EXPANDING COMMONSENSE KNOWLEDGE BASES BY LEARNING FROM IMAGE TAGS

BY

CECILIA MAUCERI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Advisor:

Professor Svetlana Lazebnik

Abstract

I present a method for learning new commonsense facts to augment existing commonsense knowledge bases by using the metadata of large online image collections. Online image collections present a source of knowledge that is supported by many contributors, has good representation of objects and their properties, and is visual. The collection's broad support of objects and object properties ensure the relevance and quality of the commonsense knowledge collected, while the visual focus provides a different subset of knowledge than typical text corpora. Using the image metadata provides a text representation of the visual information. Therefore, I can use classifiers trained on existing text-based knowledge bases to learn relationships between concepts represented in the images. I collect two datasets of more than 1 million images each, one consisting of animal images, one of room interiors. The images are tagged with relevant concepts by their owners. I train classifiers using facts from two popular commonsense knowledge bases, ConceptNet and Freebase, to classify the relationships between frequent concept pairs. The output is a list of more than 90,000 proposed facts, which are in neither source knowledge base.

Acknowledgments

Thanks to my adviser, Svetlana Lazebnik, for all your helpful advise and constructive critism that kept the project moving. Also many thanks to my husband, Steffen, for sharing countless cups of coffee and encouragement, and to my parents for always offering grammatical support and love.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

MP@1	Mean Precision at 1 (i.e. considering only the highest scoring classifier)
MP@n	Mean Precision at n (i.e. considering all n top scoring classifiers)
MR@1	Mean Recall at 1 (i.e. considering only the highest scoring classifier)
MR@n	Mean Recall at n (i.e. considering all n top scoring classifiers)
NPMI	Normalized Pointwise Mutual Information
PMI	Pointwise Mutual Information
R@1	Recall at 1
R@n	Recall at n (i.e. considering all n top scoring classifiers)
SVM	Support Vector Machine

Chapter 1

Introduction

Commonsense knowledge describes the everyday facts that humans learn through experience. Examples include simple observations such as that a ball will roll down a slope, or that sugar is sweet. Shared commonsense knowledge is an important starting point for making sense of observations. If we observe a ball rolling uphill, we assume there is some hidden agent pushing it upwards. This hypothesis may be false, perhaps we are observing the ball in an unconventional reference frame, but without commonsense knowledge, any reasoning about the scene would be like reasoning about a surrealist painting.

Similarly, artificial intelligence systems can draw on models of their environment in the form of commonsense knowledge bases to make sense of observations. Commonsense knowledge bases have well established uses in text parsing [34], word sense disambiguation [8] and improving topic modeling [28]. Such knowledge bases are often represented by a graphical network. The nodes represent concepts and edges represent the relationships between concepts. This allows for a complexly interconnected representation of the system of facts. The more complete the representation, the better the system can reason about new observations. Two popular knowledge bases with this format are Freebase and ConceptNet. Freebase is used by Google to summarize topics and suggest related search results [20]. ConceptNet is designed to support commonsense reasoning [22].

Commonsense knowledge bases are far from complete. Dong et al. found that the birthplace of 71% people on Freebase was unknown, as was the nationality of 75% [15]. Additionally, Freebase draws its facts from structured data on the web, such as Wikipedia, and therefore has more knowledge about concrete nouns, but it has less information about abstract topics such as emotions. Other databases have similar gaps resulting from their limited sources of knowledge and their limited ability to either manually or automatically extract the knowledge.

Traditional commonsense knowledge bases gather facts from text resources and stored facts in a text representation. Recent research has addressed expanding commonsense knowledge bases to include other modalities. While text-based commonsense knowledge is very useful for interpreting and parsing language, text is a representation that is specific to language. In many situations, it is easier to work directly with other representations rather than glossing them with text. For example, if the system is asked to describe how to manipulate an object based on visual input, it needs to identify possible joints, grasping points, and degrees of freedom. A text-based description would be long and cumbersome. An easier approach would be to mark the joints and grasping points in the visual space. Additionally, even when working with text, a correspondence is needed for interpreting text in terms of the visual signal and vice versa. For example, if the system is asked a question, "Is the teapot on the table?", it must visually locate the teapot and the table and understand what "on" looks like.

This work presents a strategy for automatically collecting visual commonsense knowledge to augment language-grounded commonsense knowledge bases. The goal is to identify relationships between visual concepts. For example, the concepts "teapot" and "table" are related by their relative locations. To learn similar facts, I start with the intuition that commonsense knowledge should be frequently represented in images, e.g. I should see many teapots on tables. I collect frequently co-occurring concept pairs from image tags and train classifiers to learn relationships similar to those found in existing knowledge bases.

Using images as the source of knowledge, rather than news articles or other text corpora, focuses fact collection on visual concepts. Visual concepts are often less frequent in text corpora usually because the visual context is considered obvious and an exhaustive visual description would be too lengthy. For my method, I use image tags which are often chosen by the image's owner to boost its appearance in relevant search results. While they are not an exhaustive description, the image's tags often evoke specific visual imagery. For example, given the short list, "cat", "sofa", "sunshine", and "serenity", every reader can imagine how a corresponding scene might look.

The process of using one source, e.g. text, to learn about another, e.g. images, is referred to as transfer learning. I use a training set of facts from existing knowledge bases compiled from text sources to learn relationship models. I then transfer the models to candidate facts collected from images. Transfer learning is challenging for several reasons. Different sources contain different distributions of information and usually have different representations. In this case, image tags can be used to provide a shared text representation for the two sources. However, the problem remains that the sources have very different distributions. Some of the tags for the images do not appear in the existing knowledge bases. This means the learned classifiers must be powerful enough to generalize to unknown concepts. In experiments, the classifiers have a mean 83% accuracy on cross-validation tasks, and mean 64% accuracy on unknown facts. While this is a significant drop, if a small number of facts can be identified with high confidence, humans can confirm the accuracy of the new facts before adding them to existing knowledge bases.

This work can greatly increase the completeness of the popular commonsense knowledge bases Concept-

Net and Freebase. I collect two datasets of more than 1 million images each with two themes, domestic animals and room interiors. Using classifiers trained on an average of 3,653 relevant facts retrieved from ConceptNet and Freebase, my experiments produce 69,310 new fact proposals for an animal-themed vocabulary and 40,131 new fact proposals for a room-themed vocabulary. If the preliminary result of 64% accuracy holds for these edges, the knowledge base would grow by 700%.

The primary contributions of this work are:

- Developing the idea of using transfer learning for visual commonsense knowledge.
- Collecting two datasets with more than 1 million images each and establish a procedure for extracting a vocabulary of concepts and a list of candidate facts. (Section 4)
- Presenting a method to transfer relationships from existing knowledge bases to frequent the candidate facts. (Section 3)
- Proposing 93,850 facts which are not currently in ConceptNet or Freebase. (Section 5)

Chapter 2

Related works

2.1 Information extraction

Information extraction methods support finding structured facts from unstructured source data, usually text. One of the most famous of these systems is NELL, the Never-ending Language Learner [25]. NELL builds on an initial ontology, learning new relationship types, concepts, and facts from reading websites. Mitchell et al. report that NELL has learned over 80 million facts since January 2010 [25].

NELL learns classifiers for each relationship type from the context patterns extracted from text using the Coupled Pattern Learning [6] and OpenEval [32] systems. Both systems use sophisticated frequent pattern mining to retrieve pairs of concepts that have the given relationship. The classifiers used in information extraction vary widely. Ritter et al. [29] and Dong et al. [15] use probabilistic models to expand exiting knowledge bases. Chang et al. [7] show how to use matrix factorization to transfer relationships between similar concepts. I use a much simpler one-vs-all classifier.

2.2 Visual knowledge

Visual knowledge is not well represented in text. Where it is represented, it requires additional work to identify and extract [13]. Joint image and text models have been shown to improve iconic image detection [18] and image annotation [38, 19]. Ontologies, classification systems similar to commonsense knowledge bases, have been used to improve object classification [11] and image annotation [36] as well. The value in collecting visual knowledge is significant.

2.2.1 Visual knowledge bases

There are few existing commonsense knowledge bases that specifically target visual information. These include the Never-ending Image Learner (NEIL) inspired by NELL [9], Robobrain [33], and VisKE [31]. NEIL and Robobrain collect visual concepts and relationships that correspond to a fixed set of classifiers.

For example, NEIL uses the intersection over union of two object detectors to represent the relationship "part of". VisKE is much more flexible. It has the ability to discover the validity of any relationship between two concrete nouns. The knowledge discovery method balances these two approaches. It can learn any relationship type from an existing knowledge base, making it more flexible than NEIL and Robobrain, but also more focused than VisKE. Additionally, it is not limited to noun relationships as VisKE is, but can also learn relationships between nouns and adjectives.

ImageNet, a structured collection of images with an average of 1000 images per concept[30], could also be considered a commonsense knowledge base with only hierarchical relationships between concepts. ImageNet relationships describe the broader categories concepts belong to, for example, "domestic cat is a feline" or "chair is a furniture". The method supports many more relationship types and provides images corresponding to facts, such as "cat has property black", instead of concepts, such as "cat". Additionally, ImageNet is constructed using manual annotation, I provide an automated method for finding images for concepts.

2.2.2 Image semantics

Visual commonsense knowledge bases are an extension of research in image semantics. Semantics refers to the study of meaning. Image semantics refer to the visual traits that are used to identify and describe an object or scene as well as the interactions between objects in the scene. For example, a particular object is visually identified as a "cat" because of its shape, color and texture. Additionally, it might be a "cute cat" or a "fuzzy cat". "Cuteness" and "fuzziness" are visual properties. Other properties, such as the name of the cat, are non-visual and are not part of the image semantics. The cat's location in the scene, e.g. "on the sofa", and its behavior, e.g. "sitting", can also be part of the image semantics. Image semantics are usually studied in separate pieces, such as object detection, attribute detection, pose or action detection, and modeling interaction. Image understanding systems seek to put all these pieces together for a comprehensive explanation of the structure of the image.

Of these many pieces, one of the most important inspirations for the visual knowledge discovery method is attribute detection. Object attributes describe the properties of the object without naming the object explicitly. For example, the attributes "has four legs" and "has fur" identify an object as an animal without using the name "animal". Attributes can be any visual semantic properties, such as shape, texture, color, or even higher level properties such as parts, actions, pose, and materials. The seminal works in the field of object attributes are Lampert et al.[21] and Farhadi et al.[16]. Lampert et al. proposed using high-level attributes rather than object classes to describe images because the attributes generalized better to unseen classes. Following closely afterwards, Farhadi et al. focused on automatically learning features that are discriminative for an attribute within a class. Attribute detection and attribute features are a major area of research with applications in object detection, image retrieval, and image description. Attribute detection is usually studied as a fully supervised or weakly supervised problem using image features with attribute labels and sometimes bounding boxes. My method takes a weakly supervised approach to learn attributes that correspond to the knowledge base relationships.

Commonsense knowledge emerges from attributes when those attributes can reliably describe a concept across a large collection of images. For example, "is pink" might describe a cat in one or two pictures, but in general, the attribute "is pink" would not help identify a cat. Therefore, it is not a reliable commonsense attribute. Instead, I am looking for attributes that accurately summarize a large collection of images. This is closely related to the fields of image collection summarization and iconic images. Iconic images are images that show canonical views of an object or scene. Usually, they are selected so that a small collection of iconic images can provide a clear summary of the variation present in the object or scene class. While some iconic image methods rely entirely on visual composition models, [3, 14], others include image metadata to produce more complete semantic models [27, 18]. I take inspiration from Raguram and Lazebnik and Gong et al., to use image metadata to identify frequent attributes.

Chapter 3

Method

To start, I introduce the terminology. The *vocabulary* is the set of words or phrases that I am interested in learning about. The elements of the vocabulary are referred to as *concepts*. The commonsense knowledge facts are stored in the form of directed *edges* of a commonsense graph. Each *edge* connects a pair of concepts. The start of the edge is the *source concept*. The end of the edge is the *target concept*. The edge also has a *relationship* label. For example, the edge "kitten is a cat" has the source "kitten", the target "cat", and the relationship label "IsA".

The method has five major steps.

1. Collect a dataset of several million images using a small set of search terms (Section 3.1)

2. Establish a vocabulary (Section 3.2)

    a. Mine the dataset for frequent image tags

    b. Filter unwanted tags from the list and split concatenated phrases

    c. Expand the vocabulary to include frequent phrases

    d. Label all concepts with part of speech, language, and category labels

3. Count the pairwise co-occurrence of concepts (Section 3.4)

4. Collect ground truth relationship labels from ConceptNet and Freebase (Section 3.5)

    a. Retrieve a large set of edges for each concept in the vocabulary.

    b. Select a set relationship types to learn

    c. Filter the edges using the part of speech, language, and category labels, and concept co-occurrence frequency

5. Use a multi-class SVM to predict relationship labels for an edge (Section 3.6)

    a. Represent edges using GloVe feature vectors

b. Train on high frequency edges with ground truth

c. Test on high frequency edges without ground truth

3.1 Collect image dataset

Establishing the vocabulary requires a large set of images with metadata. A randomly selected set of images may not contain any clear frequent patterns from which to learn. Therefore, I focus image retrieval using a small set of *search terms* which all belong to some common category, for example, a list of domestic animals or rooms in homes. I use the Flickr API [1] to collect images containing one of the category search terms in their title, description, or tags. I gather between several hundred thousand and one million images for each search term. The goal is to have one million images for each search term, but in some cases, it is difficult to retrieve so many. Therefore, the distribution of the retrieved images reflects the search term's overall frequency in Flickr. The combined set of images retrieved for all the search terms is referred to as the *dataset*.

3.2 Establish a vocabulary

3.2.1 Mine the dataset for frequent image tags

I select vocabulary from the tags of the images. The tags are unordered words or phrases which the image owner chooses to describe the image. Flickr also provides a few automatically generated tags based on visual classifiers. For example, Flickr automatically tags images with human faces, people, and outdoor scenes.

I define the vocabulary with a simple bag of words model. Each image's tag set is tokenized, i.e. split into individual tags. The frequency of each token is counted over the dataset. Single prolific users can have a large influence on what tokens are frequent in the dataset. Intuitively, I do not want exotic cat names, like 'Pickles', to skew frequency results because the cat's owner is a prolific contributor. Therefore, the token occurrences are counted per user rather than per image.

I use the tokens which meet a frequency threshold as the *initial vocabulary*, excluding all English stop-words, numbers and tokens composed entirely of non-roman characters from the vocabulary. I aim for an initial vocabulary of a couple thousand concepts to generate a large number of interesting edges.

3.2.2 Cleaning up the initial vocabulary

All of the concepts in the initial vocabulary are tokenized tags as retrieved from Flickr. This leads to some confusion because Flickr treats tag phrases ambiguously. Flickr either splits all the words into separate tags or it removes the spaces to create a single tag. For example, “persian cat”, might be tagged in some images as two tags, {“persian”, “cat”} and in others as a single tag, {“persiancat”}. I attempt to identify both these patterns as a single concept, “persian cat”. I have two strategies for identifying concepts with missing spaces: (1) comparing combinations of concepts and (2) searching the knowledge base.

For each concept, c, in the vocabulary, I check whether there is a pair of concepts in the vocabulary which when concatenated are equivalent to c. If there are two possible ways to split a concept, such as 'kitchen sink' or 'kitchens ink', I take the split which has the more frequent component concepts. I only check pairs of concepts because the comparison is combinatorial and hence very expensive.

I also search Freebase for all of the concepts in the vocabulary. To distinguish between different word senses, Freebase uses topic nodes, each one representing a different word sense. Each topic has several aliases or synonyms that are used to describe the same topic. For example, “new york city” has the aliases, “new york, new york”, “new york”, “nyc”, “city of new york”, and “the big apple”, among others. I remove the spaces from the aliases and compare them to the vocabulary. If one of the aliases for the retrieved topics is equivalent to a concept, I replace the concept in the vocabulary with the original alias with spaces.

Automatic tags also add noise to the initial vocabulary. Many applications and websites that post images to Flickr automatically tag the images with generated tags. Flickr, additionally, automatically tags images based on object detectors. I keep the object detector tags, but remove any other automatic tags, because such concepts are very frequent in the dataset and do not refer to the visual content of the images. Likewise, I apply filters to remove concepts related to photography, such as the makes and models of cameras and film, variations of photo, photography, or photographer, and the names of photography apps and websites. I refer to the vocabulary after this filtering step as the *filtered vocabulary*.

3.2.3 Expand the vocabulary to include phrases

Sometimes, tokenizing the tags breaks up phrases that should be treated as single token, such as ‘Maine Coon Cat’ or ‘New York City’. I already identified some phrases when splitting concatenated tags, but there may also be phrases in the vocabulary which do not occur in a concatenated form. To recover these phrases, I use two methods. I merge concepts with high pointwise mutual information and I perform a local search on the knowledge base.

Pointwise mutual information (*pmi*) is a statistical measure of dependence of two outcomes, $x \in X$ and

$y \in Y$ where X and Y are discrete random variables.

$$pmi(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

PMI is used in natural language processing to measure correlation while normalizing to account for the overall frequency of x and y [10]. For example, the concept "shearing" may always occur with "sheep", resulting in a very high conditional probability. However, "sheep" is very frequent overall in the dataset. Therefore, the PMI is low. Concepts with high PMI are more likely to be part of a phrase.

Normalized pointwise mutual information (*npmi*) defined on the range $[-1, 1]$ with -1 corresponding to x and y never occurring together, 0 corresponding to complete independence and 1 to complete co-occurrence.

$$npmi(x, y) = \frac{pmi(x, y)}{-\log(p(x, y)}$$

If the normalized pointwise mutual information of two concepts in the dataset is above 0.8, I consider it highly likely that the concepts are part of a phrase. I search Freebase for both orderings of the two concepts. If one of the orderings is in Freebase, I add the phrase to the vocabulary.

