance with ROUGE, using a variant described as ROUGE-1. We set set $d_{skip} = 4$ and to award partial credit for unigram only matches; otherwise known as ROUGE-SU4. We use ROUGE v.1.5.5 for the analysis, and configure the evaluation script to return the result for the average score for matching between the candidate and the references. A higher ROUGE score is better.

TER (Translation Error Rate) measures the number of modifications a human would need to make to transform a candidate Y into a reference sentence X. The types of modifications available are insertion, deletion, substitute a single word, and shift a word an arbitrary distance. TER is expressed as the percentage of the sentence that needs to be changed, and can be greater than 100 if the candidate is longer than the reference. More formally,

$$\text{TER} = \frac{|\text{edits}|}{|\text{reference tokens}|}$$

TER has not yet been used in the image description literature for model evaluation. We use v.0.8.0 of the TER evaluation tool, and a lower TER is

better.

Meteor is the harmonic mean of unigram precision and recall that allows for exact, synonym, and paraphrase matchings between candidates and references. It is calculated by generating an alignment between the tokens in the candidate and reference sentences, with the aim of a 1:1 alignment between tokens and minimising the number of chunks *ch* of contiguous and identically ordered tokens in the sentence pair. This alignment is based on exact token matching, followed by Wordnet synonyms, and then stemmed tokens. We can calculate precision, recall, and F-measure, where *m* is the number of aligned unigrams between candidate and reference. Meteor is defined as:

$$
\begin{aligned}
M &= (1 - Pen) \cdot F_{mean} \\
Pen &= \gamma \left(\frac{ch}{m} \right)^{\theta} \\
F_{mean} &= \frac{PR}{\alpha P + (1 - \alpha) R} \\
P &= \frac{|m|}{|\text{unigrams in candidate}|} \\
R &= \frac{|m|}{|\text{unigrams in reference}|}
\end{aligned}
$$

The results reported in the analysis are extracted from Meteor v.1.4.0 and use the package-provided free parameter settings of 0.85, 0.2, 0.6, and 0.75 for the matching components. Meteor has not yet been reported to evaluate the performance of different models on the image description task and a higher Meteor score is better.

4.4.4 *Protocol*

We performed the correlation analysis as follows. The sentence-level evaluation measures were calculated for each image–description–reference tuple in the test data. We used MultEval (Clark et al., 2011) to collect the BLEU, TER, and Meteor scores, and the ROUGE-SU4 scores were collected using the RELEASE-1.5.5.pl script. The evaluation measure scores were merged with the human judgements for each tuple, and the correlation was estimated at a sentence-level, using cor.test in R.

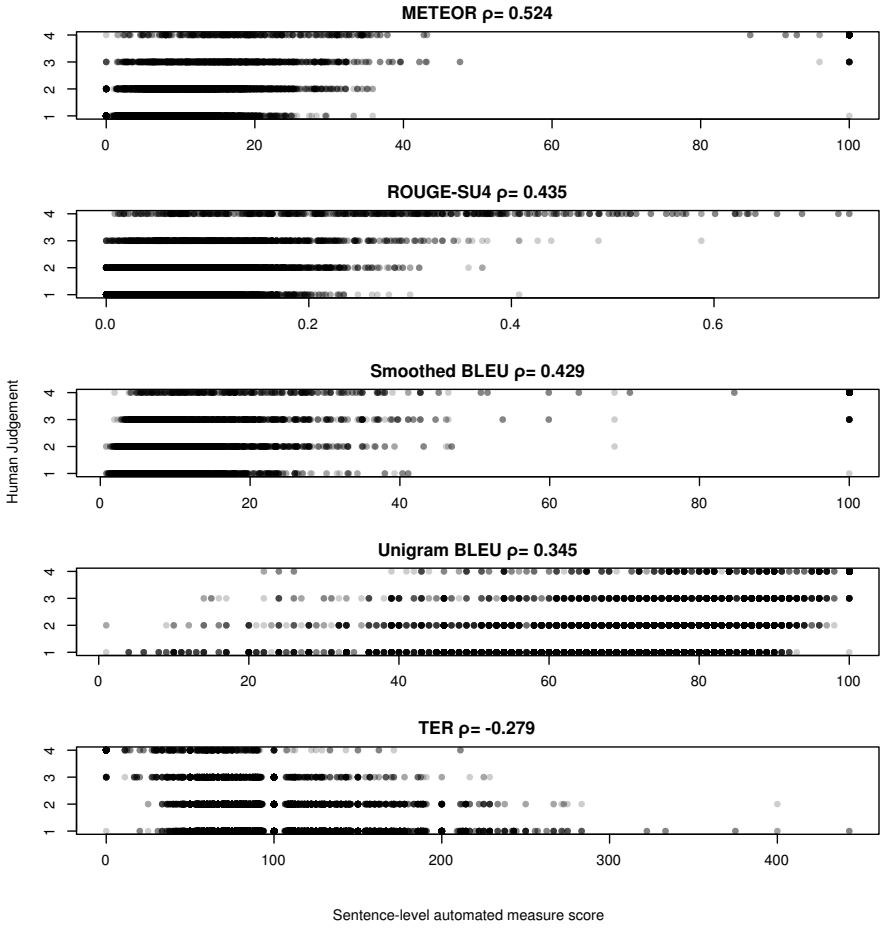
Measure	Flickr 8K Co-efficient $n = 17,466$
METEOR	0.524
ROUGE SU-4	0.435
Smoothed BLEU	0.430
Unigram BLEU	0.345
TER	-0.280

Table 4.4 *Correlation co-efficient of five automatic evaluation measures against human judgements. All correlations are significant at $p < 0.001$. The strength of the correlation co-efficient depends on the data set but the pattern holds across data sets.*

4.4.5 *Correlation Analysis*

Table 4.4 shows the correlation co-efficients between automatic measures and human judgements and Figure 4.2 shows the distribution of scores for each measure against human judgements. To classify the strength of the correlations, we followed the guidance of Dancey and Reidy (2011), who posit that a co-efficient of 0.0–0.1 is uncorrelated, 0.11–0.4 is *weak*, 0.41–0.7 is *moderate*, 0.71–0.90 is *strong*, and 0.91–1.0 is *perfect*.

On the Flickr8k data set, all evaluation measures can be classified as either *weakly* correlated or *moderately* correlated with human judgements and all results are significant. TER is weakly correlated with human judgements. However, TER could prove useful in comparing the types of differences between models, such as the difference between the number of insertions, substitutions, etc. Also, an analysis of the distribution of TER scores in Figure 4.2 shows that differences in candidate and reference length are prevalent in the image description task. Unigram BLEU is only weakly correlated against human judgements, even though it has almost universally been used as an automatic evaluation measure for the image description task. Figure 4.2 shows an almost uniform distribution of unigram BLEU scores, regardless of the human judgement. Smoothed BLEU and ROUGE-SU4 are moderately correlated with human judgements, and the correlation is much stronger than unigram BLEU. Finally, Meteor is most strongly correlated measure against human judgements.



(a) Flick8K data set, n=17,466.

Figure 4.2 *Flickr8K data set correlation of automatic evaluation measures against human judgements.* ρ *is the correlation between human judgements and the automatic measure. The intensity of each point indicates the number of occurrences that fall into that range.*

Candidate: Football players gathering to contest something to collaborating officials.
Reference: A football player in red and white is holding both hands up.

(a)

Candidate: A man is attempting a stunt with a bicycle.
Reference: Bmx biker Jumps off of ramp.

(b)

Figure 4.3 *A pair of examples in the test data which have a Meteor score 0.0 and the maximum expert human judgement. In (a) the candidate and references originate from the same image, and highlight a difference in content selection, whereas in (b) they come from different images and highlight a difference in vocabulary.*

Qualitative Analysis

Figure 4.3 shows two images from the test collection of the Flickr8K data set with a low Meteor score and a maximum human judgement of semantic correctness. The main difference between the candidates and references are in deciding *what* to describe (content selection), and *how* to describe it (realisation). We can hypothesise that in both translation and summarisation, the source text acts as a lexical and semantic framework within which the translation or summarisation process takes place. In Figure 4.3(a), the authors of the descriptions made different decisions on *what* to describe. A decision has been made to describe the role of the officials in the candidate text, and not in the reference text. The underlying cause of this is an active area of research in the human vision literature and can be attributed to bottom-up effects, such as saliency (Itti et al., 1998), top-down contextual effects (Torralba et al., 2006), or rapidly-obtained scene properties (Oliva and Torralba, 2001). In (b), we can see the problem of deciding how to describe the selected content. The reference text has used a more specific noun to describe the person on the bicycle than

the candidate text.

4.4.6 *Discussion*

There are several differences between our analysis and that of Hodosh et al. (2013). First, we report Spearman's ρ correlation coefficient of text-based automatic measures against human judgements, whereas they report agreement between judgements and text-based automatic measures in terms of Cohen's κ. The use of κ requires the transformation of real-valued scores into categorical values, and thus loses information; we use the judgement and evaluation measure scores in their original forms. Second, our use of Spearman's ρ means we can readily use all of the available data for the correlation analysis, whereas Hodosh et al. (2013) report agreement on thresholded subsets of the data. Third, we report the correlation coefficients against five evaluation measures, some of which go beyond unigram matchings between references and candidates, whereas they only report unigram BLEU and unigram ROUGE. It is therefore difficult to directly compare the results of our correlation analysis against Hodosh et al.'s agreement analysis, but they also reach the conclusion that unigram BLEU is not an appropriate measure of image description performance. However, we do find stronger correlations with Smoothed BLEU, skip-bigram ROUGE, and Meteor.

In contrast to the results presented here, Reiter and Belz (2009) found no significant correlations of automatic evaluation measures against human judgements of the *accuracy* of machine-generated weather forecasts. They did, however, find significant correlations of automatic measures against *fluency* judgements. There are no fluency judgements available for the Flickr8K data set, so we cannot measure the correlation of fluency judgements against automatic measures at this point.

In this section we performed a sentence-level correlation analysis of automatic evaluation measures against expert human judgements for the automatic image description task. We found that sentence-level unigram BLEU is only weakly correlated with human judgements, even though it has extensively reported in the literature for this task. Meteor was found to have the highest correlation with human judgements, but it requires

Wordnet and paraphrase resources that are not available for all languages. Our findings held when judgements were made on human-written or computer-generated descriptions.

The variability in what and how people describe images will cause problems for all of the measures compared in this paper. Nevertheless, we propose that unigram BLEU should no longer be used as an objective function for automatic image description because it has a weak correlation with human accuracy judgements. We recommend adopting either Meteor, Smoothed BLEU, or ROUGE-SU4 because they show stronger correlations with human judgements. We believe these suggestions are also applicable to the ranking tasks proposed in Hodosh et al. (2013), where automatic evaluation scores could act as features to a ranking function.

Given the result of this analysis, we will report ROUGE-SU4, Meteor, and smoothed BLEU results in the remainder of this chapter.

4.5 LANGUAGE GENERATION EXPERIMENTS

We now evaluate the image description models with human judgements and in an automatic setting. The human judgements were collected from untrained works on Amazon Mechanical Turk. In the automatic setting, we measure how close the model-generated descriptions are to the gold-standard descriptions using BLEU, ROUGE, and Meteor metrics.

4.5.1 *Methodology*

The task is to produce a description of an image. The PROXIMITY and CORPUS models have access to gold-standard region labels and region boundaries at test time. The STRUCTURE and PARALLEL models have additional access to the visual dependency representation of the image. These representations are either the gold-standard, or in the case of Parsed STRUCTURE and Parsed PARALLEL, produced by the image parser described in Chapter 3. Table 4.2 provides a reminder of the information the different models have access to at training time.

The experiment is performed using 10-fold cross-validation over the same 341 annotated images used for the VDR Parsing experiments (see Chapter

3.6.1 for more details on how the data was split). The VDR Parser used for the Parsed STRUCTURE and Parsed PARALLEL models is trained on the gold-standard VDRs of the training splits, and then predicts VDRs on the development and test splits. Significant differences were measured using a one-way ANOVA with Parsed PARALLEL as the reference[1], with differences between pairs of mean checked with a Tukey HSD test.

4.5.2 *Human Judgements*

We collected human judgements of the generated image descriptions using Mechanical Turk to complement the automatic evaluation. Workers were paid $0.05 to rate the quality of an image–description pair generated by one of the models using three criteria on a scale from 1 to 5:

Grammaticality: give high scores if the description is correct English and doesn't contain any grammatical mistakes.

