Semantic Spaces from Images

A master thesis presented

by

Adam Liška

September, 2013

UNIVERSITÀ DEGLI STUDI
DI TRENTO

THE UNIVERSITY OF
MELBOURNE

Erasmus Mundus European Master in
Language & Communication Technologies (LCT)

Master of Computer Science, University of Melbourne

Master of Cognitive Science, Università degli Studi di Trento

Thesis advisor
Marco Baroni, CIMeC, Università degli Studi di Trento

Author
Adam Liška

Semantic Spaces from Images

Abstract

In vector space models, the meanings of concepts or words are represented as points in high-dimensional vector spaces, also referred to as semantic spaces. These spaces are usually derived automatically from large collections of texts, directly exploiting the sets of contexts in which individual words appear and considering these contexts to be the key constituents of semantic meaning.

The weakness of existing models is that they completely rely on linguistic input, although there is a growing body of evidence that other modalities also contribute to forming semantic representations. This has led to attempts to enrich existing semantic spaces with perceptual information. However, there are two problems that need to be addressed: (1) How to incorporate perceptual information into existing vector spaces? (2) How to acquire this type of data?

In this work, we focus mostly on the latter problem and present a framework for extraction of concept representations from images. The procedure starts with a collection of images that are all tagged with a given concept. Each image is represented by a set of features called visual words. These features constitute the visual context of the given concept similarly to the linguistic context used in textual models. In this way, we arrive at having two sets of contexts for each concept: a set of linguitic contexts derived from text corpora and a set of visual contexts extracted from images. These two sources of information are then used to create two distinct conceptual representations, textual and visual, which can be combined to create a final representation of the given concept.

Key words: distributional semantics, lexical semantics, semantic memory, concept representation

Contents

List of Figures

List of Tables

Citations to Previously Published Work

Portions of Chapter 3 have appeared in the following paper:

> Elia Bruni, Ulisse Bordignon, Adam Liška, Jasper Uijlings, and Irina Sergienya. VSEM: An open library for visual semantics representation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics (ACL).

Acknowledgments

First of all, I would like to express my gratitude to Marco Baroni and Elia Bruni for leading me through the whole project, for the constant flow of ideas and for all their assistance and comments. I have enjoyed working with you very much and I hope this is not the end!

I am grateful to the LCT Consortium, namely Valia Kordoni, Bobbye Pernice and Ivana Kruijff-Korbayová, for funding my studies and giving me the opportunity to study at such diverse places as Melbourne and Rovereto. It has been an unforgettable experience and it has broadened my horizons in ways which I had never imagined. An important "thank you" goes to Tim Baldwin and Nicu Sebe, for coordinating the program at the two universities and for helping me with all the administrative issues that arised.

Last but not least, I would like to thank all my friends in Melbourne and Rovereto for making my studies enjoyable and fun and for their great company. It would not have been possible without you.

Dedicated to my family.

Chapter 1

Introduction

1.1 What is a concept and why is it important?

Every day, we encounter, interact with or talk about an uncountable number of objects. We can recognize and name the objects upon seeing or hearing, we know how to use them, we can group them together based on their similarity. Consider for example the way we think about dogs. There are dogs of numerous shapes and colours and sizes, but we regard all of them as belonging to a single class of objects. When you read a story about a dog that you have not seen before, you can draw some inferences about its behaviour and other characteristics using your mental representation of the dog as a species. This mental representation of a set of objects is called a **concept**. As such, concepts are fundamental components of human cognition and play an important role in learning, verbal communication, object recognition and many other areas (Kiefer and Pulvermüller, 2012).

For a long time, the study of concepts had been the domain of philosophy. Nowadays, however, concepts have already found their way into many other disciplines and are analyzed from many other points of view. Cognitive scientists and psychologists investigate how concepts are represented in mind and brain and how this representation affects our action. Researchers in lexical semantics, a subfield of linguistics that studies the meaning of words and its contribution to the meaning of complex expressions, consider concepts to be major constituents of word meaning (Meteyard et al., 2012). In addition to that, concepts and their representations also play an important role in artificial intelligence where making computers understand human concepts and language would represent a major step forward. It could change the very nature of human-computer interaction, expand the range of applications for computers in everyday life and bring important contributions to fields such as robotics.