For all statistical techniques, including PMI, I use Laplace smoothing to prevent zero probabilities. I model the additive quantity as a single image from a unique owner tagged with all the concepts in the current vocabulary. This produces a small positive probability $\frac{1}{n+1}$ for every pair of concepts where n is the total number owners.

Additionally, for each concept in the vocabulary, I check all of the adjacent edges in Freebase. If I find an adjacent concept composed of words already in the vocabulary, I add it as a new concept phrase. For example, if the vocabulary contains {"cat", ..., "maine", ..., "coon"...} and I find the edge "maine coon cat is a cat", I would add the phrase "maine coon cat" to the vocabulary.

Freebase is used to check the existence of a phrase because valid phrases are likely to be proper nouns, e.g. "new york city" and "maine coon cat", while erroneous phrase, e.g. "blue eyes", are likely to be descriptive. Freebase has excellent coverage of proper nouns, and therefore is fairly reliable as a resource for this task. It returns mixed results when the phrase has multiple word senses, one of which is a proper noun, for example "Sunset Park", a neighborhood in Brooklyn, or "Black Sheep", a hiphop artist.

Any new concepts must meet the frequency threshold to be added to the vocabulary. I refer to the vocabulary including the phrases as the *expanded vocabulary.*

3.3 Labels for concepts

I collect a variety of other properties of each of the concepts to facilitate relationship classification and cleaning the training set. These include part of speech, language, locations, and colors. For each concept, I collect part of speech labels, including nouns, verbs, and adjectives, from WordNet[17], via the Python Natural Language Toolkit (NLTK)[4]. This resource provides labels for an average of 86% of the vocabulary.

I also collect language labels from WordNet to identify non-English words. While the approach should be flexible for other languages as well, this implementation does not tackle non-English concepts. NLTK Wordnet interface provides dictionaries for 16 languages. I label a concept as non-English if it meets three criteria: (1) it is not in the English dictionary, (2) it is in one of the non-English dictionaries, and (3) it is not a proper noun. The third rule is necessary because most proper nouns, such as “Paris” or “France”, do not appear in the English dictionary. I use the WordNet “instance hypernym” relationship to identify a list of proper nouns.

I additionally collect Freebase type categories for each concept. The “location” types are especially useful for identifying geographic locations. Many of the concepts are geographic locations because users like to tag their images with the location where they were taken. For concepts with the location type, I also collect population statistics if available. The population is a good signal for word sense disambiguation. If the population is large, the word is more likely being used to tag a location. If the population is low and another word sense is available, the concept might not refer to the location. I use Wordnet to provide alternate word senses.

Another useful Freebase category is “color”. I am especially interested in color because it is an intrinsically visual concept. Usually, instances of color should be in the “has property” relationship. This would seem to be a straight forward case of identifying colors as adjectives. However, most colors can also be used as nouns, eg., “red is the color of a rose”. Therefore, the part of speech labels from WordNet are not as useful for filtering colors from other relationships. While the edges containing color tags collected from ConceptNet or Freebase are generally reliable and do not require much additional filtering, color filters can be helpful in eliminating proposed concept pairs in the test set, where the overall frequency of color in the dataset generates a lot of noise.

3.4 Pairwise concept co-occurrence

Once the expanded vocabulary is defined, I look at pairwise concept frequency to generate a list of high frequency *concept pairs* which may be edges. For example, if the concepts “cat” and “sofa” occur together frequently, there is likely some relationship between them. This pair is not yet considered an edge because it is un-ordered and has no relationship label. As with the vocabulary, the frequency of concept pairs can be counted in two ways, by the number of images in which both tags appear or in terms of the number of image owners who use the tags as a pair to describe an image. Counting pairs of concepts is more memory intensive than counting single concepts. I present several strategies for memory efficient counting.

To count images, a bag of words feature can be constructed for each image. These are then concatenated into a n by m term document matrix, D where n is the size of the vocabulary and m is the number of images. The matrix product $C = DD'$ is the pairwise co-occurrence matrix of the concepts. The memory required for the term document matrix is $O(nm)$ and $O(n^2)$ for the co-occurrence matrix. The matrix can also be processed in image batches, so the size of m is flexible depending on the memory constraints of the system. The co-occurrence matrix will dominate, making the memory constraint order $O(n^2)$. Using a sparse matrix representation is recommended because many concepts will never occur together.

A different approach is needed to count the number of owners. There are two possibilities. I can construct a separate term document matrix D_i for each owner. A per owner pairwise co-occurrence matrix would be calculated and thresholded to create a binary membership matrix, $C_i = (D_i D_i' > 0)$. These C_i would then be summed for all owners. This approach has similar memory complexity to the image counting method, but also requires organizing the dataset so that all images belonging to one owner can easily be retrieved. Another approach is to keep a list of unique owners for each pair of concepts. This is $O(pn^2)$ where p is the number of owners. Again, it will be smaller in practice because not every owner will contribute to every pair. The length of each list corresponds to the entry in the co-occurrence matrix for that pair. This implementation uses the latter method.

3.5 Collecting training edges from existing knowledge bases

I use the manually annotated commonsense knowledge bases ConceptNet [35] and Freebase [20]. Freebase focuses on concepts present in Wikipedia, leading to a greater coverage of proper nouns, while ConceptNet additionally collects facts from Verbosity, a game which collects commonsense knowledge from human players [37]. ConceptNet therefore has a greater coverage of more generalized commonsense knowledge. However, ConceptNet's facts can also be more subjective.

For each knowledge base, I download all edges related to concepts in the vocabulary. For ConceptNet, I use the API's text search to locate edges with the vocabulary concepts as either the source or target. I retrieve up to 10,000 edges for each vocabulary term. For Freebase, I use the API's search function to identify

Table 3.1: Description of relationship types including which knowledge bases they are drawn from, an example, and a brief definition. Relationships are listed alphabetically.

Relationship	Freebase	ConceptNet	Example and Definition
AtLocation		✓	"cat at location sofa" The source object is at the target scene or near the target object. Usually, the more portable object is the source.
AtLocationGeographic		✓	"sheep at location Scotland" Distinct from AtLocation in that the target must be a geographic location.
CapableOf		✓	"baby capable of sleep" The source is capable of performing the target
CreatedBy		✓	"bread created by baker" The source is created by the target
Causes		✓	"lightning causes thunder" The source causes the target
CausesDesire		✓	"food causes desire eat" The source causes a desire for the target
GeographicAdjective	✓		"Canada adjectival form is Canadian" The target is the adjectival form of the source.
GeographicContainment	✓	✓	"France contains Paris" A larger geographic entity contains a smaller geographic entity
HasA		✓	"baby has a toy" The target belongs to or is being used by the source. Sometimes the source and target are separate objects as opposed to PartOf.
HasProperty		✓	"cat has property white" The source can be described by the target.
IsA	✓	✓	"sofa is a chair" This is a classic hypernym relationship. The more specific concept is the source. The more general concept is the target.
LocatedNear		✓	"chair located near table" An object is located near another object. This relationship often overlaps with AtLocation, but is much less frequent in ConceptNet.
LocationOfAction		✓	"wash at location bathroom"

Continued on next page

Table 3.1 – *Continued from previous page*

Relationship	Freebase	ConceptNet	Example and Definition
			The source action takes place at the target location. This relationship often overlaps with AtLocation, but is much less frequent in ConceptNet.
MadeOf		✓	"house made of wood" Target describes the material of the source.
MotivatedByGoal		✓	"cook motivated by goal eat" The source action is motivated by the target event.
PartOf		✓	"fabric part of sofa", "baby part of family" The source is part of the target, usually, a literal piece although occasionally more conceptual such as in the second example.
ReceivesAction		✓	"cat receives action feed" The source is the recipient of the target action.
SimilarSize		✓	"dog similar size cat" Both concepts have similar sizes.
UsedFor		✓	"sofa used for relax" Target describes the use cases of the source.

the top 5 Freebase topic pages related to each vocabulary term. I then download the topic pages using the topic function. As mentioned before, each Freebase topic represents a separate word sense. I choose the first search result with an exact spelling match between the concept and one of the aliases as the word sense for that concept. For example, the topic page for the country "Turkey" is higher in the search results than the topic page for the bird "turkey". Therefore, I use the edges associated with the country rather than the bird. I limit the word senses to try to exclude unusual uses of the concept from the retrieved edges.

I take this large collection of edges and select edges where both the source and target are concepts in the vocabulary. Unfortunately, the edges collected from these resources are often quite noisy, so several layers of filters and manual intervention are used to select as clean a set of edges as possible. First, I choose a list of 19 relationships that I am interested in learning (See Table 3.1). For greater generality, I merge several Freebase relationships to make a GeographicContainment relationship, which includes edges like "France contains Paris", and a GeographicAdjective relationship which includes edges like "Canadian is the adjectival form of Canada". I also merge several Freebase scientific classification relationships and the ConceptNet IsA relationship, which includes edges like "cat is a feline". The complete list of component relationships for IsA, GeographicContainment, and GeographicAdjective is available in Appendix A.

For each of the relationships in the final list, I consider whether there should be a part of speech restriction

Table 3.2: All filters applied to relationship types. A checkmark indicates an inclusive filter, i.e., the concept must include that label. An ‘X’ indicates an exclusive filter, i.e the concept cannot have that label. Labels without either symbol are not explicitly included or excluded. Adj is short for adjective.

	Source					Target				
Relationship	Noun	Verb	Adj	Location	Color	Noun	Verb	Adj	Location	Color
AtLocation	✓			X		✓			X	
IsA	✓					✓				
UsedFor	✓							X		
GeographicContainment				✓					✓	
HasProperty	✓				X			✓		
HasA	✓					✓				
PartOf	✓			X		✓			X	
AtLocationGeographic	✓								✓	
CapableOf	✓						✓			
MadeOf	✓					✓				
ReceivesAction	✓						✓			
Causes										
GeographicAdjective				✓						
CreatedBy	✓					✓				
CausesDesire										
MotivatedByGoal										
LocatedNear	✓			X		✓			X	
LocationOfAction	✓			X			✓			
SimilarSize	✓					✓				

for either the source or the target. For example, the relationship “PartOf” can only involve nouns, so I add a noun restriction on both the concepts in the “PartOf” relationship. I also implement a few restrictions involving geographic locations, due to their frequency in the dataset, and colors, due to the visual importance of color. I also use filters to move some of the “PartOf” edges to the “GeographicContainment” relationship and to create a new relationship “AtLocationGeographic” containing edges like “cat in Paris”. Table 3.2 details the restrictions used for each relationship.

Some relationships, like “HasA”, “AtLocation” and “PartOf”, have a significant amount of overlap. To make these relationships more distinct and easier to learn, I remove the overlapping edges from the more general relationship. For example, “AtLocation” is more general than “PartOf”, so I remove the overlap between “AtLocation” and “PartOf” from “AtLocation”.

I want the training edges to be very relevant to the dataset, so I require the source and target concepts to meet a minimum co-occurrence frequency threshold. The retrieved edges with filtered relationships which meet this threshold are the *training set*.

3.6 Relationship transfer

3.6.1 Edge features

To use a classifier, I need a feature representation for the edges. Recently, there has been a focus on developing word vector representations that are good at representing analogies [24, 23, 26]. For the analogy task, the difference between the first two word vectors in the analogy should be similar to the second two word vectors in the analogy. For example, adding the difference between "cat" and "kitten" to the vector for "horse" should be very similar to the vector for "foal", as per the analogy "cat is to kitten as horse is to foal". This family of representations preserves semantic relationships in the vector space representation. Because I am interested in classifying the type of semantic relationship between two words, this kind of representation seems ideally suited to the task.

GloVe is the state-of-the-art word vector representation on the analogy task[26]. Like other representations in this family, GloVe trains a model that predicts the likelihood of terms appearing in similar contexts in text. GloVe, in particular, benefits from being a batch rather than an online method, unlike its closest competitor, skip-gram[23].

Pre-trained GloVe models are available. I selected a 300 dimensional model, trained on 42 billion tokens from Common Crawl, as it provided the most coverage of the vocabulary of any of the available models. It was also the most successful model tested by Pennington et al[26]. Some of the vocabulary terms are not covered by this model. For phrases, if all the words in the phrase are in the model, I can average the word vector representations to approximate the representation of the phrase.

I can only classify edges for which I have GloVe representations for both the source and target. Therefore, I refer to the portion of the expanded vocabulary for which I have GloVe representations as the *effective vocabulary*. Using GloVe, I experimented with two edge representations, (1) the concatenated GloVe vectors of the edge's source and target (2) the difference between the source and target vectors.

3.6.2 Classifiers

Most of the high frequency concept pairs are not present in the training set. However, their frequency in the dataset makes it likely that they share some unknown relationship. Therefore, any concept pair that meets the same frequency and filter criteria as the training set, but is not represented by an edge in the training set is a candidate pair to test for a relationship. Each candidate pair generates two *test edges*, one for each ordering of the concepts, which will be labeled by the multi-class classifier.

Using the features described in the previous section, 3.6.1, I train one-vs-all support vector machine

(SVM) classifiers for all relationships with at least 50 training edges. If an edge has multiple ground truth relationship labels, it appears in the positive training set for all relationships for which it has a label. I compare one-vs-all SVMs with Gaussian and cubic kernels. For the Gaussian SVM, I found a box constraint of 100 and a kernel scale of 27 to produce the best results.

All the filters that are applied to the training set, see Table 3.2, are also applied to the classifier output. This leaves use with a classifier score for each relationship type for each test edge. The classifier confidence is a strong signal for whether or not a directed edge exists (See Section 5.4). I consider all the classifier scores for a concept pair, i.e. all classifiers for both edge directions, and choose the maximum scoring classifier greater than zero as the *proposed edge* for that concept pair.

Chapter 4

Experimental set up

4.1 Datasets

I tested the method on two independent datasets, domestic animals and rooms in houses. I will refer to these datasets as the *animal* and *room* datasets. Choosing these sets allowed me to compare a vocabulary focused on objects to one focused on scenes. For the animal dataset, I used 11 terms collected from the '/biology/domesticated_animal' Freebase category. For the room dataset, I manually compiled a list of 7 common household rooms. In total, I collect 1,187,943 unique owners for the animal dataset and 1,109,921 unique owners for the rooms dataset. Figure 4.1 shows the number of images and the number of unique owners collected for each of the search terms.

4.2 Vocabulary

I choose a vocabulary frequency threshold of 500 unique owners per concept for the animal dataset and 200 unique owners per concept for the room dataset. These thresholds result in a 3565 concept effective vocabulary for the animal dataset and a 5897 concept effective vocabulary for the room dataset. Table 4.2 gives a more detailed break down of vocabulary statistics at each step of the collection process. Appendix B also provides the top 50 most frequent vocabulary words at each stage of collection process.

The number of non-English concepts labeled is quite small, approximately 3%. This is unsurprising as most of Flickr's traffic originates in the United States of America [2]. However, the non-English labeling process is also very conservative, labeling only concepts which meet three criteria: (1) The concept does not appear in WordNet's English dictionary, (2) the concept does appear in a WordNet language dictionary, and (3) the concept is not a proper noun. This filter fails when WordNet lacks a dictionary for the origin language for a concept. One example of a missing dictionary is German. In the animal dataset, "gans" (goose) and "ziege" (goat) are German concepts that are missed by the non-English filter. Another failure case occurs when the concept is a common noun in a non-English language and an unusual proper noun in

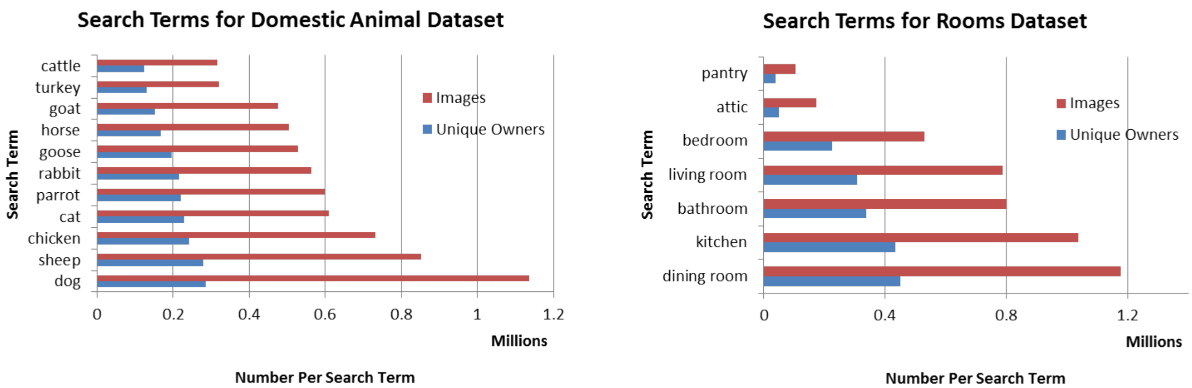


Figure 4.1: Distribution of images collected for each search term

Table 4.1: Dataset statistics. Statistics for tags per image are given after stopword filtering but before any other filters are applied.

	Domestic Animals	Rooms
Total number of images	6,536,760	4,481,772
Number of unique owners	1,187,943	1,109,921
Mean number of tags per image	6.99	5.71
Median number of tags per image	5	3

English, for example Azul, Buenos Aires. “Azul” means blue in Spanish, and is more likely being used to refer to the color than the location. Hopefully, the few concepts which escape labeling should not effect the results too greatly because they will be less frequent than their English equivalents.

Tagging an image with the location where it was photographed is very popular. Around 15% of the concepts are labeled as locations. The labeling process for locations is much more permissive, but it relies on Freebase’s completeness in cases where the concept has multiple word senses. If Freebase has no listed population for a location, the concept may not be included in the location filter. For example, “Constantinople” has no listed population and therefore, is not labeled as a location. A list of the most 200 frequent labeled locations for the room dataset is available in Appendix C.2.

Colors occur very frequently as image tags. However, the colors retrieved from Freebase frequently have multiple word senses. Some examples include “chestnut”, “terra cotta”, and “asparagus”. A complete list of labeled color concepts from both datasets is available in Appendix C.1. Therefore, it is important to use the color filter with caution. The method only applies the color filter to the source of the HasProperty relationship.