Action: give high scores if the description correctly describes what people are doing in the image.

Scene: give high scores if the description correctly describes the rest of the image (background, other objects, etc).

A total of 101 images were used for this evaluation and we obtained five judgements for each image-description pair, resulting in a total of 3,535 judgements. The 101 images were taken from the test data that corresponds with the median performing Smoothed BLEU development data on the PARALLEL model. We chose the median performing development split to ensure a conservative evaluation by human judges – they were exposed to image descriptions that were unlikely to have particularly high or low Smoothed BLEU scores.

4.5.3 *Results*

The results of the image description experiment are presented in Table 4.5, and Figure 4.4 shows example output for the models. PROXIMITY uses raw euclidean distance between object regions to determine which pairs

[1]Recall that PARALLEL uses gold-standard VDRs and Parsed PARALLEL uses the output of the image parser described in Section 3.

		
PROXIMITY	A man is beside a phone. There is also a wall and a sign in the image.	A beach is above a beach. There are also horses, a woman, and a man in the image.
CORPUS	A man is holding a sign. There is also a wall and a phone in the image.	A woman is outnumbering a man. There are also horses and beaches in the image.
STRUCTURE	A wall is above a wall. A man is beside a sign.	A man is beside a woman above a horse. A horse is beside a woman beside a beach.
PARALLEL	A man is holding a phone. A wall is beside a sign.	A man is riding a horse above a beach. A horse is beside a beach beside a woman.
GOLD	A foreign man with sunglasses talking on a cell phone. A large building and a mountain in the background.	There is a man and women both on horses. They are on a beach during the day.

Figure 4.4 *Some example descriptions produced by* PROXIMITY, CORPUS, STRUCTURE *and* PARALLEL.

	Human	PROXIMITY	CORPUS	STRUCTURE	PARALLEL
Meteor	—	13.1 ± 0.4	14.4 ± 0.6	13.4 ± 0.5	$\mathbf{18.2 \pm 0.7}^{\star}$
ROUGE-SU4	—	27.0 ± 0.6	27.7 ± 0.6	15.9 ± 0.9	$\mathbf{29.4 \pm 0.6}^{\star}$
BLEU	—	6.6 ± 0.9	8.9 ± 1.5	11.0 ± 0.7	$\mathbf{19.8 \pm 2.1}^{\star}$
Grammar	4.8 ± 0.4	3.7 ± 1.5	4.4 ± 1.1	4.1 ± 1.4	4.5 ± 1.0
Action	4.8 ± 0.6	2.1 ± 0.3	2.2 ± 1.3	2.1 ± 1.4	$\mathbf{3.4 \pm 1.6}^{\star}$
Scene	4.6 ± 0.7	3.0 ± 1.4	3.4 ± 1.3	3.0 ± 1.4	$\mathbf{3.7 \pm 1.3}^{\star}$

Table 4.5 *Automatic and human judgement evaluation for the four language generation models presented in this chapter. The results were averaged over ten random splits. It can be seen that the model that exploits image structure and the parallel corpus (*PARALLEL*) is significantly better than all other models on all measures. *: significantly better than all other models at* $p < 0.01$.

of objects should be related in a description. It is the lowest performing model as measured by the automatic measures: its Meteor and BLEU scores are significantly worse than all other models; the ROUGE-SU4 score is not significantly different from other models. However, the human judgement scores are very low: the text makes the least grammatical sense of all the models, is bad at describing the main event of the image, and the quality of the description of the background objects is lowest of all the models.

CORPUS uses an external corpus to determine which objects should be related, and the verb that should be used to relate the objects. It is significantly better than PROXIMITY at generating descriptions, as measured by Meteor and BLEU; the ROUGE-SU4 score is not significantly different compared to PROXIMITY. This model is also significantly better at producing grammatically correct sentences and action descriptions compared to the PROXMITY model. The quality of the action descriptions is not significantly different compared to the PROXIMITY model, which means the Annotated Gigaword Corpus is not a reliable resource from which to estimate the parameters for Equation 4.3.2. In Section 4.5.6 we will examine the effect of estimating the parameters of this model from the parallel image descriptions.

STRUCTURE relates image regions through the spatial relationships expressed in Visual Dependency Representations. It is significantly different

	PARALLEL	Parsed PARALLEL	STRUCTURE	Parsed STRUCTURE
Meteor	18.2 ± 0.7	17.9 ± 0.8	13.4 ± 0.5	13.9 ± 0.9
ROUGE-SU4	29.4 ± 0.6	29.6 ± 0.5	15.9 ± 0.9	$\mathbf{17.8 \pm 0.5}^{\star}$
BLEU	19.8 ± 2.1	19.2 ± 1.7	11.0 ± 0.7	11.6 ± 0.6
Grammar	4.5 ± 1.0	4.2 ± 1.3	4.1 ± 1.4	4.0 ± 1.4
Action	3.4 ± 1.6	3.3 ± 1.7	2.1 ± 1.4	1.6 ± 1.4
Scene	3.7 ± 1.3	3.5 ± 1.3	3.0 ± 1.4	3.2 ± 1.3

Table 4.6 *The effect of using automatically parsed image structures on the* PARALLEL *and* STRUCTURE *models. There is no significant difference in the generated descriptions when using automatically predicted image structures. There only significant differences between using gold-standard VDR and automatically parsed VDR were found using the* ROUGE-SU4 *measure for the* STRUCTURE *model.*

compared to CORPUS as measured using METEOR and ROUGE-SU4. It produces descriptions that humans judge to be less grammatical than CORPUS, and the quality of the action and scene sentences are equivalent to the PROXIMITY model.

PARALLEL exploits VDR image structure and the parallel image descriptions to relate objects in an image. It is significantly and substantially better than all other models using the Meteor and BLEU score measures, although not the ROUGE-SU4 measure. It produces the most grammatically correct sentences, significantly better descriptions of the main event of the image, and the best scene descriptions, compared to the other automatic models. Overall, this model exploits image structure and the parallel descriptions corpus to support the hypothesis that image structure is useful when describing 6images.

In this remainder of this chapter we will explore the effect of using automatically parsed image structures, of evaluating the quality of the sentences separately, and how the choice of corpus to estimate the verb probabilities for CORPUS or PARALLEL affect model performance.

	Human	PROXIMITY	CORPUS	Parsed STRUCTURE	Parsed PARALLEL
Meteor	—	11.2 ± 0.5	15.2 ± 1.3	15.4 ± 0.7	$\mathbf{21.7 \pm 1.5}^{\star}$
ROUGE-SU4	—	23.4 ± 0.8	24.9 ± 0.6	19.5 ± 1.3	$\mathbf{27.4 \pm 1.1}^{\star}$
BLEU	—	3.8 ± 0.3	10.1 ± 3.1	16.1 ± 1.0	$\mathbf{29.1 \pm 4.1}^{\star}$
Action	4.8 ± 0.6	2.1 ± 0.3	2.2 ± 1.3	1.6 ± 1.3	$\mathbf{3.3 \pm 1.7}^{\star}$

Table 4.7 *Automatic and human judgments of only the first sentence, which describes the main event depicted in the image. The differences between the models is now much clearer than when we use both sentences. *: significantly better than all other models at* $p < 0.01$.

4.5.4 *Automatically Parsing Image Structures*

We now study the effect of using an automatic image parser as a pre-processing step to generating image descriptions instead of using gold-standard Visual Dependency Representation. We use the VDR+IMG parser described in Chapter 3 to automatically predict the relationships between annotated objects in an image, and use the output of this process to generate descriptions. It is not possible to use the VDR+IMG+QDG parser because we are trying to generate the descriptions. Table 4.6 shows the results of this experiment. It can be seen that the only statistically significant difference between using gold-standard and automatically produced image structures is in the ROUGE-SU4 measure. We conclude that using image structures automatically predicted from region annotations is sufficient for the automatic image description task. Further results in this chapter will use the Parsed versions of the STRUCTURE and PARALLEL models.

4.5.5 *Action/Scene Sentence Evaluation*

The results in Chapter 4.5.3 did not suggest a distinction between the models using the ROUGE-SU4 evaluation measure. This was unexpected because the correlation analysis in Section 4.4 proposed a moderate correlation between this measure and human judgements. One explanation for this discrepancy could that skip-bigram matchings can occur across sentence boundaries. An example of this can be seen in the can-

	Human	PROXIMITY	CORPUS	Parsed STRUCTURE	Parsed PARALLEL
Meteor	—	11.0 ± 0.5	11.5 ± 0.6	7.9 ± 0.8	12.9 ± 1.1
ROUGE-SU4	—	21.2 ± 0.9	21.2 ± 1.0	9.5 ± 1.3	22.9 ± 0.9
BLEU	—	5.2 ± 0.9	4.8 ± 1.2	9.4 ± 1.5	5.9 ± 1.7
Scene	4.6 ± 0.7	3.0 ± 1.4	3.4 ± 1.3	3.2 ± 1.3	3.5 ± 1.3

Table 4.8 *Automatic and human judgements of only the second sentence, which describes the background scene of the image. No model is consistently significantly better than any other model.*

didate/reference pairing in (5) and (6), which would count positively towards ROUGE-SU4. However, this is not an example of a good description because the target words, *man* and *bike* are clearly not related to each other in (5). It is possible that all models are capable of generating these types of sentences and gaining artificial credit in automatic measures.

(5) The car is above the man. The bike is beside the tree.

(6) The man is riding a bike . There are trees beside the road.

This observation leads to evaluating each sentence seperately. Table 4.7 presents the results for the first sentence, which is intended to describe the main event depicted in the image. It can be seen that the differences between the models are substantial, on all automatic evaluation measures, compared to evaluating against both sentences at the same time. Parsed PARALLEL is now significantly better than all models on all measures, and the magnitude of the differences is substantially increased. This is most likely because the models are not receiving credit for matches occurring across sentence boundaries. We note that the Parsed STRUCTURE model outperforms the CORPUS model on the automatic measures because Visual Dependency Representations guides the order of generation. This means more n-grams are in the correct in the proposed description. However, the CORPUS model can generate verbs instead of spatial relationships between objects, which humans clearly prefer in the human judgements.

	Parsed PARALLEL	Parsed PARALLEL (Gigaword)	CORPUS	CORPUS (Descriptions)
Meteor	**23.8** ± **1.4**⋆	20.5 ± 1.3	15.2 ± 1.3	12.4 ± 0.7
ROUGE-SU4	**29.3** ± **1.1**⋆	26.6 ± 0.8	24.9 ± 0.8	24.7 ± 0.7
BLEU	**30.8** ± **4.0**⋆	18.3 ± 4.0	10.1 ± 3.1	7.2 ± 3.1
Grammar	4.2 ± 1.3	4.2 ± 1.3	**4.4** ± **1.1**⋆	4.1 ± 1.4
Action	**3.3** ± **1.7**⋆	2.7 ± 1.6	2.2 ± 1.3	2.1 ± 1.4
Scene	3.5 ± 1.3	3.5 ± 1.3	3.4 ± 1.3	3.1 ± 1.4

Table 4.9 *The effect of estimating verb probabilities from different text corpora. We experimented with using the external Annotated Gigaword Corpus with the Parsed* PARALLEL *model, and using the* CORPUS *model with the parallel image descriptions in our data set.*

The second sentence evaluation results are shown in Table 4.8. It can be seen that the PROXIMITY and CORPUS models are significantly better at generating Scene descriptions than the Parallel PARSED model. This finding is not consistent with the human judgements for the Scene descriptions. However, we saw in Table 4.7 that the proposed model Parallel PARSED is significantly and substantially better than the baseline models PROXIMITY and CORPUS at describing the main event of the image.

4.5.6 *Collecting Statistics from Alternative Text Corpora*

We now explore the role of the text corpus used to estimate the verb probabilities for the CORPUS and PARALLEL models. This is straightforward for the CORPUS model, which estimates subject, verb, and object probabilities from a dependency parsed corpus:

$$p(\text{verb}|o_{\text{head}}, o_{\text{child}}, \text{rel}_{\text{head-child}}) \tag{4.4}$$

$$p(\text{verb}|o_{\text{head}}, o_{\text{child}}) \tag{4.5}$$

We hypothesise that this should improve the quality of the descriptions because we are less likely to generate a verb that is never seen in the images. On the other hand, the PARALLEL models rely on alignments

between image region labels and tokens in the description to estimate the parameters of the probability distributions that govern verb selection, given a subject and object. To use an external corpus to obtain a distribution over verbs, we need to relax the requirement of having alignments, and just assume that if the tokens are lexically identical, then they refer to the same objects.