In humans, knowledge about concepts and the basic processes that act upon them is the domain of **semantic memory** (Binder and Desai, 2011; McRae and Jones, 2012). Although it is generally accepted that we learn concepts through our continuous perceptual experience of the world, there is no agreement among cognitive

scientists and neuroscientists about the actual nature of conceptual representations (Mahon and Caramazza, 2009). Semantic memory used to be regarded as a long-term amodal storage system and the concepts represented there were believed to be detached from the sensory-motor brain systems (Kiefer and Pulvermüller, 2012). However, this view is nowadays being contested and the evidence that conceptual representations are grounded in perceptual experience is growing (Barsalou, 2008). This has been accompanied by a shift from localist to distributed approaches to representations, in which a concept is believed to be encoded by an activation pattern over a set of representational units (McRae and Jones, 2012). These are better suited to encode thousands of different categories of objects and actions that humans are capable of recognizing.

1.2 Semantic Spaces

In parallel to the advances in cognitive science and neuroscience, several approaches to representing certain aspects of concepts have been developed in other fields. **Vector space models**, which have proved to be successful in many applications, represent the meanings of concepts or words as points in high-dimensional arithmetic vector spaces, also referred to as **semantic spaces**. There are at least two good reasons to use vector spaces. First, individual vector components can stand for specific features (such as size, animacy or context), which is a natural way how to characterize a concept. Second, the notion of "distance" or "similarity" between concepts reduces to the distance between representation vectors in the vector space. These implementation details are always specified by the particular model in use.

In an early attempt to *measure meaning* using semantic spaces, Osgood et al. (1957) defined each dimension by a pair of adjectives that were opposite in meaning, e.g. *happy–sad*, *hard–soft* or *tall–short*. Subjects were then asked to judge concepts against a series of such scales, and by averaging the scores across subjects the authors arrived at conceptual representations which were then used to describe various psychological phenomena.

Models of word meaning developed by computational linguists or computer scientists usually make use of large collections of texts to derive word representations in an automatic fashion. There are two main methods to do so. The first one directly exploits the sets of contexts in which the words appear and considers them to be the key constituents of semantic meaning. The dimensions of such semantic spaces stand for individual context items – for example documents or co-occurring words. In this way, the words *apple*, *fruit* and *pear* will end up having similar representation vectors because they usually appear in similar contexts – in texts about fruits, food, gardening etc. On the other hand, the word *petroleum* generally appears in different contexts and its representation will therefore differ from those of *apple*, *fruit* and *pear*. This approach is known as the **distributional method**. A theoretical justification

for this approach is the so-called *distributional hypothesis*, which states that "words that occur in similar contexts tend to have similar meanings" (Turney and Pantel, 2010, p. 143). This idea is central to our work and will be discussed in detail later in the text.

The second method is based on artificial neural networks and produces vector representations of words as a by-product of language modelling[1]. The meaning of resulting dimensions are not, however, easy to interpret and we will not discuss this method further.

1.3 Grounding Semantic Spaces in Perception

The weakness of existing semantic representation models is that they rely completely on linguistic input, although – as we mentioned earlier – there is a growing body of evidence that other modalities also contribute to forming semantic representations. This has led to attempts to enrich existing semantic spaces with perceptual information. The problem is, however, twofold: how to incorporate perceptual information and, even more importantly, how to acquire this type of data in the first place.

One possibility is to use **feature production norms**, which are lists of features that human subjects consider important or defining for given concepts. The type of information provided by feature norms is illustrated in Figure 1.1, which includes a subset of features produced for the *alligator* and *ambulance* concepts in McRae et al. (2005). Some researchers use such representations as a proxy for sensory-motor experience (Silberer and Lapata, 2012) and combine them with representations extracted from linguistic data to create **multimodal semantic spaces**.