	Animal	Room
Minimum number of unique owners for vocabulary concepts	500	200
Size of initial vocabulary	3785	5577
Size filtered vocabulary	3648	5377
Size of expanded vocabulary	3691	6036
Size of effective vocabulary	3565	5897
Number of concepts present in ConceptNet	1726	4325
Number of concepts present in Freebase	2777	2968
Number of concepts present in either knowledgebase		
Number of concepts with part of speech	3142	5398
Number of non-English concepts	100	134
Number of location concepts	535	874
Number of color concepts	59	78

Table 4.2: Vobabulary statistics at each step of the collection process.

	Domestic Animals	Rooms	Combined
Minimum number of unique owners for a training or test edge	100	100	100
Minimum number of edges per relationship	50	50	50
Number of edges retrieved from ConceptNet	37,623	60,547	64,874
Number of edges retrieved from Freebase	3456	7151	8099
Number of edges with desired relationships	12,282	18,783	20,510
Number of training edges	3915	3390	5353
Number of test edges	169,538	101,158	232,082
Number of proposed edges	69,310	40,131	93,850

Table 4.3: Edge statistics. All edges described in this table have both source and target concepts in the vocabulary.

4.3 Ground truth edges

I retrieve 12,282 edges with one of the desired relationships for the animal dataset and 18,783 edges for the room dataset. Most of these edges come from ConceptNet. These edges are further reduced using the edge frequency threshold for a minimum of 100 unique owners. The remaining edges are the training edges. Figure 4.2 shows the frequency of the each relationship types in the datasets. A minimum of 50 edge examples per relationship is required to train a classifier. Causes, CausesDesire, CreatedBy, MotivatedByGoal, GeographicAdjective, LocatedNear, LocationOfAction, and SimilarSize, do not meet the threshold. Edges with these relationships are used as negative training examples.

Table 4.3 shows the break down of edge statistics at each stage of the collection process including the number of test edges for each dataset. The test edges are concept pairs which meet the same frequency threshold as the training edges but are not in the training set. There are two orders of magnitude more test edges than training edges.

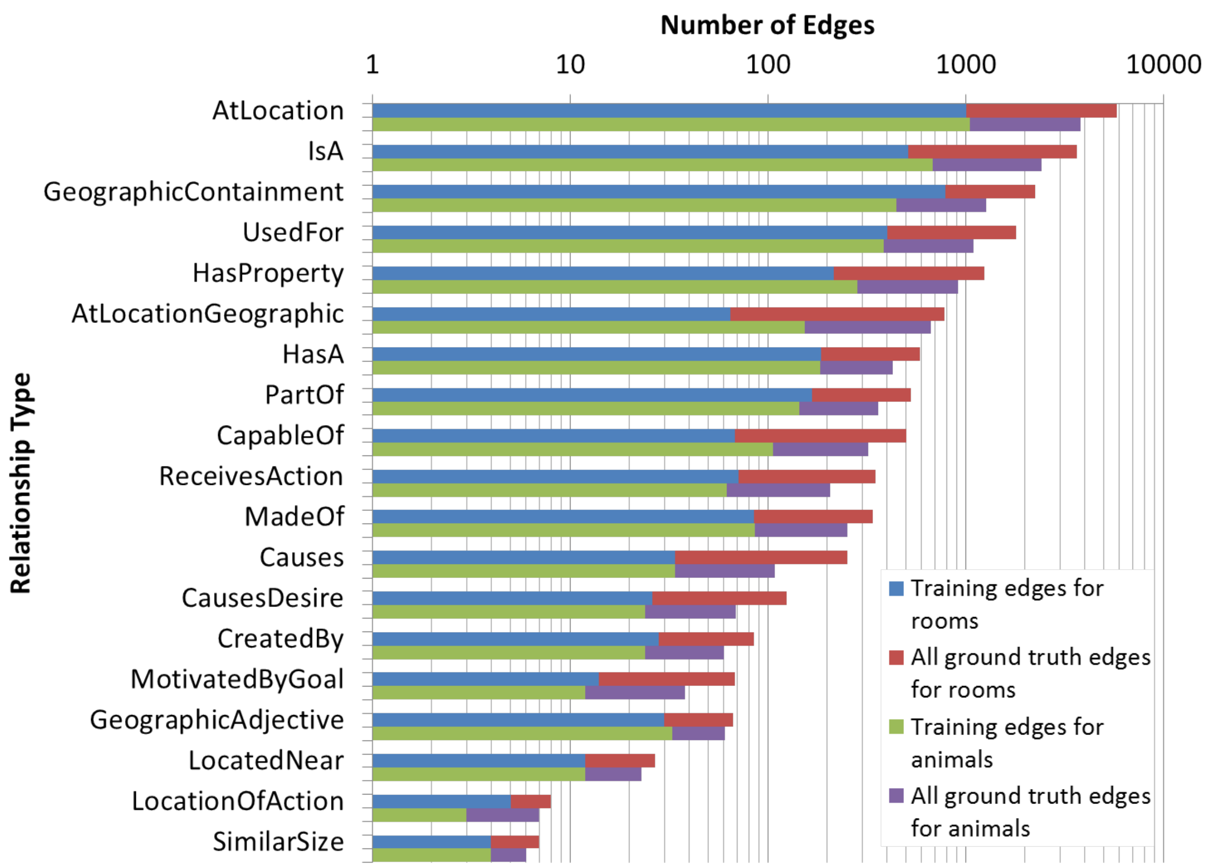


Figure 4.2: Distribution of edges over relationships

Chapter 5

Quantitative results

I perform quantitative evaluation in three ways, five-fold cross validation on the training set, inter-dataset training and testing, and hand-labeling candidate edges using a GUI interface.

5.1 Metric definitions

I use three metrics to express quantitative results: accuracy, mean precision, and mean recall. Accuracy is usually expressed as:

$$accuracy = \frac{TP + TN}{n}$$

where TP is the number of true positives, TN is the number of true negatives, n is the number of examples. In the cross-validation experiment, every edge receives some label, so disregarding relationship type, $TN = 0$. This is also equivalent to recall($recall = \frac{TP}{P}$) when $P = n$, where P is the number of examples with a positive label. The other two metrics, mean precision and mean recall, do not ignore the relationship type. Rather I average precision and recall over each classifier.

$$mean\ recall = \frac{1}{m} \sum_{r \in R} \frac{TP_r}{P_r}$$

$$mean\ precision = \frac{1}{m} \sum_{r \in R} \frac{TP_r}{GTP_r}$$

where P_r stands for the number of edges classified as relationship r from the set of all relationships R, TP_r stands for number of true positive, i.e. edges correctly classified as relationship r, GTP_r stands for number of ground truth positive, i.e. edges with the ground truth relationship r, and m is the number of relationships in R.

The datasets are unbalanced in terms of how many examples there are of each relationship. Using mean precision and mean recall gives a more accurate summary of how the individual classifiers are performing. I also report these values with different numbers of retrieved edges. It is possible for a pair of concepts to

Table 5.1: *Accuracy, Precision, and Recall for various classifiers* Columns include: Accuracy for any correct label using the highest scoring classifier, mean recall over relationships using the highest scoring classifier (MR@1), mean precision over relationships using the highest scoring classifier (MP@1), mean recall over relationships for the top three highest scoring classifiers (MR@3), mean precision over relationships for the top three highest scoring classifiers (MP@3). The highest values for each dataset are in bold.

		Animals				
Classifier	Feature Type	Accuracy	MR@1	MR@3	MP@1	MP@3
Gaussian SVM	Concatenated	**0.836**	0.624	0.829	**0.712**	0.288
	Difference	0.821	0.607	0.808	0.695	0.287
Cubic SVM	Concatenated	0.835	**0.626**	**0.839**	0.705	**0.289**
	Difference	0.828	0.601	0.811	0.700	0.286

		Rooms				
Classifier	Feature Type	Accuracy	MR@1	MR@3	MP@1	MP@3
Gaussian SVM	Concatenated	0.810	0.606	**0.826**	0.716	**0.282**
	Difference	0.810	0.603	0.808	0.721	0.281
Cubic SVM	Concatenated	0.815	**0.610**	0.819	0.719	0.280
	Difference	**0.817**	**0.610**	0.817	**0.735**	**0.282**

have multiple relationships. If I only report the highest scoring relationship (@1), I may be missing some correct edge labels. Additionally, for misclassified edges, the second most confident classifier, may have the correct classification. Therefore, I also report the mean precision and mean recall for a correct classification in the top three highest scoring classifiers (@3).

5.2 Cross-Validation

The first set of experiments use five-fold cross-validation on the training set. The training set has ground truth labels for each edge's relationship which I don't have for the test set. For five-fold cross-validation, I split the training set into five equally sized samples. I train on four of the samples and test on the fifth, repeating until I have tested on all the samples. I then average the results over all trials. This provides an estimate of how well the classifiers learn.

From the results in Table 5.1, I see that there is very little difference between the SVM classifiers. The Cubic SVM using difference features performs slightly better on the rooms dataset, so I perform the remainder of the experiments using this classifier configuration.

Figures 5.1 and 5.2 show more detailed confusion matrices from the Cubic SVM classifier for both datasets accompanied by a few examples of error cases in Tables 5.3 and 5.4. The confusion matrices show the number of times an example with a ground truth label y was labeled with the classification x. Both datasets have the greatest confusion (as a percentage of the ground truth relationship) on the same set of classifiers:

Table 5.2: *Accuracy, precision, and recall comparison for various training and test sets using the Cubic SVM and difference features* When training and testing on the same dataset, 5-fold cross-validation is used. Columns include: Accuracy for any correct label using the highest scoring classifier, mean recall over relationships using the highest scoring classifier (MR@1), mean precision over relationships using the highest scoring classifier (MP@1), mean recall over relationships for the top three highest scoring classifiers (MR@3), mean precision over relationships for the top three highest scoring classifiers (MP@3).

Training Set	Testing set	Accuracy	MR@1	MR@3	MP@1	MP@3
Animal	Animal	0.841	0.671	0.874	0.772	0.352
Room	Animal	0.758	0.529	0.780	0.612	0.318
Animal	Animal (hand-labeled)	0.662	0.368	0.710	0.405	0.213
Room	Room	0.821	0.651	0.873	0.752	0.332
Animal	Room	0.731	0.534	0.774	0.632	0.304
Room	Room (hand-labeled)	0.628	0.341	0.709	0.385	0.197

- HasA (ground truth) and AtLocation (classified) is confused an average of 17.5%
- PartOf and AtLocation is confused an average of 32.5%
- PartOf and IsA is confused an average of 15%
- ReceivesAction and UsedFor is confused an average of 17.5%
- CapableOf and UsedFor is confused an average of 15.5%

Some of the same edges are also confused in both datasets, such as "cook part of kitchen".

Several of these frequently confused relationships are somewhat ambiguously defined or contain some degree of overlap. Is a stove in a kitchen or part of a kitchen? Or both? Additionally, an edge can have multiple correct labels. A calf is a cow and a cow has a calf. The ground truth contains ambiguous relationships and is incomplete. This can also contribute to classifier confusion. For example, there are 83 AtLocation edges containing "kitchen" in the second position, compared to 6 with other relationships. Consequently, "cook part of kitchen" is classified with high confidence as "cook at location kitchen". In this case, the classifier has learned that "kitchen" is a scene with the predominating relationship AtLocation. A cook being part of a kitchen is more idiomatic than semantic. The classifier's alternative relationship proposal is completely reasonable.

The error cases fall roughly into two categories: (1) one of the concepts, in its current position, always belongs to the same relationship in the training set, or (2) the training set does not contain one or both of the concepts in their current positions and generalizes incorrectly.

For case (1), the classifier is essentially memorizing a list of concepts which always belong to the relationship in the training set. Usually, this is also specific to the position of the concept. For example in the rooms dataset, “table” is the target in 21 AtLocation edges, and only once as HasA. As a result, the classifier learns to classify this configuration as AtLocation with high confidence, and in the cross validation task, the single HasA edge, “restaurant has a table”, is classified as “restaurant at location table”. This behavior could be considered a localized overfitting caused by the limited training data.

For case (2), the classifier is being forced to generalize based on the GloVe vector features. In some cases, this leads to erroneous labels for closely related concepts. For example, “farmhouse” only occurs in the rooms dataset in one edge, “farm has a farmhouse”. However, the closely related term “farm”, occurs in seven AtLocation edges. The classifier generalizes the high confidence for “farm” to “farmhouse”, incorrectly classifying “farm has a farmhouse” as “farm at location farmhouse”. In this case, the generalization leads the classifier to make a mistake, but similar generalizations may be boosting classifier performance on other edges. For example, I have only one edge containing “aquarium”, “fish at location aquarium”, but the classifier identifies the correct label with high confidence, perhaps drawing training examples containing related concepts such as “zoo”. This illustrates how important quality training data is to the development of strong knowledge proposals.

5.3 Cross-dataset training and testing

Another way to test a classifier’s ability to generalize is training on one dataset and testing on another. The animal dataset focuses on objects while the room dataset focus on scenes. Because of their different training examples, I expect a drop in recall and precision, but if the classifiers generalize well, the drop should not be too great. Table 5.2 shows that there is an average drop of 13.5 points in mean recall and an average drop of 14 points in mean precision. These modest drops indicate the promise of this method.

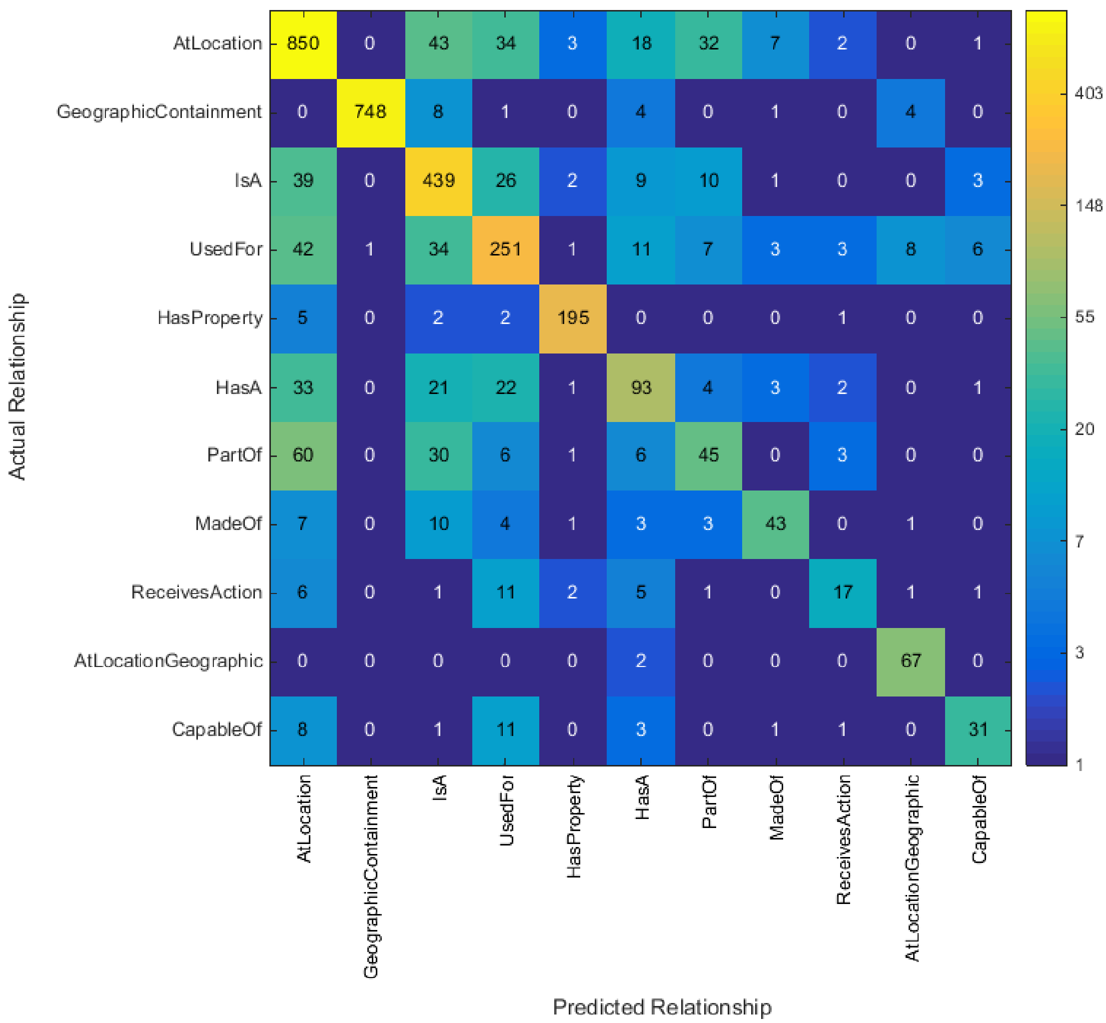


Figure 5.1: *Confusion matrix for room dataset* This confusion matrix is generated using 5-fold cross validation on the training set and the cubic svm classifiers trained with difference features. Each entry represents the number of edges with the ground truth label, y, that were classified as x. The relationships are ordered along the x and y-axis from most to least frequent, so rows on the top represent a greater number of edges than rows on the bottom. Examples of confused edges can be seen in Table 5.3

Table 5.3: Selected examples of high confidence confused edges for room dataset

Source	Target	Ground Truth Relationship	Highest Scoring Classifier Relationship	Score
restaurant	table	HasA	AtLocation	1.11
farm	farmhouse	HasA	AtLocation	0.76
cook	kitchen	PartOf	AtLocation	1.56
book	library	PartOf	AtLocation	1.07
apple	bake	ReceivesAction	UsedFor	1.56
animal	travel	CapableOf	UsedFor	0.66
baby	woman	CreatedBy	IsA	0.65