We experiment with generating descriptions using the CORPUS model trained on the parallel descriptions, and on the Parsed PARALLEL model using the Annotated Gigaword Corpus to estimate verbs. Table 4.9 shows the results of this experiment.

There is no benefit to using the parallel descriptions to estimate the parameters for the CORPUS model. This suggests that relating objects using an external corpus is not a reliable approach to description generation. The human judgements support the findings of the automatic measures.

If we use the Annotated Gigaword Corpus to predict the verb that relates a pair of objects in a Visual Dependency Representation, we observe a significant decrease in all automatic measures, and in the human judgement of the Action description. These descriptions are still significantly better than the PROXMITY and CORPUS models.

The implications of this experiment are that we don't need to obtain a parallel corpus of image descriptions if we want to generate descriptions. The descriptions generated using automatically predicted image structures and an external text corpus for verb estimation (Parsed PARALLEL Gigaword) are significantly better than not using image structures on the same text corpus (CORPUS).

4.5.7 *Revisiting Human Judgement Correlations*

Earlier in this chapter we estimated the correlation of automatic evaluation measures against human judgements on the Flickr8K data set. Now that we have collected human judgements for our data set, we can perform an identical correlation analysis. Table 4.10 shows the results of this analysis and Figure 4.5 shows the distribution of scores against human judgements.

Measure	E&K (2013) Co-efficient $n = 2,040$
METEOR	0.232
ROUGE SU-4	0.188
Smoothed BLEU	0.177
Unigram BLEU	0.097
TER	-0.044

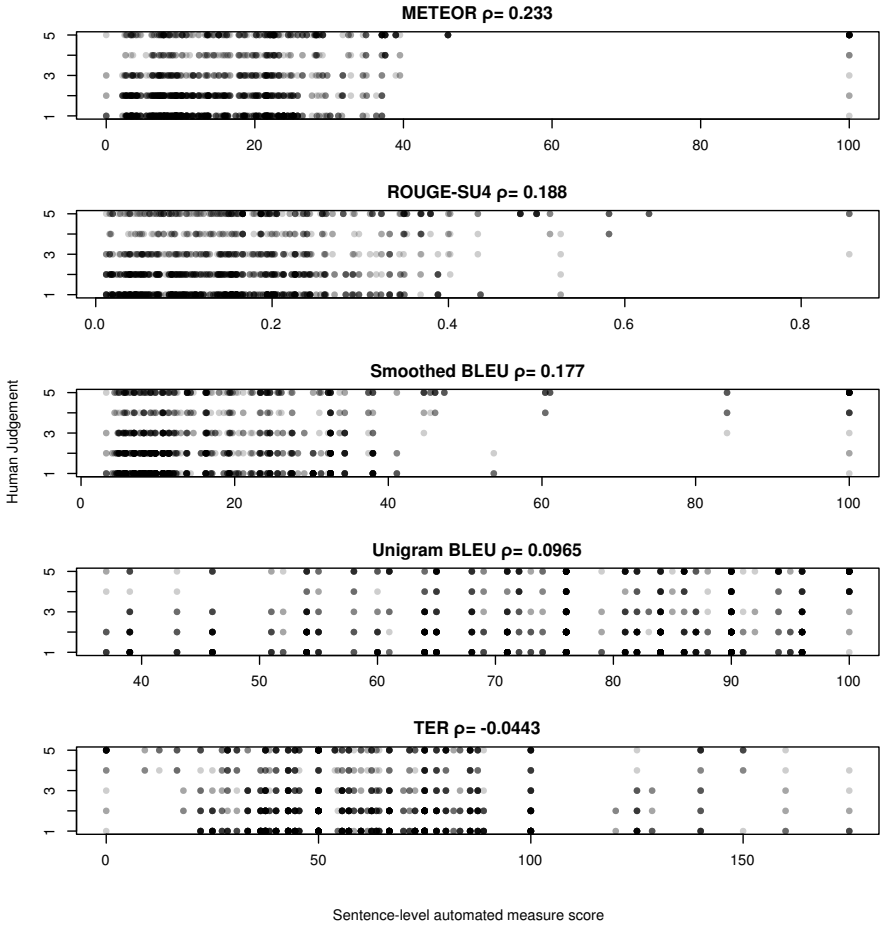
Table 4.10 *Correlation co-efficient of five automatic evaluation measures against human judgements. All correlations are significant at $p < 0.001$. The strength of the correlation co-efficient depends on the data set but the pattern holds across data sets.*

It can be seen that the correlation co-efficients are lower than on the Flickr8K data set. Regardless of the decrease in the co-efficient values, the pattern of correlation strength remains the same as on the Flickr8K data. This could be because the data set is smaller and is a less representative sample of the distribution of possible good and bad image descriptions. Alternatively, this could be because the descriptions were automatically generated by a computer and not retrieved from a list of human-written descriptions.

It was not possible to estimate the correlation of fluency judgements against automatic measures for the Flickr8K data set but we collected grammaticality judgements for our data set, which are comparable to fluency ratings. We failed to find significant correlations between grammaticality judgements and any of the automatic measures on our data set. This discrepancy could be explained in terms of the differences between the weather forecast generation and image description tasks, or because the image description data sets contain thousands of texts and a few human judgements per text, whereas the data sets of Reiter and Belz (2009) included hundreds of texts with 30 human judges.

4.6 CONCLUSIONS

In this chapter we demonstrated that Visual Dependency Representations of images can be useful in the automatic image description process. We



(a) VLT data set, n=2,040.

Figure 4.5 *Data points for the automatic evaluation measures against human judgements on the Visual and Linguistic Treebank data set. ρ is the correlation between human judgements and the automatic measure. The intensity of each point indicates the number of occurrences that fall into that range.*

performed an image description generation experiment using our corpus of images annotated with regions, Visual Dependency Representations, and human-written text. The main finding was that using our proposed structured image representation leads to significant improvements in the quality of generated descriptions compared to two competitive baseline models. This improvement remained even if we used an unrelated text corpus to estimate the verbs that should be used when realising the text, which suggests the approach will scale to different image collections. We found that using automatically predicted image structures were as reliable as using gold-standard structures, which suggests the approach can scale to larger image collections.

We also presented a correlation analysis of automatic evaluation measures against human judgements. The main finding of this analysis was that unigram BLEU, which has been almost universally used for evaluating automatic image description models, is not as strongly correlated as Meteor or ROUGE-SU4. Future research in this area should focus on using a more strongly correlated automatic measure to ensure model development is likely to be consistent with human judgements.

5

Image retrieval models typically represent images as bags-of-terms, a representation that is well-suited to matching images based on the presence or absence of terms. For some information needs, such as searching for images of people performing actions, it may be useful to retain data about how parts of an image relate to each other. If the underlying representation of an image can distinguish between images where objects only co-occur from images where people are interacting with objects, then it should be possible to improve retrieval performance. In this chapter we model the spatial relationships between image regions using Visual Dependency Representations, a structured image representation that makes it possible to distinguish between object co-occurrence and interaction. In a query-by-example image retrieval experiment on data set of people performing actions, we find an 8.8% relative increase in MAP and an 8.6% relative increase in Precision@10 when images are represented using the Visual Dependency Representation compared to a bag-of-terms baseline.

5.1 INTRODUCTION

Every day millions of people search for images on the web, both professionally and for personal amusement. The majority of image searches are aimed at finding a particular named entity, such as *Justin Bieber* or *supernova*, and a typical image retrieval system is well-suited to this type of information need because it represents an image as a bag-of-terms drawn from data surrounding the image, such as text, manual tags, and anchor text (Datta et al., 2008). It is not always possible to find useful terms in the surrounding data; the last decade has seen advances in automatic methods for assigning terms to images that have neither user-assigned tags, nor a textual description (Duygulu et al., 2002; Lavrenko et al., 2003; Guillaumin and Mensink, 2009). These automatic methods learn to associate the presence and absence of labels with the visual characteristics of an image, such as colour and texture distributions, shape, and points of

interest, and can automatically generate a bag of terms for an unlabelled image.

It is important to remember that not all information needs are entity-based: people also search for images reflecting a mood, such as *people having fun at a party*, or an action, such as *using a computer*. The bag-of-terms representation is limited to matching images based on the *presence or absence* of terms, and not the *relation* of the terms to each other. Figures 5.1(a) and (b) highlight the problem with using unstructured representations for image retrieval: there is a person and a computer in both images but only (a) depicts a person actually using the computer. To address this problem with unstructured representations we propose to represent the structure of an image using the Visual Dependency Representation. The Visual Dependency Representation is a directed labelled graph over the regions of an image that captures the spatial relationships between regions. The representation is inspired by evidence from the psychology literature that people are better at recognising and searching for objects when the spatial relationships between the objects in the image are consistent with our expectations of the world (Biederman, 1972; Bar and Ullman, 1996). In Chapter 4 we showed that encoding the spatial relationships between objects in the Visual Dependency Representation helped to generate significantly better descriptions than approaches based on the spatial proximity of objects (Farhadi et al., 2010) or corpus-based models (Yang et al., 2011). In this chapter we study whether the Visual Dependency Representation of images can improve the performance of query-by-example image retrieval models.

5.2 RELATED WORK

5.2.1 *Representing Images*

A central problem in image retrieval is how to abstractly represent images (Datta et al., 2008). A bag-of-terms representation of an image is created by grouping visual features, such as color, shape (Shi and Malik, 2000), texture, and interest points (Lowe, 1999), in a vector or as a probability distribution over the features. Image retrieval can then be performed by

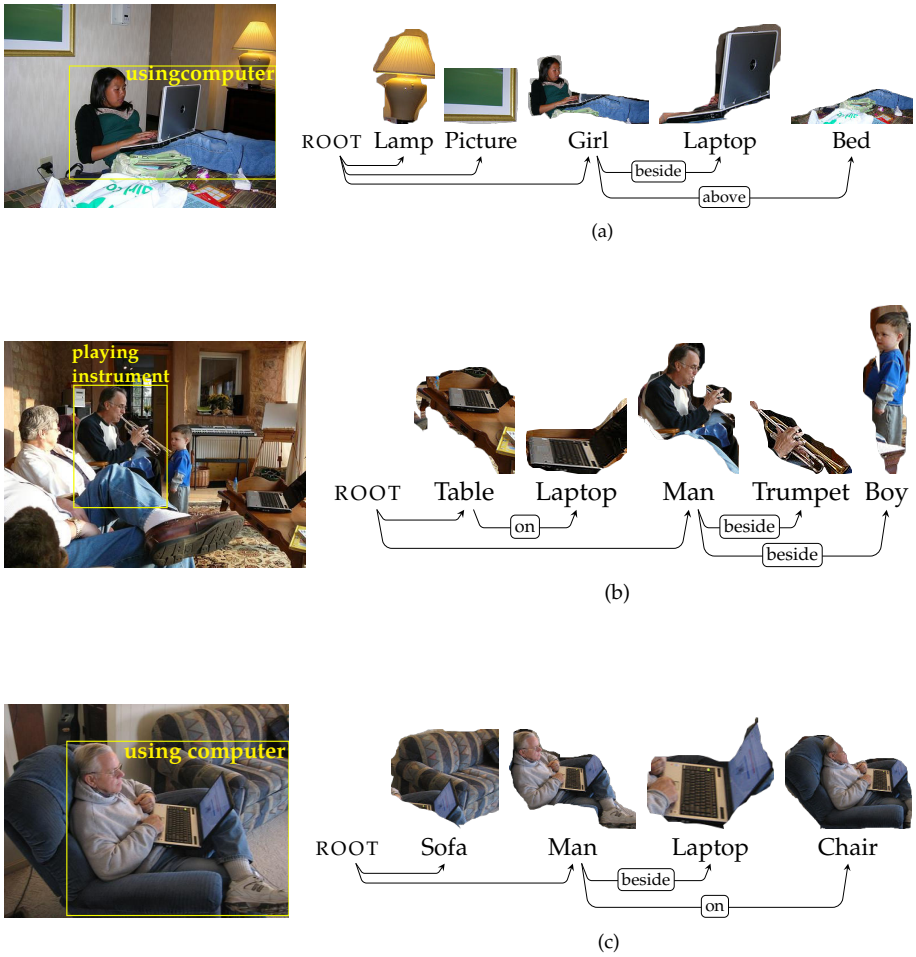


Figure 5.1 *Three examples of images depicting a person and a computer, alongside a respective Visual Dependency Representation for each image. The bag-of-terms representation can be observed in the annotated regions of the Visual Dependency Representations. In (a) and (c) there is a person using a laptop, whereas in (b) the man is actually using the trumpet. The gold-standard action annotation is shown in the yellow bounding box.*

trying to find the best matchings of terms across an image collection. Spatial Pyramid Matching is an approach to constructing low-level image representations that capture the relationships between features at differently sized partitions of the image (Lazebnik et al., 2006). This approach has proven successful for scene categorisation tasks. An alternative approach to representing images is to learn a mapping (Duygulu et al., 2002; Lavrenko et al., 2003; Guillaumin and Mensink, 2009) between the bags-of-terms and object tags. An image can then be represented as a bag-of-terms and image retrieval is similar to text retrieval (Wu et al., 2012).