Nevertheless, we think that feature norms are not sufficient. Although they can give us some initial insights into the workings of grounded models, they cannot be employed for large-scale projects due to their being too labour-instensive to produce; currently available feature norms have been collected only for a limited number of concepts. It is therefore important to explore other approaches that would give us the possibility to produce representations for large numbers of concepts. Following the work by Bruni et al. (2012a), we would like to present a framework that creates visual semantic spaces from thousands of images tagged with appropriate concepts. As the visual system is an important source of information about the surrounding environment, visual models could significantly help in grounding current semantic spaces in perception. Moreover, computer vision methods have undergone such a development in recent years, especially in the fields of image representation and object recognition, that renders this undertaking possible.

[1] A language model is a function that assigns a probability $P(w_1, ..., w_n)$ to a sequence $w_1 w_2 ... w_n$ of n words.

Alligator:

- has a mouth
- has a tail
- has jaws
- is green
- is long

Ambulance:

- has 4 wheels
- is fast
- is large
- is loud
- is white

Figure 1.1: An example of features produced by human subjects (McRae et al., 2005). In a typical feature norm set, each concept-feature pair is assigned a numerical value that represents the number of subjects that included this feature.

The proposed procedure is as follows. The model starts with a collection of images that are all tagged with a given concept. Each image can be represented by a set of features called *visual words*. These words constitute the **visual context** for the given concept similarly to the linguistic context discussed in the previous section. In this way, we arrive at having two sets of contexts for each concept: a set of linguitic contexts derived from text corpora and a set of visual contexts extracted from images. These two sources of information are then used to create two distinct conceptual representations, textual and visual, which can be combined to create a final representation of the given concept.

1.4 Structure of the Thesis

The remainder of the thesis is structured as follows. In the first part of Chapter 2 we give a detailed literature overview of semantic representation models derived from text corpora. These models are usually relatively simple to implement and introduced many ideas and techniques employed in later chapters. In the second part of the chapter, we discuss existing attempts at grounding conceptual representations in perception.

Chapter 3 gives an overview of computer vision techniques and presents the Visual Semantics toolkit (VSEM). This toolkit implements the whole image semantic representation pipeline and contains several benchmarks that can be used to test the quality of resulting semantic spaces.

In Chapter 4 we evaluate the performance of visual semantic spaces in modelling human semantic relatedness scores based on several parameters, including the number of visual words.

The VSEM toolkit constitutes just a first step in exploring the possibilities of semantic representation with visual and other multimodal information. In Chapter 5, we give a brief overview of current research problems and possible future applications, such as brain activity predictions and creation of supramodal representations using auto-encoders.

Finally, Chapter 6 provides a summary of the topics and problems covered in the thesis.

Chapter 2

Background

2.1 Distributional Method

The distributional method has its roots in linguistic theory. Zellig Harris proposed the distributional hypothesis according to which the meaning of words is at least in part derived form the contexts in which they appear:

> If [two words] A and B have almost identical environments except chiefly for sentences which contain both, we say they are synonyms: *oculist* and *eye doctor*. (Harris, 1954, p. 786)

However, this methodology is not limited only to the treatment of synonyms as Harris further proposes a method to quantify shades and differences of meaning:

> If A and B have some environments in common and some not (e.g. *oculist* and *lawyer*) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments. (Harris, 1954, p. 786)

Similar views were shared by other linguists, e.g. Firth (1957, p. 11) writes that "you shall know a word by the company it keeps."