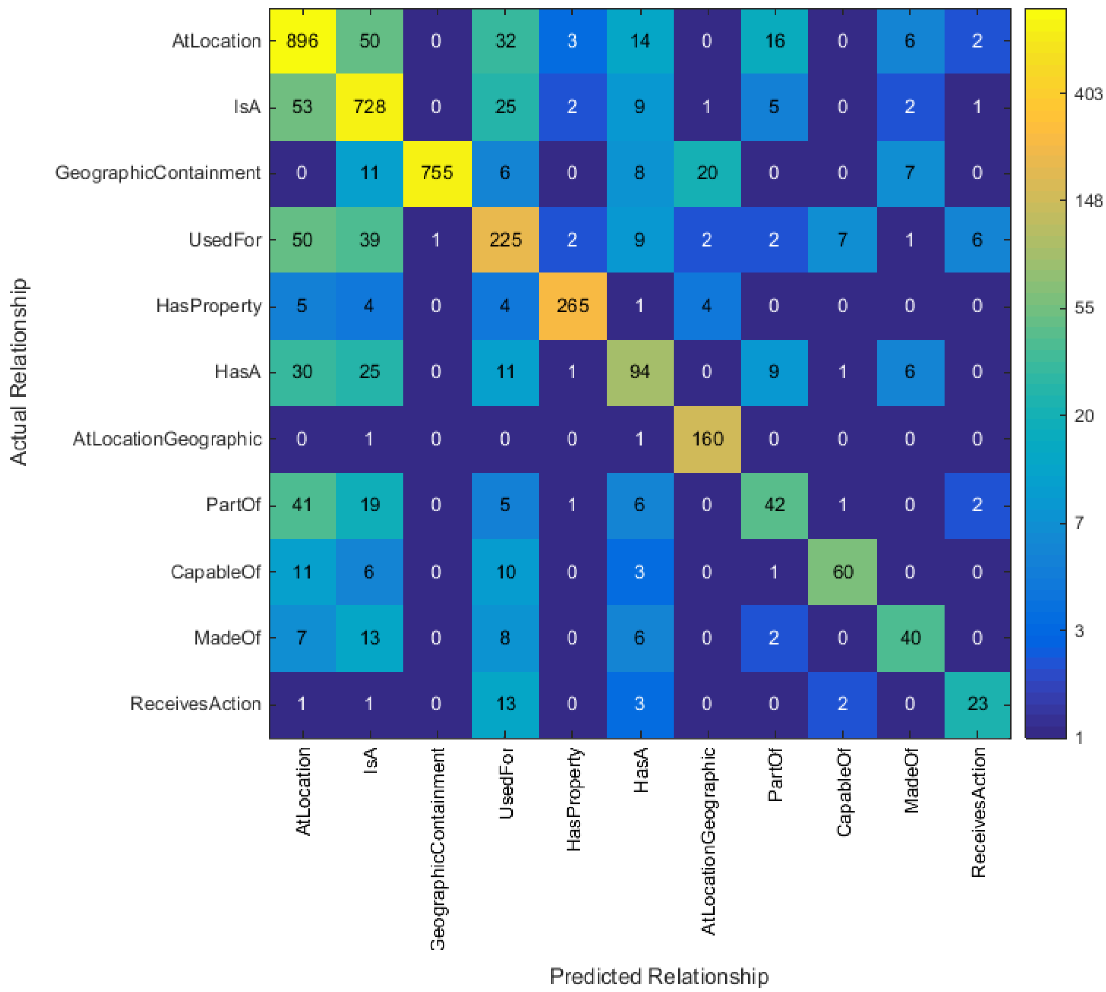


Figure 5.2: *Confusion matrix for animal dataset* This confusion matrix is generated using 5-fold cross validation on the training set and the cubic svm classifiers trained with difference features. Each entry represents the number of edges with the ground truth label, y, that were classified as x. The relationships are ordered along the x and y-axis from most to least frequent, so rows on the top represent a greater number of edges than rows on the bottom. Examples of confused edges can be seen in Table 5.4

Table 5.4: Selected examples of high confidence confused edges for animal dataset

Source	Target	Ground Truth Relationship	Highest Scoring Classifier Relationship	Score
ship	bridge	HasA	AtLocation	0.84
cook	kitchen	PartOf	AtLocation	1.18
wave	ocean	PartOf	AtLocation	0.93
pony	animal	PartOf	IsA	1.88
chicken	eat	ReceivesAction	UsedFor	0.78
animal	love	CapableOf	UsedFor	0.85
milk	mammal	CreatedBy	IsA	0.89

5.4 Hand-labeling

It is unclear whether the cross validation conclusions will hold for the test edges for several reasons. The sources from which the ground truth was collected are text-based, and the candidate edges are image-based. There are vocabulary terms which are not present in any ground truth edge. However, the test edges have one even more significant difference to the training edges; many of them should not be assigned a label. Some frequent concept pairs which are included in the test set have a strong correlation, but no direct relationship, for example "sheep" and "green". From the training edges alone, I have no way to identify such pairs, because all of the training edges have valid relationships.

To further analyze the test edges, I hand-labeled around 500 edges from each dataset using a GUI, shown in Figure 5.3. The GUI asked three questions: (1) is there a relationship between these two terms, (2) can the relationship between the two terms be illustrated with a photo such as one of those displayed below, and (3) which relationship label best describes the relationship between the two terms. Question (1) addresses edge existence. Question (2) addresses edge visualness, and Question (3) labels the edge with a relationship.

One concern when choosing which edges to label was that many edge properties are not uniformly distributed. For example, the frequency of edges has a long tail distribution as does the conditional probability of edges. The distributions for normalized pointwise mutual information and highest classifier score are closer to a skewed normal distributions. To attain a labeled sample with a large variety of edge properties, I perform a simplified version of stratified random sampling. For each property, I bin the property values in 50 bins. I then randomly sample from each bin. This provides better representation of outliers in the sample. Tabel 5.5 shows the number of labels that I collected by handlabeling labeling around 500 edges for each dataset.

The distribution of relationship types for the hand-labeled edges is somewhat similar to the relationships from the knowledge bases (See Figure 5.4). The room dataset hand-labeled relationship frequency appears to follow the same exponential trend as the knowledge base relationships, but the animal dataset follows a very different distribution. It shows spikes for the IsA and AtLocationGeographic relationships. The IsA spike probably results from the many different levels of specificity which people use to refer to animals, e.g. "bird", "chicken", and "hen". From the observations, this use of synonyms appears to be more frequent for animals than for objects found in the home. The spike in AtLocationGeographic relationships may result from people photographing animals while traveling. They are less likely to mention the geographic location of their home.

The Receiver Operator Characteristic for predicting directed edge existence show that the highest classifier score is the strongest signal for finding directed edges (See Figure 5.5). Using this signal, I propose

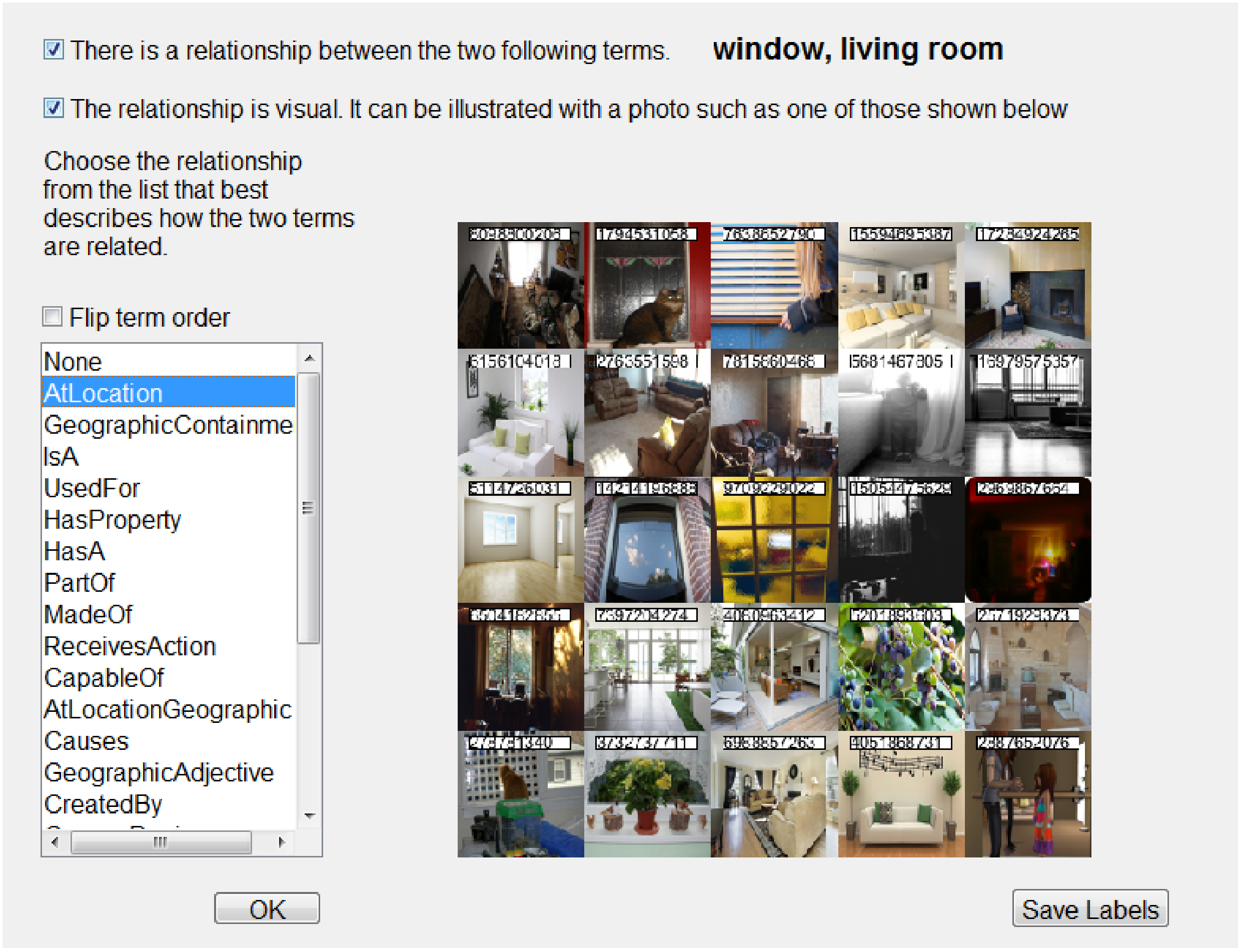


Figure 5.3: A screenshot of the GUI used for hand-labeling concept pairs

edge labels for test set edges which have the maximum classifier score over all classifiers for a concept pair. Additionally, the score must be greater than zero. For example, “cattle at location field” and “field made of cattle” are the two highest scoring classifications for the pair “field” and “cattle”. “cattle at location field” has the higher classifier score with 1.13 compared to 0.199, so the proposed label is “cattle at location field”. 69,310 edges have proposed labels in the animal dataset and 40,131 edge have proposed labels in the room dataset.

Using the proposed labels, I can further probe the accuracy of the classifiers by generating confusion matrices for the hand-labeled data. Figures 5.7 and 5.6 show these confusion matrices. Using the highest classifier score is doing a decent job of filtering out unlabeled edges. I correctly filter out 76% of the animal dataset and 67% of the room dataset while incorrectly excluding only 29% of the animal dataset and 26% of the rooms dataset. The small number of labeled examples makes it hard to draw conclusions about the individual relationships, but the AtLocation, IsA and HasProperty classifiers seem to be most powerful.

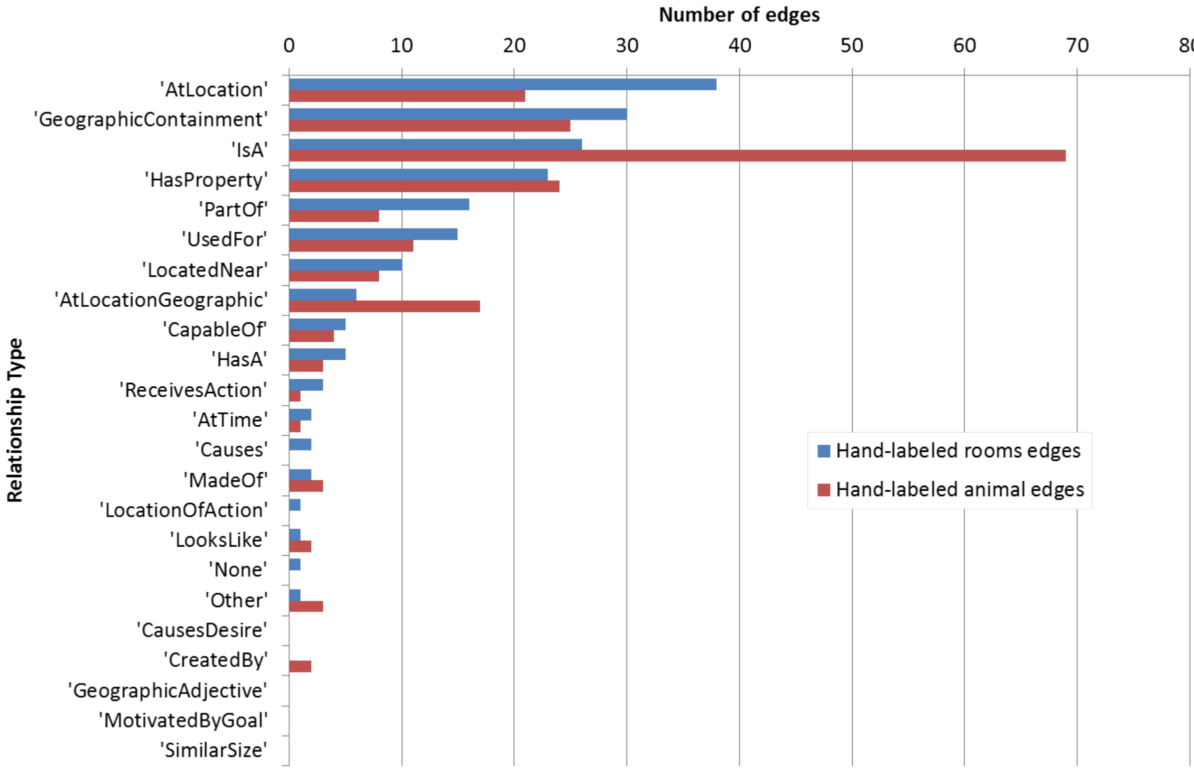
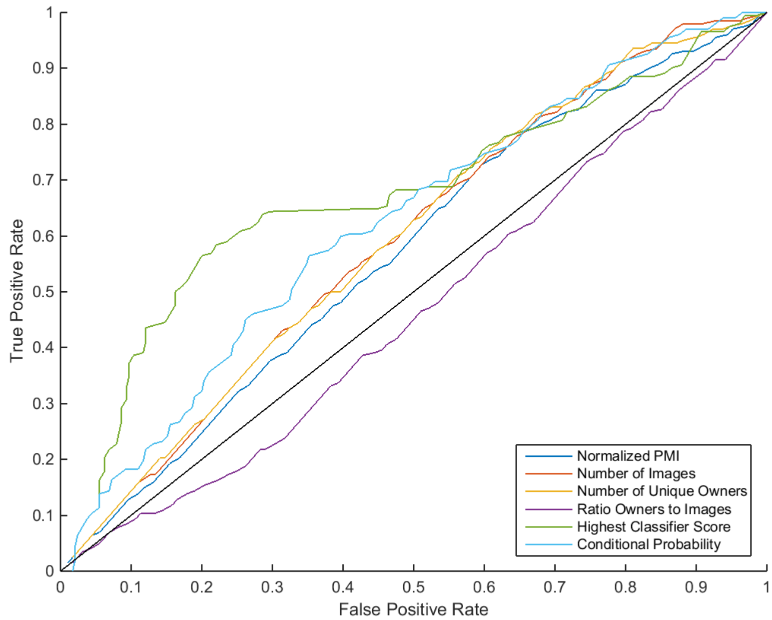


Figure 5.4: Relationship distribution of hand-labeled edges for each dataset. The animal dataset has a larger proportion of IsA edges because animals are often referred to with different levels of specificity. For example, "animal", "bird", "chicken", "hen" can all be used to refer to the same animal. The room dataset has a larger proportion of AtLocation relationships because it has a large number of inanimate objects found in the home.

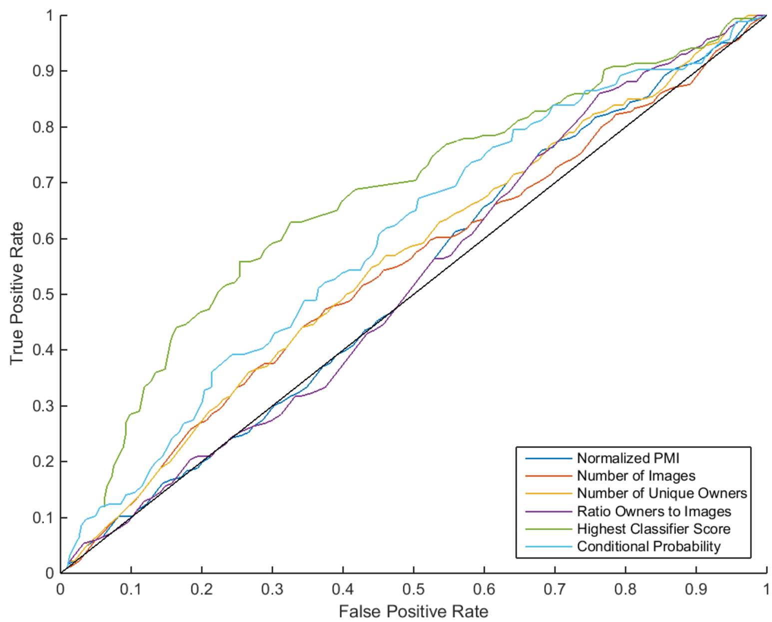
Table 5.5: Edge statistics for hand-labeled edges

	Domestic Animals	Rooms
Number of hand-labeled edges	492	490
Number of edges with a relationship	202	187
Number of visual edges	153	137

The last piece of hand-labeled information is visualness. The classifiers do not explicitly learn visualness because they have no access to a visual representation of the images, only the concept features. However, the labeling shows that the visualness of an edge is linked to its relationship type. In Table 5.6, I see that the AtLocation relationship is always visual while the GeographicContainment relationship is never visual. HasProperty and IsA are also strongly visual. The potential of visual representations is an exciting direction for future work and is discussed more in Chapter 7.