In this chapter, we represent an image as a directed acyclic graph over a set of labeled object region annotations. This representation captures the important spatial relationships between the image regions and makes it possible to distinguish between co-occurring regions and interacting regions.

5.2.2 *Still-Image Action Recognition*

One approach to recognizing actions is to learn appearance models for *visual phrases* and use these models to predict actions (Sadeghi and Farhadi, 2011). A visual phrase is defined as the people and the objects they interact with in an action. In this approach, a fixed number of visual phrase models are trained using the deformable parts object detector (Felzenszwalb et al., 2010) and used to perform action recognition.

An alternative approach is to model the relationships between objects in an image, and hence the visible actions, as a Conditional Random Field (CRF), where each node in the field is an object and the factors between nodes correspond to features that capture the relationships between the objects (Zitnick et al., 2013). The factors between object nodes in the CRF include object occurrence, absolute position, person attributes, and the relative location of pairs of objects. This model has been used to generate novel images of people performing actions and to retrieve images of people performing actions.

Most recently, actions have been predicted in images by selecting the most likely verb and object pair given a set of candidate objects detected in an

image (Le et al., 2013a). The verb and object is selected amongst those that maximize the distributional similarity of the pair in a large and diverse collection of documents. This approach is most similar to ours but it relies on an external corpus and, depending on the text collections used to train the distributional model, will compound the problem of co-occurrence of objects instead of the relationships between the objects.

The work presented in this chapter uses ground-truth annotation for region labels, an assumption similar to Zitnick et al. (2013), but requires no external data to make predictions of the relationships between objects, unlike the approach of Le et al. (2013a). The directed acyclic graph representation we propose for images can be seen as a latent representation of the depicted action in the image, where the spatial relationships between the regions capture the different types of actions.

5.3 TASK AND BASELINE

In this chapter we study the task of query-by-example image retrieval within the restricted domain of images depicting actions. More specifically, given an image that depicts a given action, such as *using a computer*, the aim of the retrieval model is to find all other images in the image collection that depict the same action. We define an action as an event involving one or more entities in an image, e.g., *a woman running* or *boy using a computer*, and assume all images have been manually annotated for objects. This assumption means we can explore the utility of the Visual Dependency Representation without the noise introduced by automatic computer vision methods. The data available to the retrieval models can be seen in Figure 5.1, and Section 5.5 provides further details about the different sources of data. The action label - which is only used for evaluation - is shown in the labelled bounding box, and the Visual Dependency Representation - not used by the baseline model - is shown as a tree at the bottom of the figure.

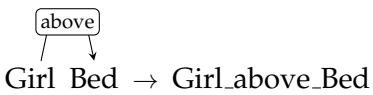
The main hypothesis explored in this chapter is that the accuracy of an image retrieval model will increase if the representation encodes information about the relationships between the objects in images. This hypothesis is tested by encoding images as either an unstructured bag-of-terms rep-

resentation or as the structured Visual Dependency Representation. The Bag-of-Terms baseline represents the query image and the image collection as an unstructured bags-of-terms vector. All of the models used to test the main hypothesis use the cosine similarity function is to determine the similarity of the query image to other images in the collection, and thus to generate a ranked list from the similarity values.

5.4 COMPARING VISUAL DEPENDENCY REPRESENTATIONS

How can we compare the Visual Dependency Representations of a pair of images. The most obvious approach is to use the labelled directed accuracy measurement used for the VDR prediction evaluation in the previous section, but we did not find significant improvements in retrieval accuracy using this method. We hypothesise that the lack of weight given to the edges between nodes in the Visual Dependency Representation results in this comparison function not distinguishing between object–object relationships that matter, such as PERSON $\overrightarrow{\text{beside}}$ BIKE, compared to ROOT $\rightarrow$ TREES. The former is a potential person–object relationship that explains the depicted event, whereas the latter is only a background object.

The approach we have adopted is to compare Visual Dependency Representations of images by decomposing the structure into a set of labelled and a unlabelled parent–child subtrees in a depth-first traversal of the VDR, alongside the unigram region labels from the Bag-of-Terms representation. The decomposition process allows use to use the same similarity function as the Bag-of-Terms baseline model, removing the confound of choosing different similarity functions. The subtrees can be transformed into tokens and these tokens can be used as weighted terms in a vector representation. An example of a labelled transformation is shown below:



The decomposed VDR representation of an image is an N-dimensional

vector[1]. The elements of this vector are the labelled and unlabelled parent-child subtrees and the unigram region labels. The weight of an element in the vector is zero if it does not occur in the image, or as its *tf* idf* value. No frequency cut-offs are applied when constructing the vectors. The *tf*-value is the number of times this element appears in the image. The *idf*-value of a $\mathtt{term}$ is calculated in an external $\mathtt{Corpus}$ (we use the plus-one smoothing variant of *idf* to avoid divide-by-zero errors): [2]

$$\mathtt{idf}(\mathtt{term}, \mathtt{Corpus}) = \log \frac{|\mathtt{Corpus}|}{\mathtt{docfreq}_{\mathtt{term}} + 1} \tag{5.1}$$

A pair of images, **i** and **j**, can then be compared by calculating the cosine similarity of the vectors representing the images:

$$\cos(\mathbf{i}, \mathbf{j}) = \frac{\mathbf{i} \cdot \mathbf{j}}{\|\mathbf{i}\| \, \|\mathbf{j}\|} \tag{5.2}$$

We now demonstrate the outcome of comparing images represented using either a vector that concatenates the decomposed transformed VDR and bag-of-terms, or a vector that contains only the bag-of-terms. In this demonstration, each term has a *tf-idf* weight of 1. The first illustration (*Similar*) compares images that depict the same underlying action: Figure 5.1 (a) and (c). The second illustration (*Dissimilar*) compares images that depict different actions: Figure 5.1 (a) and (b).

$$\mathtt{Similar} : \cos(\mathrm{VDR}_a, \mathrm{VDR}_c) = 0.56 > \cos(\mathrm{Bag}_a, \mathrm{Bag}_c) = 0.52$$

$$\mathtt{Dissimilar} : \cos(\mathrm{VDR}_b, \mathrm{VDR}_a) = 0.201 \ll \cos(\mathrm{Bag}_b, \mathrm{Bag}_a) = 0.4$$

It can be seen that when the images represent the same action, the decomposed VDR increases the similarity of the pair of images compared to the bag-of-terms representation; and when images do not represent the same action, the decomposed VDR yields a lower similarity than the bag-of-terms representation. These illustrations confirm that Visual Dependency Representations can be used to distinguish the difference between

[1] The size of N varies as a function of the specific data split

[2] In this chapter, the *idf* values are calculated in the training data.

presence or absence of objects, and the prominent relationships between objects.

5.5 DATA

We use the data set of VDR-annotated images from Chapter 2 to study whether modelling the structure of an image can improve image retrieval in the domain of action depictions. The data set contains 341 images annotated with region annotations, three visual dependency representations per image (making a total of 1,023 instances), and a ground-truth action label for each image. An example of the annotations can be seen in Figure 5.1. The image collection is drawn from the PASCAL Visual Object Classification Challenge 2011 action recognition taster and covers a set of 10 actions (Everingham et al., 2011): riding a bike, riding a horse, reading, running, jumping, walking, playing an instrument, using a computer, taking a photo, and talking on the phone.

Image Descriptions

Recall that each image is associated with three human-written descriptions collected from untrained annotators on Amazon Mechanical Turk. The descriptions do not form any part of the models presented in the current paper; they were used in the automatic image description task in Chapter 4. Each description contains two sentences: the first sentence describes the action depicted in the image, and the second sentence describes other objects not involved in the action. A two sentence description of an image helps distinguish objects that are central to depicting the action from objects that may be distractors.

Region Annotations

The images contain human-drawn labelled region annotations. The annotations were drawn using the LabelMe toolkit, which allows for arbitrary labelled polygons to be created over an image (Russell et al., 2008). The annotated regions were restricted to those present in at least one of three human-written descriptions. To reduce the effects of label sparsity, fre-

quently occurring equivalent labels were conflated, i.e., man, child, and boy $\rightarrow$ person; bike, bicycle, motorbike $\rightarrow$ bike; this reduced the object label vocabulary from 496 labels to 362 labels. The data set contains a total of 5,034 region annotations, with a mean of 4.19 ± 1.94 annotations per image.

Visual Dependency Representations

Recall that each image is associated with three descriptions, and that people were free to decide how to describe the action and background of the image. The differences between how people describe images leads to the creation of one Visual Dependency Representation per image–description pair in the data set, resulting in a total of 1,023 instances. The process for creating a visual dependency representation of an image is described in Chapter 2.3.1. The annotated dataset comprises a total of 5,748 spatial relations, corresponding to a mean of 4.79 ± 3.51 relations per image.

Action Labels

The original PASCAL action recognition dataset contains ground truth action class annotations for each image. These annotations are in the form of labelled bounding boxes around the person performing the action in the image. The action labels are only used as the gold-standard relevance judgements for the query-by-example image retrieval experiments.

5.6 EXPERIMENTS

In this section we present the results of a query-by-example image retrieval experiment to determine the utility of the Visual Dependency Representation compared to a bag-of-terms representation. In this experiment, a single image (the query image) is used to rank the images in the test collection, where the goal is to construct a ranking where the top images depict the same action as the query image.

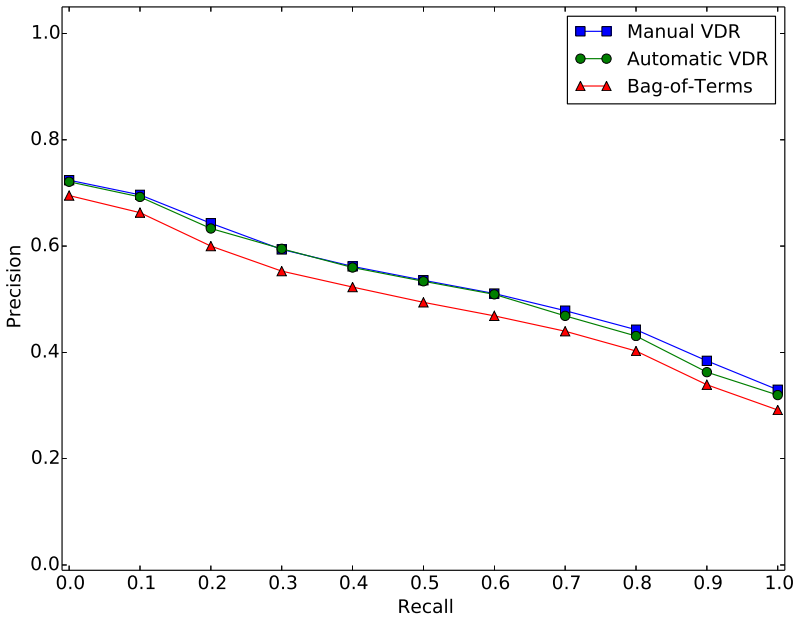


Figure 5.2 *Average 11-point precision/recall curves show that the VDR-based retrieval models are consistently better than the Bag-of-Terms model.*

5.6.1 *Protocol*

The image retrieval experiment is performed using 10-fold cross-validation in the following manner. The 341 images in the dataset are randomly partitioned into 80%/10%/10% splits, resulting in 1011 test queries[3]. For each query we compute average precision and Precision@10 of the ranked list, and use the resulting values to test the statistical significance of the results.