The notion of semantic similarity had over time become popular among psychologists to explain various psychological phenomena. However, Miller and Charles (1991) note that it was often used without understanding the processes behind semantic similarity judgments. In *Contextual Correlates of Synonymy* (1965), Rubenstein and Goodenough present an early attempt to examine experimentally the correlation between the similarity of context and the similarity of meaning postulated by the distributional hypothesis. The authors consider as evident the fact that similar words appear in similar contexts and that, conversely, very dissimilar words appear in different contexts. They are especially interested to see, though, if the proposed quantification of the similarity of meaning holds also for words in intermediate positions,

such as the oculist-and-lawyer pair mentioned by Harris. In order to investigate this, Rubenstein and Goodenough collected human judgments of semantic similarity on the scale of 0 to 4 for 65 pairs of words which range from highly similar to semantically unrelated. For each "theme word" appearing in these pairs, the authors collected 100 sentences from which the contextual distribution of the word was extracted. This distribution was represented by the set of words which appeared in the 100 sentences and possibly satisfied some additional conditions, like being a content word or appearing close to the theme word in the parse tree of the sentence. Contextual similarity of two words was then calculated as the overlap of their sets of context words. The authors present two conclusions: (1) There is a positive relationship between the similarity of context and the similarity of meaning. (2) It is safe to infer that two words are highly similar in meaning if their contexts are highly similar. However, this does not seem to hold for words of medium or low semantic similarity as they differ relatively little in overlap. Nevertheless, Rubenstein and Goodenough (1965) speculate that the second conclusion could have been affected by the experimental setup and should therefore be further investigated.

It took more than two decades before Miller and Charles tried to address the issue again in a paper titled *Contextual Correlates of Semantic Similarity* (1991). In their treatment, they redefined the way context similarity was measured and showed that it was possible to infer semantic similarity of words even for words in the middle positions. Miller and Charles proposed that semantic similarity of a pair of words is a function of the contexts in which the words occur. They defined the *conceptual representation* of a word to be the knowledge of how the word is used in language. This representation includes syntactic, semantic, pragmatic and stylistic constraints and is learned from linguistic contexts. The representation, however, is more than just a set of contexts:

> [A] word's contextual representation is not itself a linguistic context, but is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts. (Miller and Charles, 1991, p. 5)

Inspired by the distributional hypothesis of Zellig Harris, the authors stated a closely related *contextual hypothesis*:

> The similarity of contextual representations of two words contributes to the semantic similarity of those words[1]. (Miller and Charles, 1991, p. 9)

Miller and Charles claim that the limited results of Rubenstein and Goodenough (1965) are due to the measure which does not fully capture the similarity of contexts and propose an approach based on substitutability. Sentences containing both words

[1]This is a relaxed form of the *strong contextual hypothesis*, which equates semantic similarity with the similarity of contextual representations.

are collected with the two theme words removed and subjects are asked to determine which contexts are acceptable for each word. Semantic similarity is then directly proportional to the number of shared contexts. The obtained results were consistent with the contextual hypothesis and presented stronger evidence for it than co-occurrence counts. It is important to note, though, that this in itself does not present a case against the use of co-occurrence counts, as the corpus employed by Rubenstein and Goodenough (1965) was rather limited.

Together, these two experiments were important in providing psychological support for the distributional hypothesis. Nevertheless, the full strength of the methodology became fully visible only with the implementation of first large scale models. These implementations were based on the vector space model originally developed in information retrieval. This is not surprising, since the goal of information retrieval systems is to order documents in a collection according to their relevance to a given user query, which inherently requires the systems to have some notion of word and document meaning. Both the query and the documents are represented in the same vector space using the **bag-of-words method** (Manning et al., 2008). Using bag of words, a document (or a query) is regarded as an unordered collection of words in which only the frequency of words matters. Therefore, this method completely ignores word order and grammar.