(a) Animal Dataset



(b) Room Dataset

Figure 5.5: Receiver Operator Characteristic (ROC) for predicting directed edge existence. I use one of the metrics listed in the legend to predict whether or not an ordered concept pair is an edge. Pairs with a value above a certain threshold are labeled as edges. These labels are compared to the hand-labeled ground truth, counting the number of true positive and false positive labels. The curve is plotted by varying the threshold which varies the numbers of true and false positives. The area below the curve is the accuracy. The diagonal is random chance. A larger area above the diagonal indicates better prediction. In these plots, the highest classifier score is most predictive of directed edge existence. Additional plots of the corresponding thresholds and precision recall curves are available in Appendix D

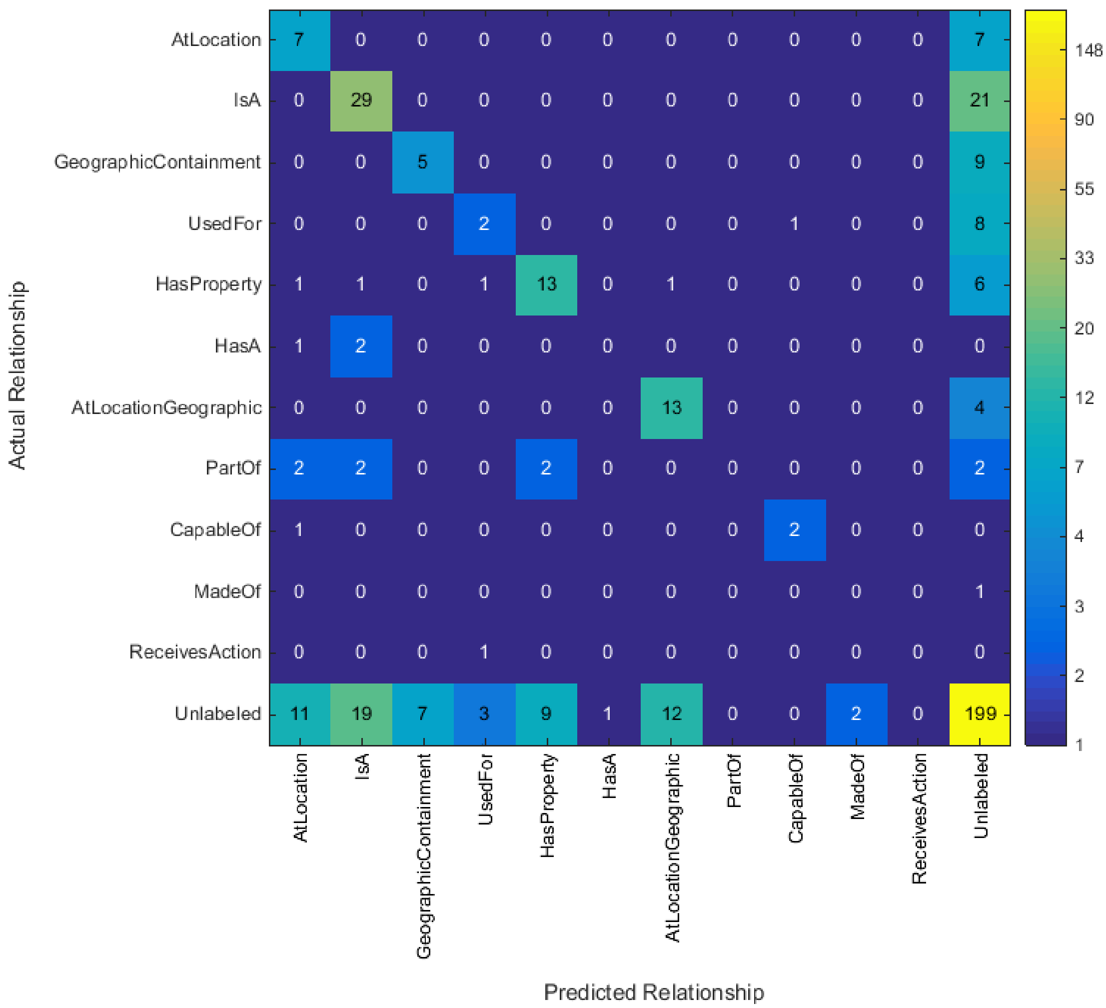


Figure 5.6: *Confusion matrix using hand-labeled relationships from animal dataset* This confusion matrix is generated using the hand-labeled relationships for the test edges. Each entry represents the number of edges with the ground truth label, y, that were classified as x. The relationships are ordered along the x and y-axis from most to least frequent, so rows on the top represent a greater number of edges than rows on the bottom.

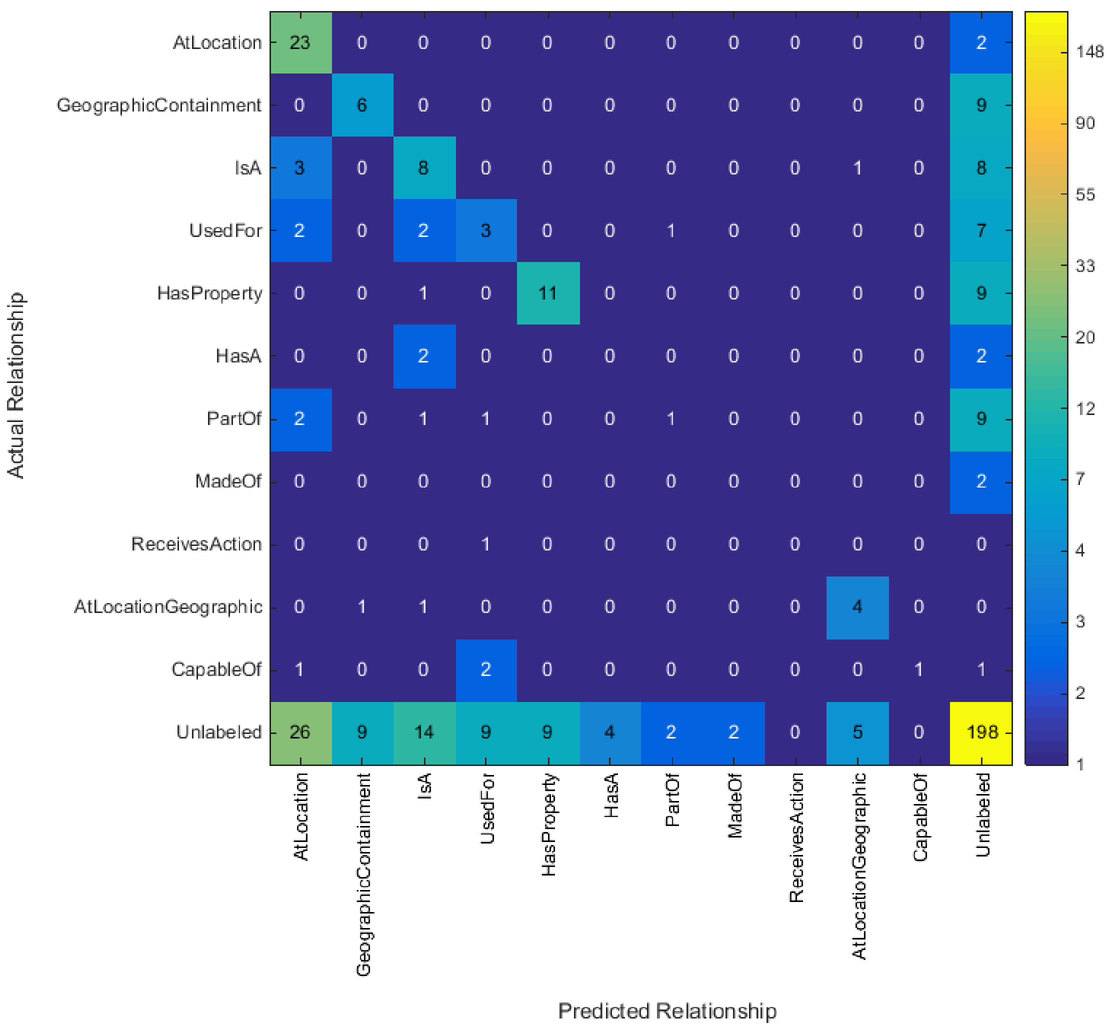


Figure 5.7: *Confusion matrix using hand-labeled relationships from room dataset* This confusion matrix is generated using the hand-labeled relationships for the test edges. Each entry represents the number of edges with the ground truth label, y, that were classified as x. The relationships are ordered along the x and y-axis from most to least frequent, so rows on the top represent a greater number of edges than rows on the bottom.

Table 5.6: Visualness of relationship types. This table reports the percent of handlabeled-edges that are labeled visual for each relationship type as well as the total number of hand-labeled edges summed over both datasets. The relationships are sorted by percent visual then by the number of examples.

Relationship	Percent Visual	Total Edges
AtLocation	1.00	59
LocatedNear	1.00	18
MadeOf	1.00	5
ReceivesAction	1.00	4
LooksLike	1.00	3
LocationOfAction	1.00	1
IsA	0.97	95
HasProperty	0.96	47
UsedFor	0.92	26
PartOf	0.92	24
CapableOf	0.89	9
HasA	0.75	8
Causes	0.50	2
AtLocationGeographic	0.09	23
GeographicContainment	0.00	55
Other	0.00	4
AtTime	0.00	3
CreatedBy	0.00	2
None	0.00	1

Chapter 6

Qualitative results

This chapter will discuss the proposed edges in detail. I retrieved 69,310 proposals for the animal dataset and 40,131 proposals for the room dataset. The quantitative analysis of hand-labeled edges in the previous chapter suggests an average accuracy of 64% for the proposed edges. I will use examples to discuss the kinds of edges that I successfully detect and the mistakes that are made.

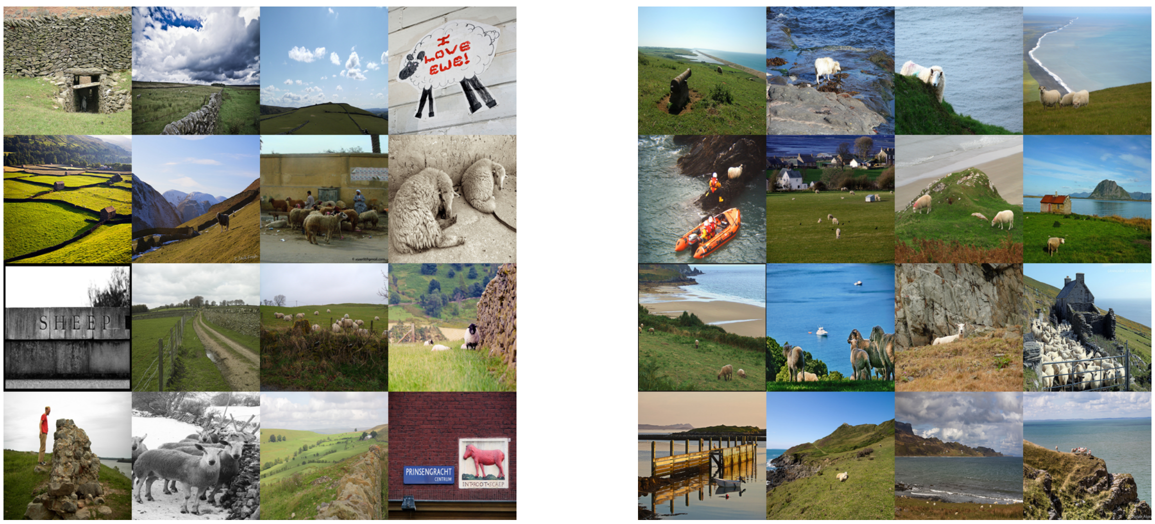
The first set of examples is drawn from the Animals dataset. In this dataset, I retrieved images with keywords from a list of domestic animals including “sheep”. Table 6.1 shows the training edges that contained the term “sheep”. This list of eleven edges is accurate “sheep” commonsense knowledge, but it is not comprehensive. I have 664 additional proposed edges for sheep. Of these, 31 are high confidence classifications. *High confidence edges* have a large score, a large number of owners, and a non-negative PMI. For the animal dataset, I use a threshold of 500 unique owners, and a score greater than 0.6 for high confidence edges. These thresholds are chosen by manual examination of the proposed edges.

Table 6.2 shows the top eight highest confidence AtLocation edges. Some of these are scenes like “sheep at location countryside” or “sheep at location field”. Others are other objects that frequently occur in images near sheep, such as, “sheep at location wall”, see Figure 6.1a or “sheep at location house”. Not all proposed edges are intuitive, for example, “sheep at location sea”, but examining the images supports the edge with a set of images where sheep graze near the ocean or are even being rescued from a rocky seashore(see Figure 6.1b). The knowledge I learn about sheep is not limited to locations. Table 6.3 show high confidence “sheep” edges for five other relationships types including parts, properties, categories, actions, and uses.

Despite the high confidence, not all edges are well supported. Figure 6.2a shows that alternate word senses may be subsumed into another frequent word sense. Most of these images show close ups of sheep ears appropriate for the fact “sheep has a ear”, but one image shows sheep shaped earrings. In this case, an appropriate relationship was proposed. However, differentiating word senses is still an open problem. Figure 6.2b shows another unsolved problem. The classifier correctly learns that green is usually a property, but incorrectly associates green with the sheep rather than the field in which the sheep is standing.

Table 6.1: All training edges containing “sheep” in animal dataset. Rows are sorted by relationship type, then by number of owners.

Source	Target	Ground Truth Relationship	Number of Owners	PMI
sheep	farm	AtLocation	11549	0.22
wool	sheep	AtLocation	5883	0.39
sheep	meadow	AtLocation	1603	0.21
sheep	fair	AtLocation	800	0.07
sheep	graze	CapableOf	2748	0.19
sheep	wool	HasA	5883	0.39
sheep	animal	IsA	15153	0.03
sheep	farm animal	IsA	2715	0.22
sheep	mammal	IsA	958	0.07
merino	sheep	IsA	348	0.21
sheep	person	IsA	135	-0.02



(a) sheep at location wall

(b) sheep at location sea

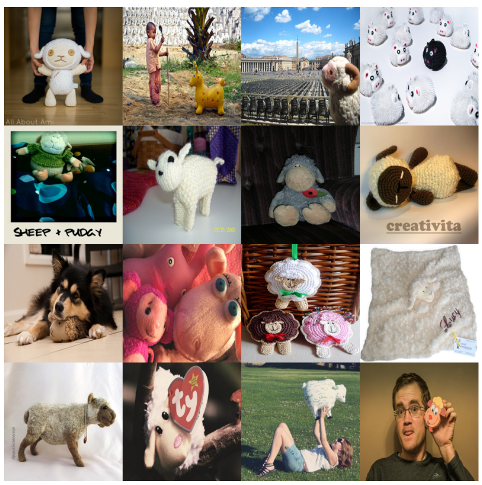
Figure 6.1: Selected images from high confidence AtLocation edges containing “sheep” from animal test set. These image grids illustrate the highlighted edges in Table 6.2. These two examples show the AtLocation relationship can be used to describe nearness as well as scene locations.

Table 6.2: Examples of high confidence AtLocation edges containing “sheep” from the animal test set. Bold rows are illustrated in Figure 6.1. Edges are the eight highest scoring proposals with more that 500 owners and NPMI greater than 0.02

Source	Target	Proposed Relationship	Score	Number of Owners	NPMI
sheep	**wall**	**AtLocation**	**2.02**	**1391**	**0.10**
sheep	fence	AtLocation	1.73	2786	0.15
sheep	countryside	AtLocation	1.70	4321	0.30
sheep	church	AtLocation	1.31	1015	0.09
sheep	**sea**	**AtLocation**	**1.25**	**2414**	**0.03**
sheep	house	AtLocation	1.17	1382	0.03
sheep	village	AtLocation	1.15	1096	0.09
sheep	field	AtLocation	1.14	8287	0.27

(a) sheep has a ear

(b) sheep has property green



(c) sheep is a toy

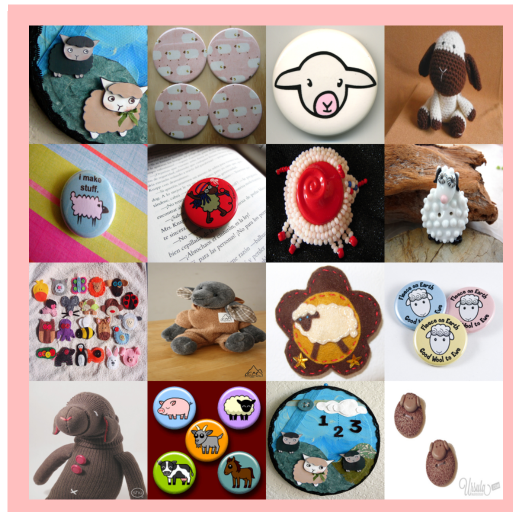
(d) sheep receives action feed

Figure 6.2: Selected images from high confidence edges containing “sheep” from animal test set. These image grids illustrate the highlighted edges in Table 6.3. These examples show the variety of images contributing to each edge. In Figure 6.1b, the images are very consistent, but the association learned is incorrect.

Table 6.3: Examples of high confidence edges containing “sheep” from the animal test set. The edges are the highest scoring proposals with more than 500 owners and PMI greater than zero for selected relationship. Bold rows are illustrated in Figure 6.4. Pink rows are misclassified.