The *training set* is used to train the VDR prediction model and to estimate inverse document frequency statistics. During the training phase, the VDR-based models have access to region boundaries, region labels and three manually-created VDRs for each training image. In the *test set*, all models have access to the region boundaries and labels for each image. Each image in the test set forms a query and the models produce a ranked list of the remaining images in the test collection. Images are marked for relevance as follows: a image at rank r is considered *relevant* if it has the same action label as the query image; otherwise it is *non-relevant*. The *dev set* was used to experiment with different matching functions and to

[3]See Chapter 3.6.1 for more details on exactly how the data was split.

	MAP	P@10
Manual VDR	$0.514^{\star\dagger}$	$0.454^{\star}$
Automatic VDR	$0.508^{\star}$	$0.451^{\star}$
Bag-of-Terms	0.467	0.415

Table 5.1 *Overall Mean Average Precision and Precision@10 images. The VDR-based models are significantly better than the Bag-of-Terms model, supporting the hypothesis that modelling the structure of an image using the Visual Dependency Representation is useful for image retrieval.* $\star$*: significantly different than Bag-of-Terms at* $p < 0.01$*;* $\dagger$*: significantly different than Automatic VDR at* $p < 0.01$*.*

optimise the feature functions used in the VDR prediction model.

5.6.2 *Models*

We compare the retrieval accuracy of three approaches: Bag-of-Terms uses an unstructured representation for each image. A *tf-idf* weight is assigned to each region label in an image, and the cosine measure is used to calculate the similarity of images. This model allows us to compare the usefulness of a structured vs. unstructured image representation. Automatic VDR is a model using the VDR+IMG prediction method from Chapter 3, and Manual VDR uses the gold-standard data described in Section 5.5. Both of the VDR-based models have a tf-idf weight assigned to the transformed decomposed terms and the cosine similarity measure is used to calculate the similarity of images.

5.6.3 *Results*

Figure 5.2(a) shows the interpolated precision/recall curve and Table 5.2 shows the Mean Average Precision (MAP) and Precision at 10 retrieved images (P@10). The MAP of the Automatic VDR model increases by 8.8% relative to the Bag-of-Terms model, and a relative improvement up to 10.1% would possible if we had a better structure prediction model, as evidenced by Manual VDR. Furthermore, if we assume a user will only view the top results returned by the retrieval model, then P@10 increases by 8.6% when we model the structure of an image, relative to using an unstructured representation; a relative improvement of up to 9.4% would be possible if we had a better image parser.

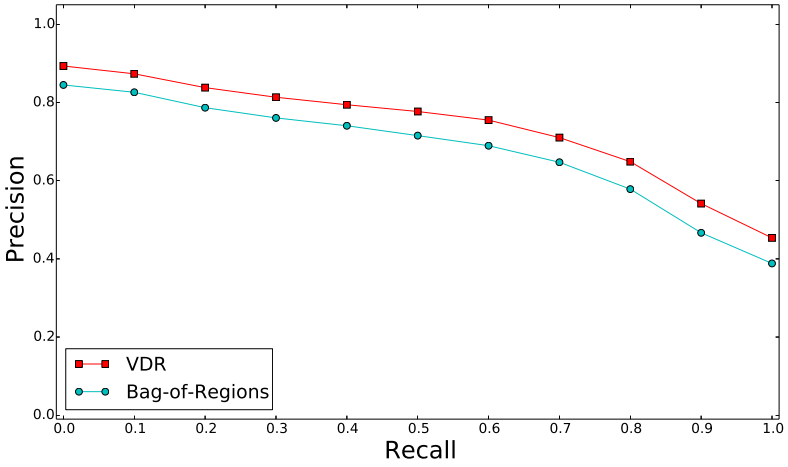
	MAP		P@10	
	VDR	Bag	VDR	Bag
Ride bike	0.721⋆	0.601	0.596⋆	0.513
Ride horse	0.833⋆	0.768	0.787⋆	0.726
Talk on phone	0.762⋆	0.679	0.666⋆	0.582
Play instrument	0.774⋆	0.705	0.634⋆	0.586
Read	0.483	0.454	0.498	0.475
Walk	0.198	0.186	0.184	0.174
Run	0.193	0.165	0.151	0.132
Jump	0.211	0.189	0.142	0.136
Use computer	0.814⋆	0.761	0.694⋆	0.648
Take photo	0.241	0.223	0.212	0.198

Table 5.2 *Mean Average Precision and Precision@10 for each action in the data set, grouped into transitive (top), intransitive (middle), and light (bottom) verbs. VDR is the Automatic VDR model and Bag is the Bag-of-Terms model. It can be seen that the Automatic VDR retrieval model is consistently better than the Bag-of-Terms model on both MAP and Precision@10. ⋆: the Automatic VDR model is significantly different than Bag-of-Terms at* $p < 0.01$.

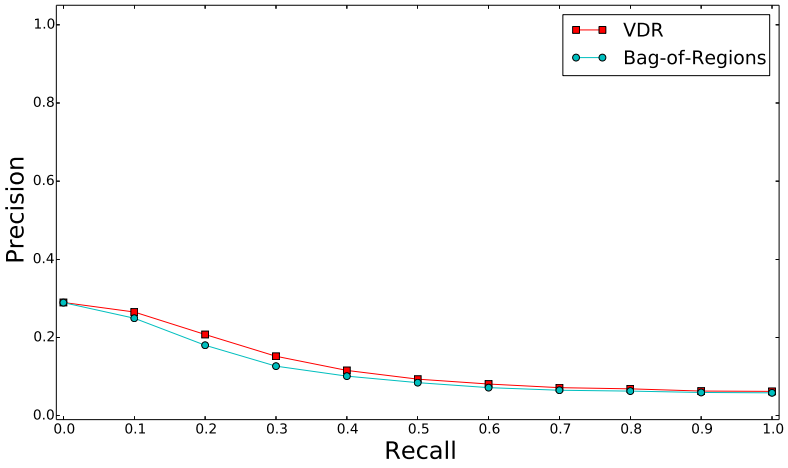
To determine whether the differences are statistically significant, we perform the Wilcoxon Signed Ranks Test on the average precision and P@10 values over the 1011 queries in our cross-validation data set. The results support the main hypothesis of this chapter: structured image representations allow us to find images depicting actions more accurately than the standard bag-of-terms representation. We find significant differences in average precision and P@10 between the Bag-of-Terms baseline and both Automatic VDR ($p < 0.01$) and Manual VDR ($p < 0.01$). This suggests that structure is very useful in the query-by-example scenario. We find a significant difference in average precision between Automatic VDR and Manual VDR ($p < 0.01$), but no difference in P@10 between Automatic VDR and Manual VDR ($p = 0.442$).

5.6.4 *Retrieval Performance by Type of Action and Verb*

We now analyse whether image structure is useful when the action does not require a direct object. The analysis presented here compares the Bag-of-Terms model against the Automatic VDR model because there was no



(a)



(b)

Figure 5.3 *Precision/recall curves grouped by the type of verb. The solid lines represent the Automatic VDR model; the dashed lines represent the Bag-of-Terms model; y-axis is Precision, and the x-axis is Recall. (a) Images depicting transitive verbs benefit the most from the Visual Dependency Representation and are easiest to retrieve. (b) Intransitive verbs are difficult to retrieve and there is is a negligible improvement in performance when using Visual Dependency Representation.*

significant difference in P@10 between the Automatic and Manual VDR models. Table 5.2 shows the MAP and Precision@10 per type of action. Figure 5.3 shows the precision/recall curves for (a) transitive verbs and (b) intransitive verbs.

In Figure 5.3(a), it can be seen that the actions that can be classified as transitive verbs benefit from exploiting the structure encoded in the Visual Dependency Representation. The only exception is for the action *to read*, which frequently behaves as an intransitive verb: *the man reads on a train*. The consistent improvement in both the entirety of the ranked list and at the top of the ranked list can be seen in the MAP and P@10 results in Table 5.2.

Figure 5.3(b) shows that there is a small increase in retrieval performance for intransitive verbs compared to the transitive verbs. We conjecture this is because there are fewer objects to annotate in an image when the verb does not require a direct object. The summary results for the intransitive verbs in Table 5.2 confirm the small but insignificant increase in MAP and P@10.

Finally, the light verbs, shown at the bottom of Table 5.2(c), exhibit variable behaviour in retrieval performance. One reason for this could be that if the light verb encodes information about the object, as in *using a computer*, then the computer can be annotated in the image, and thus it acts as a transitive verb. Conversely, when the light verb conveys information about the outcome of the event, as in the action *take a photograph*, the outcome is rarely possible to annotate in an image, and so no improvements can be gained from structured image representations.

5.6.5 *Discussion*

In our experiments we observed that all models can achieve high precision at very low levels of recall. We found that this happens for testing images that are almost identical to the query image. For such images, objects that are unrelated to the target action form an effective context, which allows this image to be placed at the top of the ranking. However, near-identical images are relatively rare, and performance degrades for higher levels of

recall.

It is surprising that image retrieval using automatically predicted VDR model is statistically indistinguishable from the manually crafted VDR model, given the relatively low accuracy of our VDR prediction model: 61.3% by the labelled dependency attachment accuracy measure. One possible explanation could be that not all parts of the VDR structure are useful for retrieval purposes, and our VDR prediction model does well on the useful ones. This observation also suggests that we are unlikely to achieve better retrieval performance by continuing to improve the accuracy of VDR prediction. We believe a more promising direction is refining the current formulation of the VDR, and exploring more sophisticated ways to measure the similarity of two structured representations.

5.7 CONCLUSION

In this chapter we argued that a limiting factor of retrieving images depicting actions is the unstructured bag-of-terms representation typically used for images. In a bag-of-terms representation, images that share similar sets of regions are deemed to be related even when the depicted actions are different. We proposed that representing an image using the Visual Dependency Representation (VDR) can prevent this type of misclassification in image retrieval. The VDR of an image captures the region–region relationships that explain what is happening in an image, and it can be automatically predicted from a region-annotated image.

In a query-by-example image retrieval task, we found that representing images as automatically predicted VDRs resulted in statistically significant 8.8% relative improvement in MAP and 8.6% relative improvement in Precision@10 compared to a Bag-of-Terms model. There was a significant difference in MAP when using manually or automatically predicted image structures, but no difference in the Precision@10, suggesting that the proposed automatic prediction model is accurate enough for retrieval purposes. Future work will focus on using automatically generated visual input, such as the output of the image tagger (Guillaumin and Mensink, 2009), or an automatic object detector (Felzenszwalb et al., 2010), which will make it possible to tackle image ranking tasks (Hodosh et al., 2013). It

would also be interesting to explore alternative structure prediction methods, such as predicting the relationships using a conditional random field (Zitnick et al., 2013), or by leveraging distributional lexical semantics (Le et al., 2013b).

CONCLUSIONS

6

The central claim of this thesis was that it would be useful to capture the relationships between image regions for tasks that involved understanding the action depicted in an image. We proposed to model the spatial relationships between image regions using the Visual Dependency Representation of images , which was introduced in Chapter 2. The central claim was tested in an automatic image description experiment in Chapter 4, and a query-by-example image retrieval experiment in Chapter 5. We found statistically significant improvements on these extrinsic tasks over unstructured baselines, and these improvements held when we used automatically predicted Visual Dependency Representations using the image parser from Chapter 3.

The main contribution of this thesis was the novel Visual Dependency Representation of images introduced in Chapter 2. This structured representation encodes the spatial relationships between regions of an image; it draws heavily from dependency syntax for natural language, and from studies showing that humans are better at recognising objects in images when the object is placed in a spatially consistent context (Biederman, 1972; Bar and Ullman, 1996). The Visual Dependency Representation makes it possible to distinguish between co-occurring image regions, and image regions that occur together to depict an action, by encoding the relationships between the regions using eight possible spatial relationships. In Chapter 3 we showed how to automatically predict the Visual Dependency Representation of an image using a statistical dependency parser (McDonald et al., 2005a) modified to exploit features from the image regions and parallel image descriptions. We found the best Visual Dependency Representation prediction performance when the parser extracted features from both the visual and linguistic modalities. However, we were unable to use this variant in our extrinsic evaluations because we either wanted to generate the linguistic modality (Chapter 4), or avoid it entirely (Chapter 5).

The first test of the central thesis claim was presented in an image description experiment in Chapter 4. We found significant improvements for both automatic evaluation measures and human judgements in the descriptions generated from a Visual Dependency Representation of an image compared to unstructured baselines. An important finding in this chapter was that Visual Dependency Representations also outperformed the baselines when relying on an external corpus to govern the generation the verb that relates a pair of objects. In this experiment we observed significant decreases in automatic measures and human judgements when replacing the parallel corpus with the external corpus. However, the decreased results were still significantly better than the state-of-the-art models that relied on the external corpus or spatial proximity.