A simple information retrieval system can be created as follows. The document collection is represented by a matrix D with as many rows as there are vocabulary terms and as many columns as there are documents. The entry $D_{i,j}$ represents the number of occurrences of term i in document j. Such a matrix is called the **term-document matrix**. In an equal fashion, q_i is the frequency of term i in the query $\mathbf{q}$. The similarity between the query and the document $\mathbf{D_{*,i}}$, i.e. the i-th column of the collection matrix, is calculated as the cosine of the vectors in the hyperspace:

$$sim(\mathbf{D_{*,i}}, \mathbf{q}) = \frac{\mathbf{D_{*,i}} \cdot \mathbf{q}}{\|\mathbf{D_{*,i}}\| \cdot \|\mathbf{q}\|} \tag{2.1}$$

The more similar the document is to the query, the smaller is the angle between the representation vectors of the query and the document. The cosine is 1 for vectors "pointing" in the same direction and 0 for orthogonal vectors. The same matrix can be also used to measure similarity between documents. If two documents have a similar topic, they contain similar words and their representation vectors (columns in the matrix D) are similar.

When comparing two documents under this model, all words are given the same weight. However, it is clear that less frequent words like "beekeeping" or "robotics" are more informative about the topic of a document than words such as "is", "the" or "give". We can address this problem in two ways. A simple solution is to create a list of stop words to remove from the bag-of-words representations. Nevertheless, this is not an optimal solution, because the words that are left still vary in their

usefulness. A better solution is to employ a weighting scheme, which would adjust the collected co-occurrence counts. One such scheme is the Term Frequency-Inverse Document Frequency (TF-IDF) model (Manning et al., 2008). It adjusts the term frequency value $D_{i,j}$ by multiplying it with inverse document frequency, which is a measure of how common the term i is across all documents. Using TF-IDF, rare terms are weighted up and their contribution to the final similarity score increases, whereas very frequent terms are dampened.

So far, the focus has been put on documents. However, we can move from looking at columns of the term-document matrix to looking at its rows. Each row can be, in fact, considered a distributional representation of the corresponding vocabulary term, with columns – documents – being context items in the sense of the distributional hypothesis. Moreover, the representation using the $t \times d$ term-document matrix is probably redundant and the distribution of a word in a collection of documents could be approximated by a smaller set of *latent factors*. Typically, the number of columns is reduced from the order of thousands of documents to several hundred that represent the most important latent factors. This technique was first introduced in Deerwester et al. (1990) as part of the *latent semantic indexing* model for information retrieval. The authors employed the **Singular Value Decomposition** (SVD) method to discover the latent factors. Using SVD, a $t \times d$ matrix M can be decomposed into:

$$M = TSD^T,$$

where T is a $t \times r$ matrix providing a representation for terms, S is a $r \times r$ diagonal matrix with the singular values of M sorted from largest to smallest, D is a $d \times r$ matrix of document representations, and r is the rank of the matrix M. Note that the representation spaces of documents and terms have both r dimensions. Setting the smallest singular values to 0, we can reduce the dimensionality of the matrix, leaving only those dimensions that account for most of the variation in the original dataset. Dimensionality reduction has also the added advantage of reducing noise.

Drawing on presumed similarities between information retrieval in external systems and in the human mind, Landauer and Dumais (1997) interpreted the techniques of latent semantic indexing from a cognitive point of view and employed them to model vocabulary learning in children. Their **latent semantic analysis** (LSA) model – based on a word-paragraph matrix – performed just as good as second-language English speakers on a multiple-choice synonym test and closely modelled the estimated improvement rate per paragraph of seen text for schoolchildren. Landauer and Dumais offer two interpretations of these results: (1) Most of the knowledge needed to answer vocabulary tests is derivable from co-occurrence statistics. (2) Inner workings of the LSA model are analogous to the mechanisms through which humans acquire knowledge.

Singular value decomposition is central to LSA and its goal is not just the reduction of noise and computational complexity. Landauer and Dumais give the following

example. Suppose a communicator generates text by selecting words that are near each other in his own high-dimensional semantic space. We as receivers create estimates of distances between words by using word-paragraph co-occurrence statistics. Although these are only rough estimates, they can be improved if we start modelling the words in a space of the same or similar dimensionality as the one of the communicator. By fitting the representations to a space with less dimensions, we make a more effective use of the constraints available in the text. Landauer and Dumais have for example shown that presenting extra text to the model improves indirectly the representation of words not present in this new text. The reduced representation could also be crucial for producing good similarity estimates among pairs of words that have never been seen together. Through a series of experiments on the synonym test, Landauer and Dumais (1997) have shown that the model performs best when the number of dimensions is in the low hundreds.