Source	Target	Proposed Relationship	Score	Number of Owners	NPMI
sheep	**ear**	**HasA**	**1.07**	**583**	**0.01**
sheep	**green**	**HasProperty**	**0.93**	**6450**	**0.11**
sheep	young	HasProperty	0.75	508	0.03
sheep	**toy**	**IsA**	**0.92**	**1650**	**0.02**
sheep	livestock	IsA	0.69	1917	0.22
sheep	ram	IsA	0.62	3140	0.32
sheep	**feed**	**ReceivesAction**	**0.79**	**897**	**0.05**
sheep	train	ReceivesAction	0.69	510	0.01
land	sheep	UsedFor	0.82	551	0.11
farmland	sheep	UsedFor	0.74	665	0.21
wood	sheep	UsedFor	0.74	1188	0.02
barn	sheep	UsedFor	0.74	1739	0.14

(a) sheep at location fine art



(b) sheep has a button

(c) sheep made of plastic

Figure 6.3: Selected images from low confidence edges containing “sheep” from animal test set. These image grids illustrate the highlighted edges in Table 6.4. Twos of these examples show unusual but accurate edges. The third, Figure 6.3b, shows some toy sheep with buttons, but is predominantly sheep images on buttons or sheep shaped buttons.

Table 6.4: Low confidence edge proposals containing "sheep" from animal dataset. These edges have a score less than 0.6, fewer than 500 owners, and a positive NPMI. Low confidence edges have many more erroneous labels than high confidence edges. Bold rows are illustrated in Figure 6.3. Pink rows are misclassified. Ellipsis indicates hidden examples.

Source	Target	Proposed Relationship	Score	Number of Owners	NPMI
sheep	west	AtLocation	0.41	375	0.02
sheep	wind	AtLocation	0.40	349	0.07
sheep	horizon	AtLocation	0.40	293	0.10
sheep	lane	AtLocation	0.37	164	0.09
sheep	**fine art**	**AtLocation**	**0.36**	**104**	**0.00**
...					
sheep	vista	AtLocation	0.02	193	0.10
sheep	figure	AtLocation	0.01	181	0.01
sheep	**button**	**HasA**	**0.40**	**138**	**0.04**
sheep	horizontal	HasA	0.33	129	0.05
sheep	flare	HasA	0.21	111	0.04
sheep	embroidery	HasA	0.19	125	0.01
sheep	crochet	HasA	0.16	313	0.04
sheep	cloudy	HasProperty	0.57	463	0.11
sheep	calm	HasProperty	0.44	150	0.04
sheep	friendly	HasProperty	0.40	108	0.00
sheep	curious	HasProperty	0.34	287	0.03
sheep	peaceful	HasProperty	0.19	251	0.08
sheep	overcast	HasProperty	0.14	104	0.06
sheep	lonely	HasProperty	0.05	174	0.04
sheep	rare	HasProperty	0.01	102	0.02
sheep	baby animals	IsA	0.53	337	0.07
sheep	track	IsA	0.52	377	0.05
sheep	analogue	IsA	0.51	167	0.05
sheep	tag	IsA	0.47	346	0.01
sheep	angel	IsA	0.44	214	0.02
...					
sheep	outback	IsA	0.01	130	0.07
sheep	national trust	IsA	0.00	477	0.16
sheep	top	IsA	0.00	154	0.01
sheep	**plastic**	**MadeOf**	**0.28**	**170**	**0.00**
sheep	drive	ReceivesAction	0.33	263	0.02

If I consider low confidence edges, there are many more erroneous classifications. Table 6.4 shows some low confidence edge examples. While a larger proportion of these are incorrectly labeled, there are still a few unusual but accurate edges in the list. Figure 6.3 shows the supporting images for some of these less intuitive edges.

"Sheep" was one of the search terms that I used to retrieve the Flickr images and the 4th most frequent concept in the animal dataset with 78,748 unique owners. However, most of the vocabulary were not initial search terms and are not as frequent. It is important to show that the method can also learn from less frequent concepts.

"Bird" is the 16th most frequent vocabulary term with roughly half the number of unique owners compared to "sheep". Interestingly, I have 89 training edges containing the term "bird", many more than "sheep" (See Appendix Table E.1 for the full list). This may be because "bird" is a broader animal category than "sheep". The search keywords for the animal dataset include three kinds of bird: turkey, goose, and parrot.

Table E.2 shows high confidence proposed edges for "bird". The classifiers' proposals appear to learn from the different species of birds in the dataset. The locations learned, such as lake, river, ocean, pond, and wetland, may come from the "goose" images. While the location "jungle" and the many colorful properties may come from the "parrot" images. For example, Figure 6.5b shows images of brightly colored parrots from the set of images with the relationship "bird has property bright". Other proposals represent all species of bird, such as "bird has a foot". In Figure 6.4a, I see both parrot and goose feet illustrating "bird has a foot".

Let's consider a much less frequent concept, "farmland". "Farmland" occurs in 1,531 images. I have 6 training edges for "farmland". For less frequent concepts, I receive many fewer proposals because they co-occur with fewer other concepts. For "farmland", there are 16 proposals. Interestingly, I learn some facts which are very different from the training examples. In the training set, I have "cow at location farmland" and "horse at location farmland". In the proposed edges, I have "bull at location farmland", which is a similar relationship, but I also have "farmland used for sheep" and "farmland used for cattle". This shows the ability of the classifiers to generalize between concepts. The complete list of training examples and a selection of high confidence edges for "farmland" is available in Appendix E.1.

The number of training edges also effects the confidence and accuracy of the proposed edges. The concept "gosling" is not present in any training edge. In Table 6.5, there are many more misclassified high confidence proposals than in the previous examples. For example, the classifiers prefer "swim at location gosling" with a score of 0.87 to "gosling capable of swim" with a score of -0.22. However, the classifiers do correctly identify that "gosling" is an animal, bird, duck, and baby, and that they are frequently found in lakes.

Similar results from frequent to infrequent concepts using the terms "sofa", "mirror", "garlic", and "new

(a) bird has a foot

(b) bird has property bright

Figure 6.4: Selected images from high confidence edges containing “bird” from animal test set. These image grids illustrate the highlighted edges in Table E.2. These two examples show that the “bird” facts are being learned that apply to all species of birds, e.g. “bird has a foot”, and to specific species of birds, e.g. “bird has property bright”.

year” from the room dataset are detailed in Appendix E.2. The two datasets have different vocabularies and different training sets, but when they do overlap it can provide some insight into the different kinds of knowledge that they are learning. A good example of that overlap is the concept “baby”. “Baby” is very frequent in the animal dataset and in the room dataset. In the animal dataset, it occurs in images picturing a wide variety of species, while in the room dataset it usually refers to human babies.

Two edges are learned by the classifiers for both datasets: “mother has a baby” (Figures 6.5a and 6.5b) and “baby has property white” (Figures 6.5c and 6.5d). The classifiers arrive at the same knowledge from very different starting points, but there are slight differences. “Mother has a baby” is a universal fact that applies to all animals including humans, so it can be learned equally well from both datasets. However, “baby has property white” in the animal dataset refers to white fur or feathers, while “baby has property white” in the room dataset refers to either white walls or clothing or a light flesh tone. The animal dataset edge is more consistent with the images. For “baby has property white”, the datasets learn two slightly different relationships with the same text gloss. This is an area where visual models of the relationships could clarify meaning.

For the proposed edges that are specific to one dataset, the differences between non-human animal and human babies emerge. Figure 6.6a shows the animal dataset edge “baby has property fuzzy”. Figure 6.6b shows the room dataset edge “baby at location kitchen”. Starting with the same training edges (See Appendix Table E.12), the two datasets learn different knowledge, suggesting that animal babies are fuzzier

Table 6.5: Selected examples of high confidence edges containing "gosling" from the animal test set. There are no training edges containing "gosling" in the rooms dataset. Pink rows are misclassified. The italic row not a proposed relationship. It is included to show the score of the reverse edge.

Source	Target	Proposed Relationship	Score	Number of Owners	PMI
water	gosling	AtLocation	1.89	1067	0.18
swim	gosling	AtLocation	0.87	319	0.23
gosling	*swim*	*CapableOf*	*-0.22*		
young	gosling	AtLocation	0.83	362	0.31
gosling	lake	AtLocation	0.74	691	0.18
gosling	park	AtLocation	0.60	381	0.10
gosling	cute	HasProperty	0.77	715	0.12
gosling	animal	IsA	2.44	698	0.01
gosling	bird	IsA	2.03	3431	0.24
gosling	grass	IsA	1.12	423	0.11
gosling	wild life	IsA	1.02	1097	0.22
gosling	waterfowl	IsA	0.90	500	0.35
gosling	pond	IsA	0.41	549	0.22
gosling	duck	IsA	0.37	360	0.10
gosling	canadian goose	IsA	0.27	461	0.34
gosling	baby	IsA	0.27	1266	0.30

than human babies and less likely to be in kitchens. Tables of high confidence proposed edges for both datasets are available in Appendix Table E.13 and E.14.

Exploring the proposed edges concept by concept shows the variety of knowledge learned, but how are individual classifiers behaving? Tables 6.6 and 6.7 show the top four highest scoring edge proposals for each relationship type. Here many of the highest scoring predictions use the same concepts. For example, "feed" as the source produces a high score from the RecievesAction classifier in the animal dataset. This is similar to the behavior explored in Section 5.2 where classifiers appear to be learning patterns of concepts. In many cases, the pattern matching is successful. For example, "kitchen" is usually a location and "cute" is usually a property. However, it fails when the classifier is trained on fewer edge examples, as in "black used for love" and "woman made of glass" from the animal dataset.

(a) "mother has a baby" in the animal dataset

(b) "mother has a baby" in the room dataset

(c) "baby has property white" in the animal dataset

(d) "baby has property white" in the room dataset

Figure 6.5: Selected images from edges containing "baby" learned by both test sets. Each rows shows the same edges with the examples from the animal dataset on the left and the examples from the room dataset on the right. We can see that while the knowledge comes from very different sources, but the general facts about babies are the same.

(a) "baby has property fuzzy" in the animal dataset

(b) "baby at location kitchen" in the room dataset

Figure 6.6: Selected images from high confidence edges containing "baby" learned by only one dataset. The edges are only learned by one dataset's classifiers, suggesting that baby animals are more fuzzy and that they are less likely to be found in kitchens than human babies.

Table 6.6: Top four highest scoring proposed edges in room dataset for each relationship. Edges are sorted by relationship then by score.

Source	Target	Proposed Relationship	Score	Number of Owners	Normalized PMI
sunlight	countryside	AtLocation	2.87	680	0.18
horn	zoo	AtLocation	2.89	608	0.08
wool	farm	AtLocation	3.01	1381	0.16
sunlight	park	AtLocation	3.02	663	0.01
bird	swim	CapableOf	1.43	1334	0.13
animal	graze	CapableOf	1.43	1021	0.06
animal	swim	CapableOf	1.55	576	0.04
animal	fly	CapableOf	1.77	797	0.08
animal	tongue	HasA	1.76	784	0.05
wild life	beak	HasA	1.84	741	0.22
wild life	wing	HasA	1.86	1086	0.20
wild life	feather	HasA	2.14	1768	0.20
pet	cute	HasProperty	2.06	8104	0.24
hound	cute	HasProperty	2.06	19634	0.05
farm animal	cute	HasProperty	2.07	799	0.12
wild life	black	HasProperty	2.12	754	0.03
ara	bird	IsA	2.56	6307	0.24
ara	animal	IsA	2.84	2175	0.12
black sheep	animal	IsA	2.92	744	0.03
anser anser	bird	IsA	3.27	716	0.21
art work	stone	MadeOf	0.99	770	0.06
sun rise	water	MadeOf	1.02	1280	0.19
woman	glass	MadeOf	1.08	980	0.06
holiday	sand	MadeOf	1.19	650	0.16
sea	wave	PartOf	0.64	2073	0.39
canine	head	PartOf	0.67	622	0.08
window	automobile	PartOf	0.81	510	0.05
wing	duck	PartOf	0.83	543	0.15
cow	feed	ReceivesAction	0.95	1063	0.15
goose	feed	ReceivesAction	0.97	916	0.02
goat	feed	ReceivesAction	0.98	1085	0.13
bird	feed	ReceivesAction	1.20	1588	0.07
market	travel	UsedFor	1.86	700	0.18
black	love	UsedFor	1.92	515	0.05
paint	love	UsedFor	1.97	1175	0.04
eye	love	UsedFor	2.28	690	0.02

Table 6.7: Top four highest scoring proposed edges in room dataset for each relationship. Edges are sorted by relationship then by score.

Source	Target	Proposed Relationship	Score	Number of Owners	Normalized PMI
vegetarian	kitchen	AtLocation	3.12	521	0.05
salt	kitchen	AtLocation	2.90	770	0.17
curtain	home	AtLocation	2.81	584	0.10
toy	bedroom	AtLocation	2.79	911	0.07
infant	smile	CapableOf	0.95	515	0.28
boy	sit	CapableOf	0.93	667	0.21
man	sleep	CapableOf	0.91	637	0.15
man	reflect	CapableOf	0.87	740	0.14
bedroom	leg	HasA	1.97	1082	0.11
garden	window	HasA	1.63	502	0.13
apartment	window	HasA	1.57	998	0.07
dwelling	window	HasA	1.49	2166	0.19
pet	cute	HasProperty	2.06	953	0.42
food	yellow	HasProperty	2.02	787	0.06
lifestyle	white	HasProperty	2.02	644	0.19
day	white	HasProperty	1.95	762	0.08
mutt	animal	IsA	2.57	1604	0.31
red	vegetable	IsA	2.34	691	0.27
pup	animal	IsA	2.29	1604	0.32
dude	beverage	IsA	2.28	558	0.17
night	glass	MadeOf	1.31	954	0.03
art	wood	MadeOf	1.29	558	0.07
view	glass	MadeOf	1.21	1294	0.08
book	glass	MadeOf	1.17	565	0.02
leaf	nature	PartOf	1.60	586	0.34
frame	living room	PartOf	1.49	591	0.14
bloom	living room	PartOf	1.39	1237	0.00
african american	living room	PartOf	1.33	856	0.10
meal	eat	ReceivesAction	0.51	712	0.49
tomato	cook	ReceivesAction	0.41	802	0.25
dwelling	design	ReceivesAction	0.23	685	0.13
interior	paint	ReceivesAction	0.06	857	0.04
kitchen	eat	UsedFor	2.69	2305	0.18
kitchen	dine	UsedFor	2.49	757	0.04
bed	relax	UsedFor	2.10	894	0.24
bathroom	relax	UsedFor	2.06	559	0.06

Chapter 7

Future Work

7.1 Human annotation

The small hand-labeled sample shows that the proposed edges have an average accuracy of 64%. This is not accurate enough to add the edges directly to ConceptNet or Freebase, but the proposed edges can be quickly and efficiently reviewed by human annotators. Human annotation is already extensively used to add edges to ConceptNet and Freebase [35, 20]. However, human annotators are likely to miss more unusual facts like "sheep at location sea".

The method proposes very complete list of potential edges three orders of magnitude smaller than the number of all possible concept pairs. Human labelers are expensive and are better at making easy judgments, such as answering yes or no questions. Therefore, presenting them with a smaller list of more likely edges is more effective use of their time and energy. Once labeled, the edges can be added to the knowledge bases and used to further improve the classifiers.

7.2 Including visual representations in classification

One of the most important future goals for this project is incorporating visual representations of the images in the classification. The visual information in the images may prove helpful for word sense disambiguation, learning more robust classifiers, and identifying the most interesting edges. Currently, the method does not distinguish between different word senses, such as "Turkey" the country and "turkey" the bird. This is a limitation of using only the tags of the image, which do not indicate which sense is intended, and of the GloVe vector, which also uses the same representation for both word senses. However, the images for each word sense would differ greatly. By adding visual features to the method, I could leverage these differences to distinguish between the two senses, hopefully leading to more robust and accurate classification.

7.3 Building visual edge detectors

Another benefit of using visual representations would be the potential to provide edge detectors for images without tags or with incomplete tags. Many images on the internet are not tagged, and tagging is another expensive manual annotation task. I could train edge detectors that would learn a visual model of an edge from the tagged images in the datasets. I could then label untagged images with appropriate concepts and edges. NEIL [9], Divvala et al. [12], and VisKE [31], all learn visual models of the concepts and knowledge that they identify.

7.4 Frequent concept sets

Beyond simple concept co-occurrence, I can also examine frequent concept sets or sequences of more than two concepts using techniques from Data Mining. The subfield of frequent pattern mining has developed several algorithms for efficiently discovering frequent sets. Borgelt et al. [5] provides a good overview of different algorithms to extract frequent concept sets from the dataset.

Frequent concept sets provide us with interesting clusters of concepts which might indicate commonly co-occurring edges. Identifying commonly co-occurring edges might help predict whether or not a edge should exist. For example, the common concept cluster of "sheep", "grass", and "green", currently produces three edges, "sheep at location grass", "grass has property green", and "sheep has property green". The last edge, "sheep has property green" is a misclassification resulting in part from the high correlation between "sheep" and "green". Knowing that this correlation results from the latent concept "grass" might help limit such misclassifications.

Chapter 8

Conclusions

In conclusion, I have contributed a method for extracting potential commonsense edges in the form of concept pairs from a large collection of images, learning models of relationships from existing commonsense knowledge bases, and transferring the relationships to the concept pairs. This is the first work which attempts to learn visual commonsense knowledge from image collections through transfer learning. For the experiments, I collected two datasets with more than one million images each, extracted concept vocabularies from those datasets, collected a training set of edges from Freebase and ConceptNet containing concepts from the vocabulary, and proposed 93,850 new edges with an accuracy of 63%.

These proposed edges could be quickly and efficiently reviewed by human annotators. Then number of proposed edges is several orders of magnitude smaller than the number of possible edges presenting a more efficient review task. Future work would incorporate visual representations of the images into the method opening the door for learning visual edge detectors. Visual commonsense knowledge has great potential both for extending existing knowledge bases and for new applications in image understanding.

Appendix A

Merged relationships

This appendix lists the relationships from Freebase and ConceptNet which are merged to create the relationship types we use for training.