The second test of the central claim can be found in Chapter 5, where we showed that the Visual Dependency Representation also improved the performance of query-by-example image retrieval. We compared our approach to a bag-of-terms baseline and found improvements in the quality of the entire ranked list, and in the top 10 images in the ranked list. In a post-hoc analysis, we found that Visual Dependency Representations were especially useful for finding images depicting transitive verbs, but there was no significant improvement for intransitive verbs.

The main limitation of this thesis was the use of gold-standard image region annotations. We decided to work with gold-standard data because automatic image annotation models are too noisy to study the potential value of a structured representation of an image. Several groups working in this area have used automatic object detectors based on the deformable parts model of Felzenszwalb et al. (2010), or learned correlations between visual features and image descriptions (Hodosh et al., 2013). In some cases, those contributions have used twenty pre-trained detection models (Yang et al., 2011), or have been restricted to evaluating on images where the object detector performance was deemed to be tolerable (Kuznetsova et al., 2012). Neither of these options seemed reasonable because we have more than 400 different types of objects in our data set (Chapter 2).

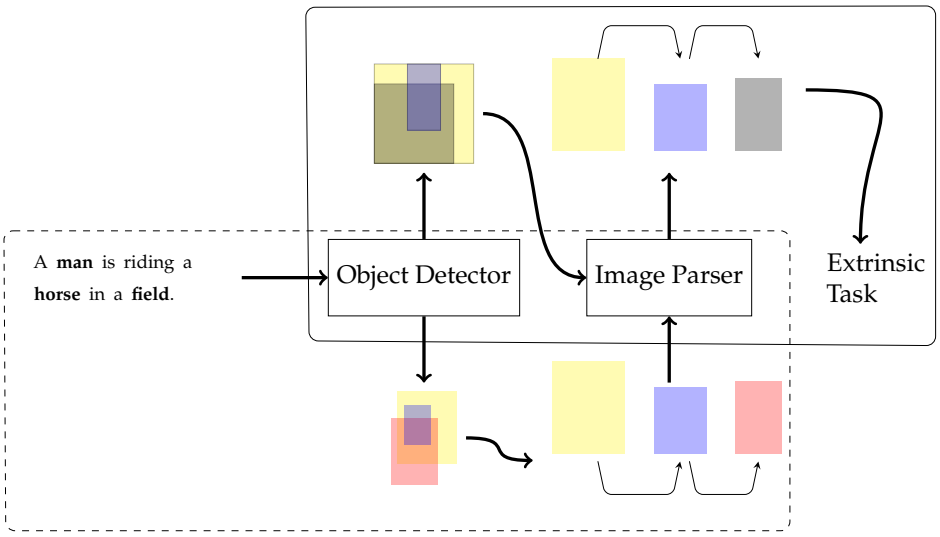


Figure 6.1 *An overview of the proposed future work. The nodes inside the dashed rectangle highlight future work on using an automatic object detector to predict the visual input, which is piped into a pre-trained image parser, and the predicted Visual Dependency Representation over the automatically predicted visual input is piped into the extrinsic tasks. The nodes inside the solid rectangle highlight work on using an image description to constrain an object detector to create training data from external images to train the image parser.*

6.1 FUTURE WORK

The annotation process described in Chapter 2 contains three steps: (a) collecting image descriptions, (b) image region anntotation, and (c) gold-standard Visual Dependency Representations. It is relatively cheap to obtain multiple descriptions of an image from Amazon Mechanical Turk, but expensive to train annotators for (b) and (c). Our future work will be focused on reducing the costs associated with human annotation.

The first point for future work is to reduce the reliance on gold-standard image region annotations. The deformable parts object detector (Felzenszwalb et al., 2010) would be the best fit for our overall framework of encoding the spatial relationships between labelled regions of an image. The most significant challenge is training a sufficient number of detection models to cover the range of objects annotated in our data set: we have over 400 types of annotated objects and there is only 20 pre-trained detection models. It may be possible to train additional detection models using the bounding-box annotated data in the ImageNet Large Scale Visual Recognition Challenge, or from the wider range of annotated data in

ImageNet. It will require a significant investment in time and expertise to optimise the parameters of visual detection models, but would provide a crucial understanding of how the Visual Dependency Representation can tolerate noisy computer vision models. An alternative approach would be to label the images with nouns from automatic image taggers (Lavrenko et al., 2003; Guillaumin and Mensink, 2009), which are generally more accurate than object detectors. Recall from Chapter 3 that our image parser is still a good predictor of image structure if we only have the labels of objects in the images.

If we can successfully integrate automatically extracted visual input into the process of predicting Visual Dependency Representation, then we can think about reducing the reliance on gold-standard Visual Dependency Representations to train the image parser. This would be especially useful if we are to evaluate on the Flickr8K (Hodosh et al., 2013) or SBU Captioned Photo Dataset (Ordonez et al., 2011). At training time, the image descriptions can be used to restrict the application of the pre-trained object detectors. The output of the object detectors are labelled bounding boxes, from which we can automatically predict the dependencies in the Visual Dependency Representation, and thus produce semi-supervised training data. The semi-supervised training data could either be used to supplement the gold-standard data, or to train an image parser from scratch. The trained image parser could then be used on the automatically predicted objects, as described above. In essence, this would create an (almost) fully automatic approach to the tasks studied in this thesis.

An additional avenue for future work is whether the useful spatial relationships captured in the Visual Dependency Representation generalise to images that do not depict actions. This could be in the form of scene type classification (Choi et al., 2010), which is well-studied and has some very competitive unsupervised results using the Spatial Pyramid Matching (Lazebnik et al., 2006) or scene gist (Oliva and Torralba, 2001).

Finally, it would be interesting to determine whether the salient object–object relationships encoded in the Visual Dependency Representation are actually useful beyond encoding the entire object–object relationship graph. Throughout this thesis, we have assumed that it will be useful to

construct Visual Dependency Representations that are, in essence, tree-like structures. However, it is obvious that a fully-connected graph can be trivially constructed by simply enumerating all possible object–object relationships. A candidate for exploring this avenue of future research would be the image retrieval experiments presented in Chapter 5.

Appendices

A

CHEAT SHEET

The cheat sheet is to be used for reference for experienced annotators and is not a substitute for reading the entire document.

- Always start from the first image description. The importance of this cannot be stressed enough due to its importance in evaluating consistency between annotators.
- Label an object with the first word used to refer to it.
  - An exception to this rule is when the initial reference to an object is completely incorrect. An example of a completely incorrect reference is describing a *car* as a *bike*, as compared to describing a car as a *vehicle*.
- If it is not clear which object is being referred to, *don't* guess, just skip it out and make a note of your decision.
- Don't spend too much time annotating an image. We do not expect it will take more than five minutes to annotate an image.
- Remember to reduce plural nouns to singular nouns when you decide to annotate individual objects. If a description refers to *trees* and you annotated individual trees then the label of those polygons should be *tree*.

A.1 INTRODUCTION

Image annotation is the task of drawing labelled polygons on an image. Annotation is often done with a fixed vocabulary of labels in mind, such as the twenty object classes in the PASCAL Visual Object Classification Challenge, or even is isolation of linguistic stimulus.

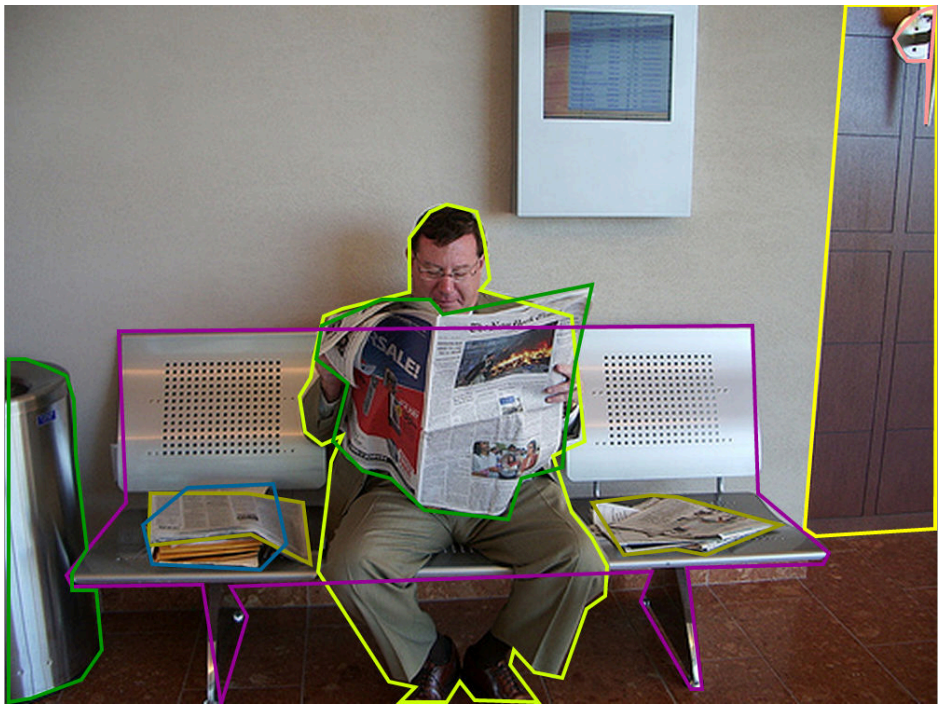
In this task, you will annotate images with the guidance of an image description. The description contains information about the type of action being performed, the actor and the object, and information about the context in which the action is taking place. An example of an image that has been annotated with the guidance of an image description can be seen in Figure B.1. This annotation process is briefly outlined below, with each step explained in more detail in the remainder of this document:

1. Find and verify the existence of the objects referred to in the image description.
2. Draw an accurate polygon on the image for each object in the description and label it.

A.2 NOUNS AND OBJECTS

Each image is presented alongside a pair of sentences which describe the image, as shown in Figure B.1. The first sentence describes the action taking place and the actor and the object involved in the action. The action is almost always a transitive verb, which means it requires both an actor and an object. The second sentence describes the context and any other interesting objects in the image. The first step is to read the image description, identify which objects occur in both the description and the image, and to earmark this objects for annotation.

The description in Figure A.1(b) refers to the following objects: man, bench, newspaper, papers, documents, bench, stainless steel waste can, bench, panelled wall, and scone light. The same bench is referred to three times and the bin, wall, and light are referred to using compound nouns. Where possible, compound nouns should be reduced to the head of the noun: panelled wall $\rightarrow$ wall, for example. Figure A.1(c) shows the list of



(a) An annotated image of a man reading a newspaper.

A man is sitting on a bench reading a newspaper.

Additional papers and documents are on the bench, a stainless steel waste can is beside the bench, and a panelled wall with scone light is in the background.

(b) A corresponding image description.

man, bench, paper, paper, bin, wall, light, newspaper

(c) The nouns extracted from the description.

Figure A.1 *An image (a) has been annotated with the guidance of a description (b). Accurate, although not perfect, polygons have been drawn around each noun or compound noun appearing in the description. The polygon labels are shown (c); the labels are simplified/generalised version of the original nouns/compound nouns in the description.*

nouns extracted from the image description.

You might come across a few corner cases:

- I cannot find an object from the description in the image...
  - Sometimes objects were hallucinated into existence by the people who wrote the image descriptions. See Figure A.2 for an example of a table that has been inferred but is not visible.
- A noun in the description is given in its plural form but I can clearly

see multiple instances of the noun in its singular form...

- If it is going to be easy to draw a polygon around each instance of the object, then reduce the noun to its singular form. See Figure A.4(a) for an example.
- If it is difficult to draw a polygon around individual instances of an object, then maintain the plural form. See Figure A.4(b) for an example.

A man is reading the newspaper.
There are a few people to his right, and a few tables and chairs.

Figure A.2 *The few tables in this image description have been inferred but they are not actually visible in the image. They are said to have been hallucinated and you do not need to annotate them.*

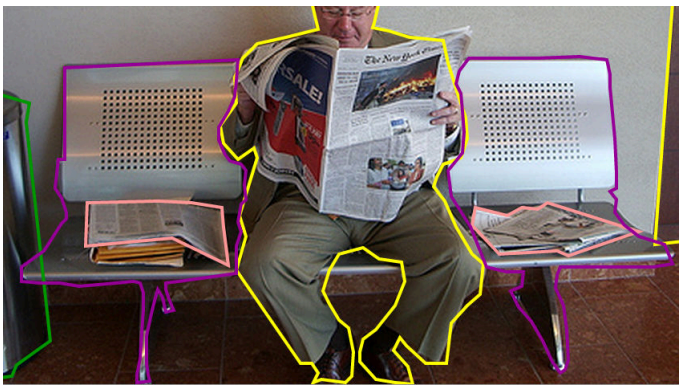


Figure A.3 *The bench is occluded by the man. If you were to follow the outline of the man, you would split the bench polygon in half, even though the bench is a single object. Instead, imagine the man was not there and draw the outline as shown in Figure A.1.*

(a) The description contains the word “trees” but it would be difficult to annotate each tree individually.

(b) The description contains the word “trees” but it is relatively easy to annotate each tree individually. The noun “trees” is reduced to “tree”.

Figure A.4 *An example of when to reduce a plural noun to a singular noun*

Polygons are drawn on the image using the LabelMe tool. LabelMe is a web-based tool for image annotation that will allow you to draw polygons an image using your mouse. It is as easy as pointing and clicking to create the outline of the polygon. When you have finished drawing the polygon, you will be prompted for a label, which you extracted from the description in the previous step. You might come across a few corner cases:

- A target object is occluded by another object.
  - If you can draw a complete polygon around the target object while following the line of the occlusion, then follow the occlusion line.
  - If the occluding object splits the target object into multiple polygons, just imagine the occluding object is not there. Figure A.3 shows an example of splitting a single object into two parts. Avoid this!

B

CHEAT SHEET

The cheat sheet is to be used for reference for experienced annotators and is not a substitute for reading the entire document.

- Do not attempt to create a visual dependency tree for an image where multiple people are performing an action. We will cover these types of trees after the guidelines have been revised.
- Pay special attention to indirect references such as he, her, they, and them when creating the tree.

B.1 INTRODUCTION

Image parsing is the task of producing a structured representation of an image. This structured representation, referred to as a *dependency graph*, is created with the guidance of a set of labelled polygons and an image description. The dependency graph is defined by labelling the geometric dependencies between the polygons. The image description was collected during a previous study, and the image has already been annotated with the set of labelled polygons.

In this task, you will create a dependency graph to represent the relationships between the annotated objects in an image. You will have access to the original image, the annotations, and the pair of sentences used to describe the image. An example of a parsed image can be seen in Figure B.1(a), given the annotated image in Figure B.1(b), and the pair of sentences in Figure A.1(b).

The remainder of this document outlines the geometric dependency grammar, how to use *dotty* to produce a dependency graph, and a step-by-step example of how Figure B.1(a) was produced.

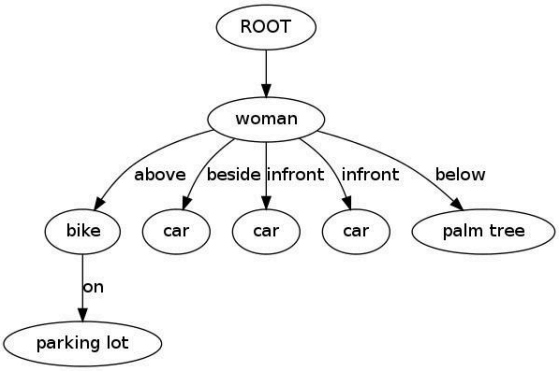
B.2 GEOMETRIC DEPENDENCY GRAMMAR

The Geometric Dependency Grammar, shown in Table B.1, defines the set of geometric relationships between pairs of objects in an image. Each relation is defined with some representative examples.

B.3 PROCESS

An image dependency graph is constructed in a series of steps, which define the geometric relationships between pairs of labelled polygons in an image. You will be presented with: the image; the set of labelled polygons; and the original image description.

All of the arcs are labelled with the guidance of the image description and the Geometric Dependency Grammar in Table B.1. It is important to note that the direction of an arc must be from the *head* to the *complement*. The head is the node that already exists in the graph, the complement is the



(a) The image dependency graph.

(b) An annotated image of a woman riding a bike.

A girl is riding a bike in a parking lot . Behind her are parked cars and palm trees .

(c) The image description with the annotated object labels highlighted.

Figure B.1 *An image (b) has been annotated with the guidance of a description (c). These labelled polygons have been used to create an image dependency graph (a), with respect to the annotations and the description.*

Relation	Description & Example
X $\overrightarrow{\text{on}}$ Y	Most of the pixels of polygon X overlap with polygon Y. In Figure B.1(b), the bike is **on** the parking lot.
X $\overrightarrow{\text{surrounds}}$ Y	Most of the pixels of polygon X overlap with polygon Y but X is much larger than Y. In Figure B.3(b), the couch **surrounds** the cat.
X $\overrightarrow{\text{beside}}$ Y	If the angle between the centre of mass of X and the centre of mass of Y lies between 315° and 45° or 135° and 225° then X is **beside** Y. In Figure B.3(a), the man is **beside** the keyboard.
X $\overrightarrow{\text{above}}$ Y	If the angle between X and Y lies between 45° and 135° then X is **above** Y. In Figure B.1(b), the girl is **above** the bike.
X $\overrightarrow{\text{below}}$ Y	If the angle between X and Y lies between 225° and 315° then X is **below** Y. In Figure B.1(b), the hedge (not annotated) is **below** the palm tree.
X $\overrightarrow{\text{infront}}$ Y	The Z-axis relationship between the objects is dominant. In Figure B.1(b), the girl is **infront** of the cars.
X $\overrightarrow{\text{behind}}$ Y	The Z-axis relationship between the objects is dominant. In Figure B.1(b), the car is **behind** the girl.
X $\overrightarrow{\text{opposite}}$ Y	Similar to *beside*, but used when there is a substantial distance between X and Y. In Figure B.1(b), the boy on the bike (not annotated) is **opposite** the cars.

Table B.1 *The Geometric Dependency Grammar defines seven relations between pairs of annotated polygons. Figure B.2 shows how the angles are defined. All relations are considered with respect to the centre of a polygon.*

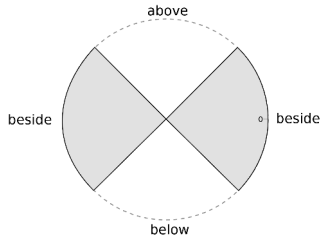
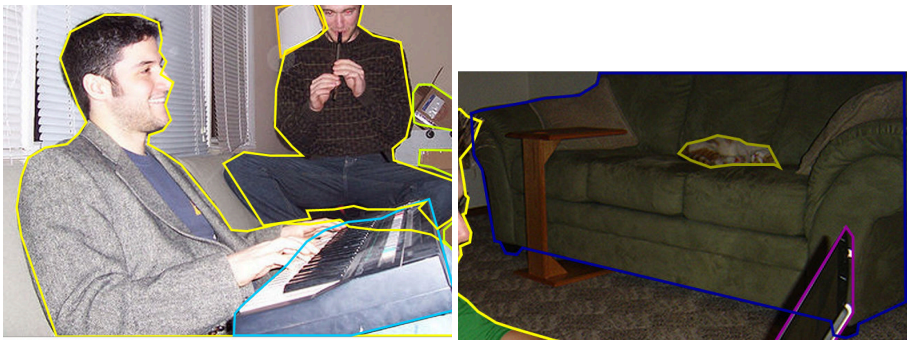
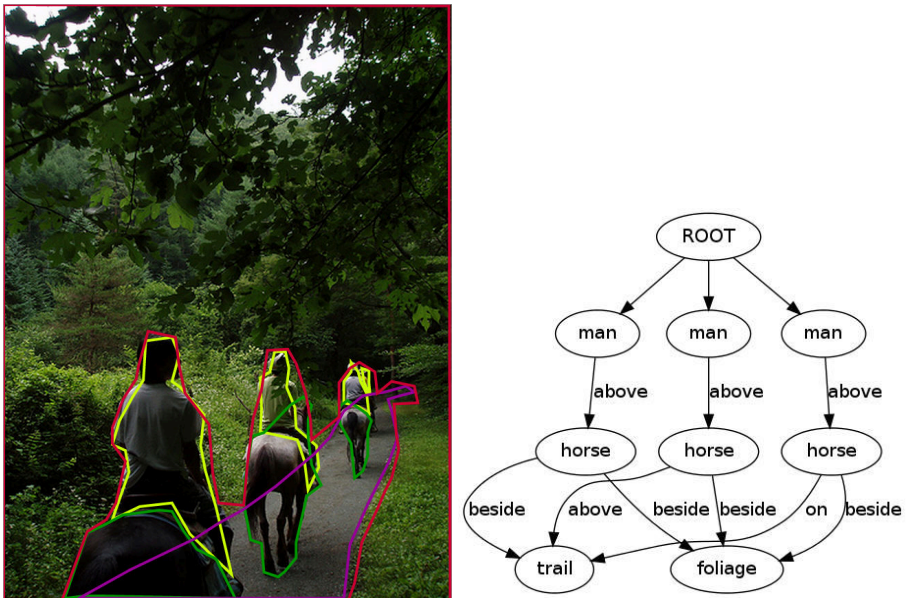


Figure B.2 *Geometric relationships inside the grey areas are beside, relationships outside the grey areas are above or below. 0° is shown on the right-side of the circle.*



(a) A man is **beside** a keyboard. (b) The couch **surrounds** the cat.

Figure B.3 *Supplementary examples to motivate the relations in the GDG.*



(a) An image with multiple subjects. (b) The corresponding dependency graph.

Figure B.4 *If there are multiple subjects in an image, and multiple labelled polygons, then connect multiple nodes to the root node.*

node you have just added to the graph.

1. Always start by drawing a ROOT node. It should be attached to the subject of an image with an unlabelled arc. An example of this can be seen in Figure B.6(b).

    - The subject of an image can usually be found near the centre of the image. This is an artefact of how people compose photographs. In most of your work, the subject of an image will be the person (or people) performing the action(s).

    - An image might have multiple subjects, as shown in Figure B.4. In this instance, draw as many nodes as there are subjects.

2. If the subject is engaged in a transitive action (an action that requires an object), draw a new labelled node on the graph and draw a labelled arc between the subject node and the new node. Figure B.6(c) is an example of expressing the relationship between the woman and the bike.

    - The arc is labelled with the geometric relationship between the centre of the subject and the centre of the object.

3. Work through the remaining polygons, that relate to the current description, and add them to the graph.

    - The next object in the description is the parking lot. A new node is added to the graph for this object and an arc labelled *on* is drawn from the bike to the parking lot.

    - The second sentence refers to *parked cars*, which the annotator has reduced to car, car, and truck. Three new nodes have been added to the graph and labelled arcs have been drawn between the **woman** node and these new nodes.

    - Finally, a node labelled **palm tree** is added to the graph and a labelled arc is drawn between the **woman** and this new node.

Note that you should attach objects in the second sentence to the ROOT node unless there is an explicit relationship expressed between those objects and the objects already ih the graph. A common sentence construction is "In the background, there are flags and a building". The flag and

building nodes should be attached to ROOT and not to any other node since no relationship is expressed in the sentence.

Finally, you should try to **not** represent the relationships between all pairs of objects, as shown in Figure B.5; but you should represent the relationships between the objects, with respect to the description, as shown in Figure B.1(a). In general, we are not interested in graphs; **however**, you might find it impossible to avoid drawing a graph when there are multiple subjects in an image (see Figure B.4).

B.4 DOTTY

dotty is a simple graph editor. Type **dotty** in a terminal to open the application. Note that you will need to disable Num Lock on the keyboard to use the application. dotty presents itself a small white screen, which is where you will draw the dependency graph.