A.1 IsA

IsA is composed of Freebase relationships:

- /biology/organism_classification/higher_classification
- /biology/organism_classification/lower_classifications
- /biology/domesticated_animal/breeds
- /biology/animal_breed/breed_of

and ConceptNet relationship:

- IsA

A.2 GeographicContainment

GeographicContainment is composed of Freebase relationships:

- /location/administrative_division/country
- /location/administrative_division/first_level_division_of
- /location/administrative_division/second_level_division_of
- /location/location/partially_containedby
- /location/location/containedby

- /location/country/administrative_divisions
- /location/country/first_level_divisions
- /location/location/contains
- /location/location/partially_contains
- /location/country/capital

and ConceptNet relationships:

- PartOf if both source and target have location labels
- AtLocation if both source and target have location labels

A.3 GeographicAdjective

Geographic Adjective is composed of Freebase relationships:

- /location/location/adjectival_form
- /location/country/iso3166_1_shortname
- /location/country/fifa_code
- /location/country/iso_alpha_3
- /olympics/olympic_participating_country/ioc_code

A.4 AtLocationGeographic

AtLocationGeographic is composed of ConceptNet relationships:

- AtLocation if the source or target has a location label, but not both

Appendix B

Vocabulary

This appendix shows the vocabulary filtering and extention process for the top 50 most frequent concepts in each dataset. Details of the full process are given in Section 3.2.

Table B.1: Top 50 concepts from the room dataset at various stages of collection. The initial vocabulary is collected from the image tags only excluding English stopwords. The filtered vocabulary removes camera vocabulary, numbers, non-roman characters, and automatic Flickr tags (except for vision tags), and splits concatenated phrases when possible. The extended vocabulary adds phrases that are frequent in the dataset found using high PMI and local search on Freebase. The effective vocabulary is the vocabulary terms in the extended vocabulary for which we have GloVe representations. The frequency counts are shown for the extended vocabulary.

Initial Vocab	Filtered Vocab	Extended Vocab	Effective Vocab	Frequency
kitchen	kitchen	kitchen	kitchen	96741
bathroom	bathroom	bathroom	bathroom	61139
square	bedroom	bedroom	bedroom	45726
iphoneography	house	house	house	28745
squareformat	home	home	home	27680
bedroom	food	food	food	26798
instagramapp	room	room	room	13110
uploaded:by=instagram	window	window	window	12379
house	white	white	white	19290
home	living room	living room	living room	32153
food	red	modern architecture	modern architecture	1274
room	art	red	red	16707
light	christmas	art	art	14004
window	blue	christmas	christmas	12948
white	green	blue	blue	14544
livingroom	mirror	green	green	16192
red	cooking	mirror	mirror	7418
portrait	bed	cooking	cooking	15617
canon	hotel	bed	bed	16219
art	black	hotel	hotel	14386
christmas	living	black	black	12080
uploaded:by=flickrmobile	restaurant	living	living	2236

Continued on next page

Table B.1 – *Continued from previous page*

Initial Vocab	Filtered Vocab	Extended Vocab	Effective Vocab	Frequency
blue	family	restaurant	restaurant	15397
green	cat	family	family	14503
mirror	girl	cat	cat	13158
cooking	night	girl	girl	13231
bed	water	night	night	12491
nikon	new	water	water	12177
hotel	wedding	new	new	3585
flickriosapp:filter=nofilter	table	wedding	wedding	13561
black	party	table	table	6222
living	travel	party	party	10464
restaurant	sink	travel	travel	13059
family	apartment	sink	sink	1926
cat	interior	apartment	apartment	12175
girl	selfportrait	rock music	rock music	1025
night	london	interior	interior	8558
water	yellow	building construction	building construction	534
new	vintage	selfportrait	selfportrait	11763
wedding	reflection	london	london	11427
table	old	yellow	yellow	11559
party	summer	black girl	black girl	1657
travel	dinner	vintage	vintage	8628
sink	architecture	reflection	reflection	11001
apartment	flowers	old	old	8597
interior	toilet	chicago illinois	chicago illinois	966
blackandwhite	people	summer	summer	10528
film	winter	dinner	dinner	9963
selfportrait	design	architecture	architecture	9761
london	usa	flowers	flowers	10673

Table B.2: Top 50 concepts from the animal dataset at various stages of collection. The initial vocabulary is collected from the image tags only excluding English stopwords. The filtered vocabulary removes camera vocabulary, numbers, non-roman characters, and automatic Flickr tags (except for vision tags), and splits concatenated phrases when possible. The extended vocabulary adds phrases that are frequent in the dataset found using high PMI and local search on Freebase. The effective vocabulary is the vocabulary terms in the extended vocabulary for which we have GloVe representations. The frequency counts are shown for the extended vocabulary.

Initial Vocab	Filtered Vocab	Extended Vocab	Effective Vocab	Frequency
dog	dog	dog	dog	203792
cat	cat	cat	cat	163431
vision:outdoor=	vision:outdoor=	vision:outdoor=	sheep	95139
sheep	sheep	sheep	horse	78745
horse	horse	horse	rabbit	55859
rabbit	rabbit	rabbit	animal	69836
animal	animal	animal	chicken	66504
square	turkey	turkey	animals	60749
turkey	chicken	chicken	bird	59544
chicken	animals	animals	nature	61051
iphoneography	bird	bird	dogs	58519
squareformat	nature	nature	cats	56653
instagramapp	dogs	dogs	goat	48432
uploaded:by=instagram	cats	cats	goose	40975
animals	goat	goat	geese	41878
bird	goose	goose	birds	49781
nature	vision:sky=	vision:sky=	parrot	47039
dogs	geese	geese	white	33059
cats	birds	birds	cute	45374
goat	vision:mountain=	vision:mountain=	food	40962
goose	parrot	parrot	black	20076
vision:sky=	white	white	green	38494
geese	cute	cute	farm	29412
birds	food	food	pet	37840
vision:mountain=	black	black	water	37593
parrot	green	green	zoo	34126
white	farm	farm	horses	34476
cute	pet	pet	bunny	5306
food	water	water	snow	28512
black	zoo	zoo	blue	24501
green	horses	horses	park	19622
farm	bunny	bunny	wild life	31723
pet	snow	snow	puppy	30183
water	blue	blue	winter	29113

Continued on next page

Table B.2 – *Continued from previous page*

Initial Vocab	Filtered Vocab	Extended Vocab	Effective Vocab	Frequency
zoo	park	park	portrait	27697
horses	wildlife	wild life	red	25381
canon	vision:text=	vision:text=	cattle	23955
bunny	puppy	puppy	grass	27463
snow	winter	winter	landscape	26553
nikon	portrait	portrait	art	23325
blue	red	red	beach	26060
park	cattle	cattle	summer	25259
wildlife	grass	grass	sky	19815
vision:text=	landscape	landscape	pets	23966
puppy	art	art	lake	22120
winter	beach	beach	sun set	23294
portrait	vision:plant=	vision:plant=	cow	22078
red	summer	summer	kitten	22821
cattle	sky	sky	travel	21937
grass	pets	pets	spring	20957

Appendix C

Labeled vocabulary examples

This appendix provides some examples from the color and location vocabulary labels.

C.1 Colors

The following lists all the vocabulary labeled as colors from both datasets.

- amber
- aqua
- asparagus
- beige
- berry
- black
- blond
- blue
- brick
- bronze
- brown
- burgundy
- camel
- cardinal
- champagne
- charcoal
- cherry
- chestnut
- chocolate
- clear
- coffee
- copper
- coral
- cranberry
- cream
- dark blue
- denim
- eggplant
- fawn
- flame
- gold
- gray
- green
- grey
- iris
- jade
- lavender
- lemon
- lilac
- lime
- lion
- magenta
- magnolia
- midnight
- mint
- mustard
- navy blue
- olive
- orange
- orchid
- peach
- pear
- pearl
- pink
- pomegranate
- pumpkin
- purple
- raspberry
- red
- red hair
- rose
- ruby
- rust
- sage
- scarlet
- sepia
- silver
- stone
- straw
- tan
- teal
- terra cotta
- tomato
- transparent
- turquoise
- vanilla
- vert
- violet
- wheat
- white
- wine
- yellow

C.2 Locations

The following lists the top 200 most frequent vocabulary labeled as geographic locations from the room dataset. The total number of locations in the room dataset is 874. The animals dataset contains similar locations with only 97 different locations.

- london
- chicago illinois
- usa
- california
- japan
- france
- new york
- bath
- paris
- italy
- china
- england
- nsw australia
- europe
- san francisco
- canada
- spain
- chicago
- germany
- vienna austria
- florida
- tokyo
- valencia
- seattle
- mexico
- australia
- texas
- india
- thailand
- red house
- washington
- new york city
- morning sun
- old city
- barcelona
- scotland
- asia
- ireland
- berlin
- united states
- amsterdam
- los angeles
- portland
- pittsburgh pennsylvania
- turkey
- las vegas
- hong kong
- manhattan
- boston
- singapore
- toronto
- taiwan
- brooklyn
- oregon
- hawaii
- africa
- amsterdam museum
- fuji
- vancouver
- austin
- mobile
- colorado
- vegas
- america
- vietnam
- red square
- italia
- bangkok
- arizona
- argentina
- shanghai
- virginia
- malaysia
- michigan
- sweden
- nashville
- beijing
- madrid
- sydney
- ontario
- philadelphia
- belgium
- rome
- netherlands
- portugal
- greece
- edmonton alberta
- ohio
- melbourne
- new orleans

- illinois
- taipei
- brazil
- reunion
- georgia
- pennsylvania
- austria
- city park
- victoria
- beach park
- dublin
- wales
- edinburgh
- istanbul
- korea
- switzerland
- united kingdom' 'new zealand
- philippines
- brasil
- deutschland
- atlanta
- show room
- san diego
- peru
- montreal
- roman empire
- washington dc
- nevada
- massachusetts
- indonesia
- vienna
- norway
- wisconsin
- bali
- venice
- stone county
- tennessee
- kyoto
- buenos aires
- starbucks coffee
- russia
- new river
- new house
- prague
- orlando
- north west
- maine
- maryland
- sarasota florida
- wash
- pearl river
- miami
- chile
- denmark
- minnesota
- alaska
- north carolina
- north yorkshire
- nice
- poland
- south africa
- cambridge
- oxford
- cuba
- manchester
- church village
- old castle
- stockholm
- budapest
- finland
- frankfurt germany
- seoul
- caribbean
- brussels
- dallas
- black rock
- houston
- glass bowl
- hollywood
- louisiana
- florence
- denver
- new jersey
- morocco
- utah
- iceland
- indiana
- costa rica
- copenhagen
- cambodia
- egypt
- baltimore
- memphis
- starbucks
- glasgow
- munich
- minneapolis
- louvre museum
- israel
- model city
- disneyland
- tuscany
- new mexico
- green lake
- home park
- osaka
- river thames
- thames river
- missouri

Appendix D

Predicting edge existence

This appendix contains precision recall curves and threshold plots to accompany the ROC curves in Figure 5.5

I use one of the metrics listed in the legend to predict whether or not an ordered concept pair is an edge. Pairs with a value above a certain threshold are labeled as edges. These labels are compared to the hand-labeled ground truth. The curve is plotted by varying the threshold which varies the precision and recall of the prediction. Figure D.1 use the same values as Figure 5.5a and Figure D.2 uses the same values as 5.5b. The figure on the left shows the precision recall curve. The figure on the right shows thresholds corresponding to ROC and precision recall curves. In the threshold plot, the number of images, number of owners, and highest classifier score are normalized. The other metrics show their actual values.

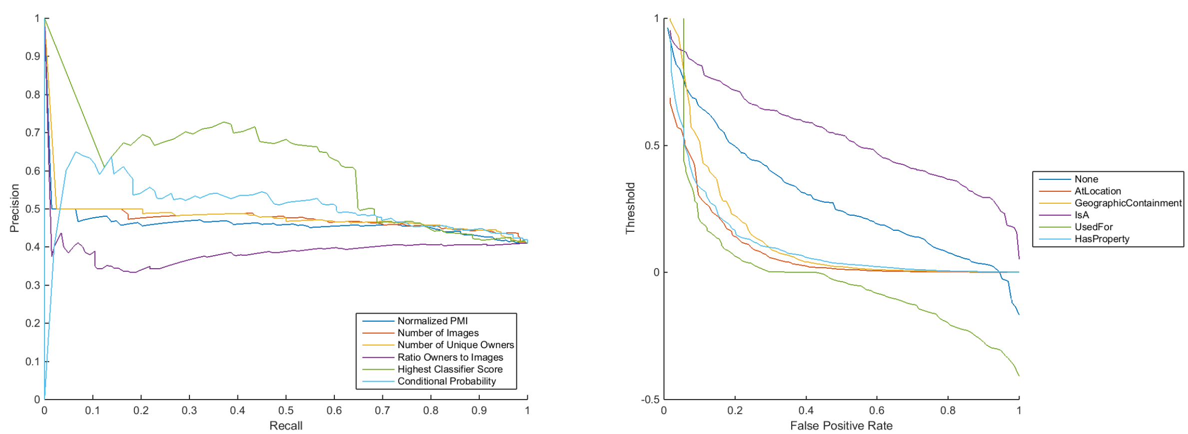


Figure D.1: Examining directed edge prediction for the animal dataset

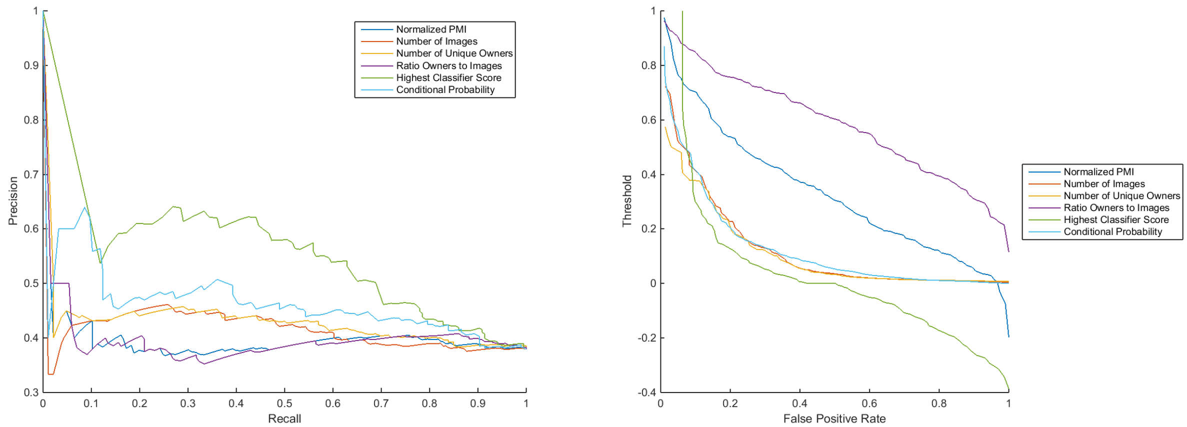


Figure D.2: Examining directed edge prediction for the room dataset

Appendix E

Proposed edge details

E.1 Proposed edges from the animal dataset

This appendix contains more complete results from the animal dataset to accompany Chapter 6.

Table E.1: All training edges with "bird" as the source in animal dataset. Rows are sorted by relationship type, then by number of owners. There are also 47 edges with "bird" as the target in the training set, all with the relationship IsA, representing different species of bird. These examples accompany Figure 6.4 and Table E.2.

Source	Target	Ground Truth Relationship	Number of Owners	PMI
bird	zoo	AtLocation	8596	0.20
bird	tree	AtLocation	5613	0.08
bird	sky	AtLocation	3331	0.02
bird	sea	AtLocation	2870	0.00
bird	beach	AtLocation	2339	-0.04
bird	cage	AtLocation	1752	0.24
bird	wild	AtLocation	1683	0.12
bird	nest	AtLocation	1416	0.22
bird	field	AtLocation	1190	-0.09
bird	wood	AtLocation	962	-0.03
bird	forest	AtLocation	799	0.00
bird	countryside	AtLocation	529	-0.10
bird	bush	AtLocation	281	0.03
bird	state park	AtLocation	224	-0.04
bird	air	AtLocation	189	-0.03
bird	roof	AtLocation	146	-0.05
bird	lawn	AtLocation	142	-0.02
bird	flight	CapableOf	4752	0.28
bird	perch	CapableOf	1040	0.27
bird	walk	CapableOf	946	-0.07
bird	land	CapableOf	835	-0.03
bird	feather	HasA	8594	0.32
bird	beak	HasA	3608	0.38

Continued on next page

Table E.1 – *Continued from previous page*

Source	Target	Ground Truth Relationship	Number of Owners	PMI
bird	eye	HasA	2453	0.09
bird	egg	HasA	1001	0.02
bird	claw	HasA	374	0.14
bird	tail	HasA	305	0.00
bird	wing	HasA;PartOf	4318	0.25
bird	cute	HasProperty	2525	-0.08
bird	beautiful	HasProperty	1905	0.09
bird	pretty	HasProperty	763	0.09
bird	cool	HasProperty	319	0.01
bird	animal	IsA	24615	0.18
bird	pet	IsA	4048	-0.01
bird	owl	IsA	1202	0.19
bird	food	IsA	1079	-0.16
bird	mammal	IsA	490	-0.05
bird	predator	IsA	259	0.15
bird	flock	PartOf	1849	0.12
bird	watch	UsedFor;CapableOf	156	-0.04
bird	eat	UsedFor;CapableOf;ReceivesAction	1122	0.03
bird	fly	UsedFor;HasProperty;CapableOf	2568	0.26

Table E.2: Selected examples of high confidence edges containing "bird" from the animal test set. The edges are the highest scoring proposals with more than 500 owners and PMI greater than zero for selected relationship. Bold rows are illustrated in Figure 6.4. Pink rows are misclassified.

Source	Target	Proposed Relationship	Score	Number of Owners	PMI
bird	lake	AtLocation	1.41	6474	0.13
bird	river	AtLocation	1.18	3130	0.08
bird	ocean	AtLocation	0.85	2870	0.03
bird	pond	AtLocation	0.80	3556	0.17
bird	wetland	AtLocation	0.74	1385	0.20
bird	jungle	AtLocation	0.69	611	0.14
bird	swim	CapableOf	1.41	1334	0.13
bird	hunt	CapableOf	0.30	606	0.02
bird	**foot**	**HasA**	**0.67**	**568**	**0.03**
bird	baby	HasA	0.27	2126	0.02
bird	plumage	HasA	0.27	525	0.27
bird	**bright**	**HasProperty**	**1.15**	**723**	**0.14**
bird	young	HasProperty	0.93	782	0.086
bird	yellow	HasProperty	0.86	4239	0.16
bird	black	HasProperty	0.85	2645	0.05
bird	green	HasProperty	0.85	7270	0.14
bird	white	HasProperty	0.80	4894	0.09
bird	blue	HasProperty	0.80	5417	0.14
bird	female	HasProperty	0.65	2014	0.10
bird	male	HasProperty	0.61	1972	0.15
bird	feed	ReceivesAction	1.19	1588	0.07
anser anser	bird	IsA	3.27	716	0.21
ara	bird	IsA	2.57	6307	0.24
duckling	bird	IsA	2.17	508	0.17
gosling	bird	IsA	2.03	3431	0.24
amazon	bird	IsA	1.86	827	0.21
branta canadensis	bird	IsA	1.85	1812	0.27

Table E.3: Training edges containing “farmland” in animal dataset

Source	Target	Ground Truth Relationship	Number of Owners	PMI
farmland	farm	UsedFor	1061	0.31
farmland	country	AtLocation	243	0.34
farmland	countryside	AtLocation	378	0.43
farmland	land	IsA	119	0.38
cow	farmland	AtLocation	959	0.24
horse	farmland	AtLocation	175	-0.04

(a) farmland used for sheep

(b) bull at location farmland

Figure E.1: Selected images from high confidence edges containing “farmland” from animal test set. These image grids illustrate the highlighted edges in Table E.4. These two examples show two possible relationships between “farmland” and animals. A similar AtLocation relationship, “cow at location farmland” is present in the training data, but there is no similar UsedFor relationship.

Table E.4: Selected examples of high confidence edges containing “farmland” from the animal test set. The edges are the highest scoring proposals with more than 300 owners and PMI greater than zero for selected relationship. Bold rows are illustrated in Figure 6.4.

Source	Target	Proposed Relationship	Score	Number of Owners	PMI
farmland	**sheep**	**UsedFor**	**0.74**	**665**	**0.21**
farmland	cattle	UsedFor	0.60	959	0.31
bull	**farmland**	**AtLocation**	**1.24**	**959**	**0.20**
grass	farmland	AtLocation	0.90	341	0.27
tree	farmland	AtLocation	0.87	369	0.17

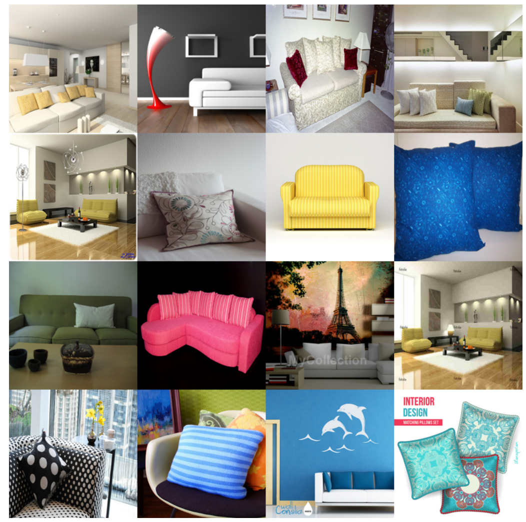
E.2 Proposed edges from the room dataset

This section contains more complete results from the room dataset to accompany Chapter 6.

Table E.5: All training edges containing "sofa" in room dataset. Rows are sorted by relationship type, then by number of owners. These examples accompany Figure E.2 and Table E.6.

Source	Target	Ground Truth Relationship	Number of Owners	PMI
sofa	living room	AtLocation	6898	0.49
sofa	house	AtLocation	1194	0.14
sofa	home	AtLocation	1354	0.16
cat	sofa	AtLocation	509	0.05
sofa	leg	HasA	193	0.18
sofa	chair	IsA	1630	0.36
sofa	leather	MadeOf	443	0.44
cushion	sofa	PartOf	499	0.46
fabric	sofa	PartOf	314	0.19
sofa	cushion	ReceivesAction	499	0.46
sofa	relax	UsedFor	1064	0.38
sofa	comfort	UsedFor	771	0.38
sofa	sit	UsedFor	398	0.35
sofa	sleep	UsedFor	362	0.14

(a) Male AtLocation Sofa



(b) Pillow PartOf Sofa

Figure E.2: Selected images from high confidence edges containing "sofa" from room test set. These image grids illustrate the highlighted edges in Table E.6. Figure E.2a shows "male" being used as a synonym for "man" in stock photo tagging.

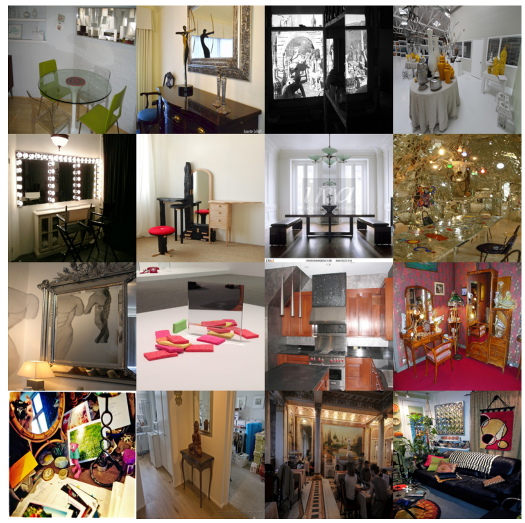
Table E.6: Selected examples of high confidence edges containing "sofa" from the room test set. The edges are the highest scoring proposals with more than 500 owners and PMI greater than zero for selected relationship. Bold rows are illustrated in Figure E.2.

Source	Target	Proposed Relationship	Score	Number of Owners	PMI
sofa	apartment	AtLocation	2.75	871	0.20
sofa	wall	AtLocation	1.58	615	0.22
sofa	rug	AtLocation	0.83	1098	0.35
sofa	table	AtLocation	0.68	1203	0.36
sofa	window	AtLocation	0.65	1638	0.23
dog	sofa	AtLocation	1.69	580	0.11
male	**sofa**	**AtLocation**	**1.45**	**778**	**0.22**
female	sofa	AtLocation	1.39	1329	0.22
boy	sofa	AtLocation	1.27	778	0.11
girl	sofa	AtLocation	1.06	1329	0.08
sofa	white	HasProperty	0.94	1013	0.15
sofa	furniture	IsA	0.32	1377	0.39
dwelling	sofa	HasA	1.16	1354	0.34
pillow	**sofa**	**PartOf**	**0.17**	**1030**	**0.36**

Figure E.3: Training edges containing “mirror” in room dataset

Source	Target	Ground Truth Relationship	Number of Owners	PMI
mirror	bedroom	AtLocation	2283	0.24
mirror	wall	AtLocation	317	0.14
mirror	glass	IsA;MadeOf	719	0.11

(a) Mirror AtLocation Bed



(b) Mirror IsA Sculpture

Figure E.4: Selected images from high confidence edges containing “mirror” from room test set. These image grids illustrate the highlighted edges in Table E.7

Table E.7: Selected examples of high confidence edges containing “mirror” from the room test set. The edges are the highest scoring proposals with more than 300 owners and PMI greater than zero for selected relationship. Bold rows are illustrated in Figure E.4. Pink rows are misclassified.

Source	Target	Proposed Relationship	Score	Number of Owners	PMI
mirror	home	AtLocation	1.92	452	0.02
mirror	room	AtLocation	1.29	450	0.12
mirror	**bed**	**AtLocation**	**0.56**	**759**	**0.18**
mirror	dwelling	AtLocation	0.54	452	0.08
mirror	living room	AtLocation	0.02	731	0.08
mirror	white	HasProperty	1.66	530	0.09
mirror	blue	HasProperty	1.37	316	0.05
mirror	**sculpture**	**IsA**	**0.74**	**416**	**0.02**
mirror	reflect	UsedFor	0.15	2673	0.30
self	mirror	AtLocation	0.60	777	0.30
female	mirror	AtLocation	0.39	1054	0.10
male	mirror	AtLocation	0.36	359	0.06

Table E.8: Training edges containing "garlic" in room dataset

Source	Target	Ground Truth Relationship	Number of Owners	PMI
garlic	kitchen	AtLocation	967	0.21
garlic	dinner	AtLocation	175	0.23
garlic	food	IsA	561	0.30
garlic	spice	IsA	164	0.40
garlic	ingredient	IsA	138	0.45

(a) garlic at location knife

(b) still life made of garlic

Table E.9: Selected images from high confidence edges containing "garlic" from room test set. These image grids illustrate the highlighted edges in Table E.10.

Table E.10: Selected examples of high confidence edges containing "garlic" from the room test set. The edges are the highest scoring proposals with more than 100 owners and PMI greater than zero for selected relationship. Bold rows are illustrated in Figure E.9. Pink rows are misclassified.

Source	Target	Proposed Relationship	Score	Number of Owners	PMI
garlic	**knife**	**AtLocation**	**1.56**	**107**	**0.26**
garlic	pan	AtLocation	1.09	113	0.20
garlic	garden onion	AtLocation	0.39	370	0.41
garlic	white	HasProperty	1.01	108	0.09
garlic	green	HasProperty	0.90	141	0.14
garlic	vegetable	IsA	0.92	278	0.38
garlic	herb	IsA	0.64	141	0.40
garlic	vegetarian food	IsA	0.43	104	0.38
chilli pepper	garlic	HasA	0.56	266	0.36
vegetarian cuisine	garlic	HasA	0.56	104	0.40
vegetarian	garlic	HasA	0.55	104	0.26
still life	**garlic**	**MadeOf**	**0.33**	**152**	**0.30**
recipe	garlic	MadeOf	0.24	126	0.36
salt	garlic	UsedFor	0.17	102	0.37

Figure E.5: Selected images from edge "new year has property happy". The images contain three different "happy new year" themes: written messages, fancy meals, and groups of celebrating people, showing how different image motivs can represent the same fact.

Table E.11: Selected examples of high confidence edges containing "new year" from the rooms test set. There are no training edges containing "new year" in the rooms dataset. Bold rows are illustrated in Figure E.5. Pink rows are misclassified.

Source	Target	Proposed Relationship	Score	Number of Owners	PMI
new year	**happy**	**HasProperty**	**0.19**	**109**	**0.24**
new year	woman	IsA	0.54	160	0.02
new year	female	IsA	0.31	160	0.01
new year	holiday	IsA	0.23	208	0.10
new year	fun	UsedFor	0.85	119	0.06
new year	party	UsedFor	0.48	547	0.21
new year	celebration	UsedFor	0.17	173	0.33

E.3 Proposed edges dataset comparison

This section contains more complete results from the cross dataset comparison of edges containing the concept "baby" to accompany Chapter 6 and specifically Figure 6.6.

Table E.12: Training edges containing "baby" in both datasets. Rows are sorted by relationship type, then by number of owners. These examples accompany Figure 6.6 and Tables E.13 and E.14.

Source	Target	Ground Truth Relationship	Number of Owners	PMI
baby	home	AtLocation	374	0.027
toy	baby	AtLocation	340	0.17
baby	house	AtLocation	224	-0.055
baby	crib	AtLocation	156	0.445
baby	rug	AtLocation	108	0.067
baby	play	CapableOf	382	0.246
baby	sleep	CapableOf	311	0.154
baby	laugh	CapableOf	116	0.212
woman	baby	HasA	1189	0.10
cat	baby	HasA	234	0.04
animal	baby	HasA	196	0.14
baby	hair	HasA	110	0.03
baby	cute	HasProperty	861	0.361
baby	happy	HasProperty	306	0.248
baby	young	HasProperty	182	0.244
baby	small	HasProperty	119	0.184
baby	mammal	IsA	255	0.14
puppy	baby	IsA	104	0.16
baby	family	PartOf	737	0.203
sleep	baby	UsedFor	311	0.15

Table E.13: Selected Examples of High Confidence edges containing "baby" from the room test set. Bold rows are illustrated in Figure 6.6. Pink rows are misclassified. Green rows are learned by the classifiers for the animal dataset as well. Ellipsis indicates a number of excluded high confidence edges.

Source	Target	Proposed Relationship	Score	Number of Owners	PMI
baby	bedroom	AtLocation	1.81	840	0.08
baby	kitchen	AtLocation	1.80	1000	0.01
baby	bed	AtLocation	0.84	386	0.09
baby	smile	CapableOf	0.61	515	0.26
baby	**white**	**HasProperty**	**1.34**	**488**	**0.06**
baby	living room	PartOf	0.67	550	0.06
dwelling	baby	HasA	1.23	374	0.056
caucasian	baby	HasA	0.85	488	0.230
man	baby	HasA	0.82	914	0.026
female	baby	HasA	0.78	1189	0.109
...					
mother	**baby**	**HasA**	**0.42**	**787**	**0.366**
bed	baby	UsedFor	0.59	386	0.092
bath	baby	UsedFor	0.34	406	0.188

Table E.14: Selected Examples of High Confidence edges containing "baby" from the animal test set. Bold rows are illustrated in Figure 6.6. Pink rows are misclassified. Green rows are learned by the classifiers for the room dataset as well. Ellipsis indicates a number of excluded high confidence edges.

Source	Target	Proposed Relationship	Score	Number of Owners	PMI
baby	farm	AtLocation	1.12	863	0.028
baby	field	AtLocation	1.11	308	0.002
baby	grass	AtLocation	0.30	862	0.073
baby	feather	HasA	1.13	362	0.052
baby	green	HasProperty	1.11	760	0.019
...					
baby	**white**	**HasProperty**	**0.99**	**693**	**0.031**
baby	blue	HasProperty	0.93	483	0.021
baby	sweet	HasProperty	0.92	596	0.220
baby	fluffy	HasProperty	0.88	537	0.213
baby	fuzzy	HasProperty	0.82	315	0.237
baby	adorable	HasProperty	0.81	662	0.276
baby	bird	IsA	1.42	2126	0.019
baby	female	IsA	0.71	1242	0.069
baby	male	IsA	0.64	857	0.024
baby	pet	IsA	0.60	1038	0.027
calf	baby	IsA	0.56	380	0.200
grey	baby	UsedFor	0.90	305	0.039
natural light	baby	UsedFor	0.78	403	0.081
gray	baby	UsedFor	0.73	305	0.066
wild life	baby	HasA	1.57	904	0.034
farm animal	baby	HasA	0.88	392	0.135
domestic cat	baby	HasA	0.73	2266	0.013
mother	**baby**	**HasA**	**0.39**	**983**	**0.330**

References

[1] Flickr services. Web, July 2015.

[2] flickr.com site overview. Web, July 2015.

[3] Tamara L Berg and Alexander C Berg. Finding iconic images. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2009.

[4] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[5] Christian Borgelt. Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):437–456, 2012.

[6] Andrew Carlson, Justin Betteridge, Richard C Wang, Estevam R Hruschka Jr, and Tom M Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 101–110. ACM, 2010.

[7] Kai-Wei Chang, Wen tau Yih, and Chris Meek. Multi-relational latent semantic analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL Association for Computational Linguistics, October 2013.

[8] Junpeng Chen and Juan Liu. Combining conceptnet and wordnet for word sense disambiguation. In *IJCNLP*, pages 686–694, 2011.

[9] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1409–1416. IEEE, 2013.

[10] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[11] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *Computer Vision–ECCV 2014*, pages 48–64. Springer, 2014.

[12] Santosh K Divvala, Alireza Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3270–3277. IEEE, 2014.

[13] Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C Berg, et al. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 762–772. Association for Computational Linguistics, 2012.

[14] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A Efros. What makes paris look like paris? *ACM Transactions on Graphics (TOG)*, 31(4):101, 2012.

[15] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM, 2014.

[16] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

[17] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[18] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *arXiv preprint arXiv:1212.4522*, 2012.

[19] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *Computer Vision–ECCV 2014*, pages 529–545. Springer, 2014.

[20] Google. Freebase API. `https://developers.google.com/freebase`, 2014.

[21] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 951–958. IEEE, 2009.

[22] Hugo Liu and Push Singh. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[24] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013.

[25] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543, 2014.

[27] Rahul Raguram and Svetlana Lazebnik. Computing iconic summaries of general visual concepts. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[28] Dheeraj Rajagopal, Daniel Olsher, Erik Cambria, and Kenneth Kwok. Commonsense-based topic modeling. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 6. ACM, 2013.

[29] Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378, 2013.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April 2015.

[31] Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1456–1464, 2015.

[32] Mehdi Samadi, Manuela M Veloso, and Manuel Blum. Openeval: Web information query evaluation. In *AAAI*. Citeseer, 2013.

[33] Ashutosh Saxena, Ashesh Jain, Ozan Sener, Aditya Jami, Dipendra K Misra, and Hema S Koppula. Robobrain: Large-scale knowledge engine for robots. *arXiv preprint arXiv:1412.0691*, 2014.

[34] Lei Shi and Rada Mihalcea. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Computational linguistics and intelligent text processing*, pages 100–111. Springer, 2005.

[35] Robert Speer and Catherine Havasi. Conceptnet 5: A large semantic network for relational knowledge. In *The Peoples Web Meets NLP*, pages 161–176. Springer, 2013.

[36] Munirathnam Srikanth, Joshua Varner, Mitchell Bowden, and Dan Moldovan. Exploiting ontologies for automatic image annotation. In *Proceedings of the 28^{th} annual international ACM SIGIR conference on Research and development in information retrieval*, pages 552–558. ACM, 2005.

[37] Luis Von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78. ACM, 2006.

[38] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.