B.4.1 *Drawing and labelling a node*

Note that most of the work in drawing and labelling nodes can be done with the xml2dot script. A left-click places a node in the graph. *Left-click on dotty to place a new node.* A node can be given a label by right clicking on the node, selecting **set attr**, and typing **label=xyz**, where **xyz** is the label you to assign to the node. *Right-click on the node you just placed, select* **set**

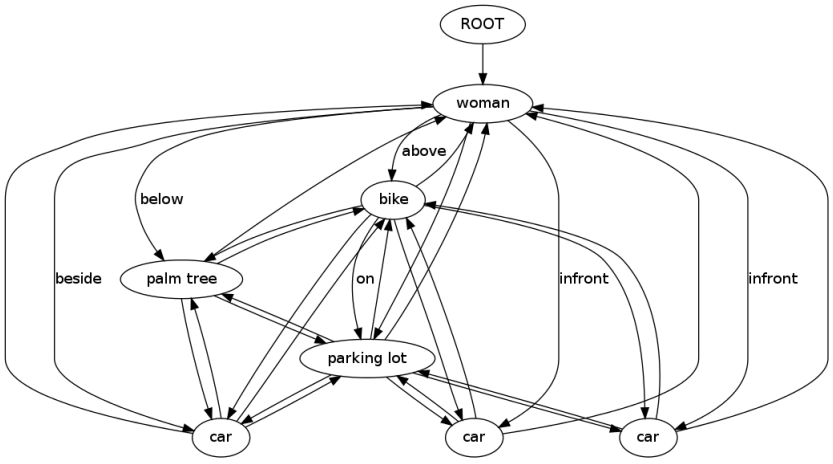
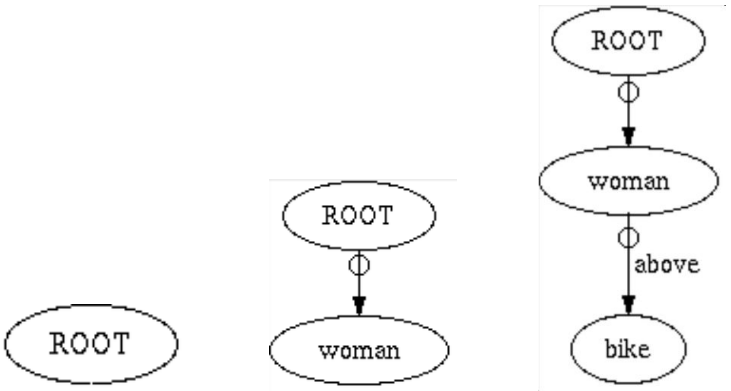


Figure B.5 *Try to avoid creating a fully-connected graph. We want to obtain graphs that use image descriptions to guide their configuration.*



(a) A single, labelled, node in dotty.

(b) Two nodes in dotty, with an arc between them.

(c) Three nodes in dotty, with an arc and a labelled arc.

Figure B.6 *How to use dotty to create a dependency graph*

attr, *type* **label=ROOT** *and press Return.*

The result of your interactions with dotty are not immediately visible. You need to right-click on an empty area of the dotty window and select **do layout**. *Move your mouse away from the node, right-click, and select bf do layout.*

You should see something similar to Figure B.6(a).

B.4.2 *Drawing an arc between two nodes*

To draw an arc, there needs to be at least two nodes in a graph. *Left-click in a blank part of* **dotty** *to add a second node to the graph, right-click on the new node, select* **set attr**, *type* **label=woman** *and press Return.* Remember that you won't see *woman* as the label for this newly created node until you select **do layout** from the right-click context menu.

Move the mouse to the node labelled ROOT, middle-click and drag from ROOT to the new node and release the middle mouse button. If you redo the graph layout, you should see something similar to Figure B.6(b).

B.4.3 *Drawing a labelled arc between nodes*

An arc between two nodes can be labelled or unlabelled. The previous section showed how to create an unlabelled arc. *Left-click to add a node to*

the graph, right-click on the new node, select **set attr**, *type* **label=bike** *and press Return, and then redo the layout. Move the mouse to the node labelled woman, middle-click and drag from woman to bike and release the middle mouse button. Redo the layout. Right-click on the small circle on the arc between the woman and bike nodes, select* **set attr** *and type* **label=above**. If you redo the graph layout, you should see something similar to Figure B.6(c).

B.4.4 *Saving a graph*

Right-click on an empty space of dotty and select **save as**. Type in a name for the graph you are creating and press Return.

Bach, F. R. and Jordan, M. I. (2002). Kernel Independent Component Analysis. *Journal of Machine Learning Research*, (3):1–48.

Bar, M. and Ullman, S. (1996). Spatial Context in Recognition. *Perception*, 25(3):343–52.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, 177(4043):77–80.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270, Ann Arbor, Michigan, U.S.A.

Choi, M. J., Lim, J. J., Torralba, A., and Willsky, A. S. (2010). Exploiting Hierarchical Context on a Large Database of Object Categories. In *2010 IEEE Conference on Computer VIsion and Pattern Recognition*, pages 129 – 136, San Francisco, CA, USA.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hauge.

Clark, J., Dyer, C., Lavie, A., and Smith, N. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, U.S.A.

Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, Washington, D.C., USA.

Dancey, C. and Reidy, J. (2011). Statistics Without Maths for Psychology. page 175. Prentice Hall, 5th edition.

Das, D. and Smith, N. A. (2009). Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 468–476, Suntec, Singapore.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, Florida, USA. Ieee.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, Edinburgh, Scotland, U.K.

Dixon, R. (2005). *A Semantic Approach to English Grammar*. Oxford University Press, Oxford, England, 2nd edition.

Duygulu, P., Barnard, K., de Freitas, J. F. G., and Forsyth, D. A. (2002). Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*, pages 97–112, Copenhagen, Denmark.

Elliott, D. and Keller, F. (2013). Image Description using Visual Dependency Representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, U.S.A.

Everingham, M., Gool, L. V., Williams, C., Winn, J., and Zisserman, A. (2010). The PASCAL Visual Object Classes Challenge. *International Journal of Computer Vision*, 88(2):303–338.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2011). The PASCAL Visual Object Classes Challenge 2011.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: generating sentences from images. In *Proceedings of the 15th European Conference on Computer Vision*, pages 15–29, Heraklion, Crete, Greece.

Fei-Fei, L., Fergus, R., and Perona, P. (2004). Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In *Workshop on Generative-Model Based Vision at the 2004 IEEE Conference on Computer Vision and Pattern Recognition*, pages 178–178, Washington, DC, USA.

Fei-Fei, L. and Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *2005 IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, San Diego, CA, USA.

Felzenszwalb, P. F., Girshick, R. B., 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.

Feng, Y. and Lapata, M. (2010). Topic Models for Image Annotation and Text Illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839, Los Angeles, California, U.S.A.

Gimpel, K. and Smith, N. A. (2009). Feature-rich translation by quasi-synchronous lattice parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 219–228, Edinburgh, Scotland, U.K.

Grauman, K. and Darrell, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1458—-1465, Beijing, China.

Grubinger, M., Clough, P., Müller, H., Deselaers, T., and Bank, W. (2006). The IAPR TC-12 Benchmark : A New Evaluation Resource for Visual

Information Systems. In *International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval at the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy.

Guillaumin, M. and Mensink, T. (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE 12th International Conference on Computer Vision*, pages 309–316, Kyoto, Japan.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Klein, D. and Manning, C. D. (2004). Corpus-based induction of syntactic structure: models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 478–485, Barcelona, Spain.

Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., and Guadarrama, S. (2013). Generating Natural-Language Video Descriptions Using Text-Mined Knowledge. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 541–547, Bellevue, Washington, USA.

Kübler, S., McDonald, R., and Nivre, J. (2009). *Dependency Parsing*. Morgan and Claypool Publishers.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1601–1608, Colorado Springs, Colorado, U.S.A.

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., and Choi, Y. (2012). Collective Generation of Natural Image Descriptions. In *Proceedings of*

the 50th Annual Meeting of the Association for Computational Linguistics, pages 359–368, Jeju Island, South Korea.

Lan, T., Yang, W., Wang, Y., and Mori, G. (2012). Image retrieval with structured object queries using latent ranking SVM. In *Proceedings of the 12th European Conference on Computer Vision*, pages 1–14, Firenze, Italy.

Lavrenko, V., Manmatha, R., and Jeon, J. (2003). A Model for Learning the Semantics of Pictures. In *Advances in Neural Information Processing Systems 16*, Vancouver and Whistler, British Columbia, Canada.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, New York, NY, USA.

Le, D., Bernardi, R., and Uijlings, J. (2013a). Exploiting language models to recognize unseen actions. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 231–238, Dallas, Texas, U.S.A.

Le, D. T., Uijlings, J., and Bernardi, R. (2013b). Exploiting language models for visual recognition. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 769–779, Seattle, Washington, U.S.A.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, U.S.A.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL '04*, pages 605–612, Barcelona, Spain.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, pages 1150–1157, Washington, D.C., USA.

Manning, C. D., Raghavan, P., and Schutze, H. (2008). Scoring, term weighting, and the vector space model. In *Introduction to Information Retrieval*, pages 100—-122. Cambridge University Press, 1st edition.

Marsh, E. E. and White, M. D. (2003). A taxonomy of relationships between images and text. *Journal of Documentation*, 59(6):647–672.

McDonald, R., Crammer, K., and Pereira, F. (2005a). Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 91–98, University of Michigan, U.S.A.

Mcdonald, R. and Nivre, J. (2011). Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005b). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada.

Minnen, G., Carroll, J., and Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Mensch, A., Berg, A., Berg, T., and Daum, H. (2012). Midge : Generating Image Descriptions From Computer Vision Detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France.

Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction in Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada.

Nister, D. and Stew, H. (2006). Scalable Recognition with a Vocabulary Tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2161—-2168, New York, NY, USA.

Nivre, J., Hall, J., and Nilsson, J. (2004). Memory-based dependency parsing. In *HLT-NAACL 2004 Workshop: Eighth Conference on*

Computational Natural Language Learning, pages 49–56, Boston, Massachusetts, USA.

Oliva, A. and Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems 24*, Granada, Spain.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL '02*, pages 311–318, Philadelphia, Pennsylvania, U.S.A.

Philbin, J., Chum, O., and Isard, M. (2007). Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA.

Quirk, C. and Menezes, A. (2005). Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 271–279, Sydney, Australia.

Radford, A. (2004). *English Syntax: An Introduction*. Cambridge University Press.

Randell, D., Cui, Z., and Cohn, A. (1992). A spatial logic based on regions and connection. In *Principles of Knowledge Representation and Reasoning*, pages 165–176.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using Amazon's Mechanical Turk. In *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk at 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 139–147, Los Angeles, California.

Regneri, M., Rohrbach, M., Wetzel, D., and Thater, S. (2013). Grounding Action Descriptions in Videos. *Transactions of the Association of Computational Linguistics*, 1:25–36.

Reiter, E. and Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 4(35):529–558.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77(1-3):157–173.

Sadeghi, M. A. and Farhadi, A. (2011). Recognition Using Visual Phrases. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1745–1752, Colorado Springs, Colorado, U.S.A.

Saxena, A., Chung, S. H., and Ng, A. Y. (2006). Learning Depth from Single Monocular Images. *Neural Information Processing Systems 18*, 18:1161–1168.

Shatford, S. (1986). Analysing the Subject of a Picture: A Theoretical Approach. *Cataloging & Classification Quarterly*, 6(3):39–62.

Shi, J. and Malik, J. (2000). Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

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*, pages 572–582, Sofia, Bulgaria.

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, South Korea.

Smeulders, A. and Worring, M. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1–32.

Smith, D. A. and Eisner, J. (2006). Quasi-synchronous grammars: alignment by soft projection of syntactic dependencies. In *Workshop on Statistical Machine Translation in Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 23–30, New York City, New York, U.S.A.

Smith, D. A. and Eisner, J. (2009). Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 822–831, Suntec, Singapore.

Snover, M., Dorr, B., and Schwartz, R. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA.

Steedman, M. (1996). *Surface Structure and Interpretation*. MIT Press.

Tesnière, L. (1953). *Esquisse d'une syntaxe structurale*. Librairie C. Klincksieck.

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychologial Review*, 113(4):766–786.

Turgenev, I. (1862). *Fathers and Children*. The Russian Messenger, Moscow, Russia.

Wang, M. and Smith, N. (2007). What is the Jeopardy model? A quasi-synchronous grammar for QA. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 22–32, Prague, Czech Republic.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Wu, L., Jin, R., and Jain, A. K. (2012). Tag Completion for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):716–727.

Yang, Y., Teo, C. L., Daumé III, H., and Aloimonos, Y. (2011). Corpus-Guided Sentence Generation of Natural Images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454, Edinburgh, Scotland, UK.

Yu, C.-N. J. and Joachims, T. (2009). Learning structural SVMs with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1169–1176, Montreal, Quebec, Canada.

Zhang, Y., Jia, Z., and Chen, T. (2011). Image retrieval with geometry-preserving visual phrases. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 809–816, Colorado Springs, Colorado, U.S.A.

Zitnick, C., Parikh, D., and Vanderwende, L. (2013). Learning the Visual Interpretation of Sentences. In *IEEE International Conference on Computer Vision*, pages 1681–1688, Sydney, Australia.