So far, the meaning of words has been modelled by word-document or word-paragraph matrices. However, the context could be further reduced to single words. Schütze (1992) describes a word-word matrix model in which the distribution of a word is represented by the frequency of its co-occurrence with other words in window of a certain size. The co-occurrence counts are collected by moving this window over a text corpus. Dimensionality reduction through SVD can be applied to this representation too, although Schütze believes that its use is driven mostly by practical memory usage considerations rather than by the need to improve or smooth the representations themselves. A similar methodology is used by Lund and Burgess in their Hyperspace Analogue to Language (HAL) model (Lund and Burgess, 1996). By analysing the high-dimensional neighbourhoods of words and clustering their representations, the authors show that the co-occurrence procedure is successful in extracting general semantic information from text and that the distance in the semantic space correlates with reaction times from lexical priming studies.

Many new ways to build semantic models have been proposed in recent years. For example, neither HAL nor LSA make explicit use of linguistic information. This issue was only addressed later by models that tried to filter or link context words by syntactic relations (Padó and Lapata, 2003; Baroni and Lenci, 2010). Moreover, there have been attempts to redefine semantic representation in probabilistic terms. Griffiths et al. (2007) describe one such probabilistic approach based on the Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003). In **topic models**, each document in a corpus is represented as a probability distribution over a set of topics. The topics are, in turn, defined as distributions over words. The difference between topics resides in different probabilities that they assign to words. A topic that we could label as "finance" would give high probability to words like *bank*, *money*, and *interest*, whereas a "beekeeping" topic would prefer *bees*, *honey*, and *hive*. A semantic representation of a word in this model is defined as the distribution of the word over these abstract topics. The topics are automatically induced from co-occurrence matrices and their number can be varied, giving rise to a different way of achieving

Feature norms:	Distributional models:
• have stripes	• live in jungle
• have teeth	• can kill
• are black	• risk extinction

Figure 2.1: Difference between human-elicited feature norms and information extracted from textual distributional models (Baroni et al., 2010).

dimensionality reduction.

2.2 Embodied Models of Meaning

Semantic models based on the distributional method were successfully employed in many areas, ranging from applications in artificial intelligence to modelling psychological phenomena, and have therefore shown to be useful sources for inferring semantic representations. Their success is all the more surprising given that the only input they require is a large collection of text. Nevertheless, their cognitive plausibility has been questioned since the resulting representations are based purely on linguistic data and are therefore disembodied[2]. This issue can be illustrated by the difference between how humans describe concepts and what kind of information is provided by distributional models (Figure 2.1). Distributional models emphasize abstract and encyclopedic information, whereas human-produced feature norms tend to give higher importance to grounded, sensory-motor properties of objects (Andrews et al., 2009; Baroni and Lenci, 2008).

Andrews et al. (2009) claim that knowledge acquired from language has to relate to the world around us in order to be pragmatically useful. Moreover, numerous neuroimaging studies show activations in sensory-motor cortices during language processing (Binder and Desai, 2011), which raises the question of whether these systems are involved in semantic representations and if yes, then to what extent. The activations could be explained by simulations of modal states captured during perception, action, and introspection, on which **embodied theories of cognition** base meaning and cognition in general (Gallese and Lakoff, 2005; Barsalou, 2008). This view is, however, criticized by Mahon and Caramazza (2008), who claim that the activations in sensory-motor areas can be accounted for by disembodied theories too. In addition, embodied theories are often challenged on the representation of abstract

[2]For that reason, distributional semantic models can be criticized on the same grounds as pure symbolic systems, see for example Harnad (1990).

concepts that are not easily connected to perceptual experience.

Binder and Desai (2011) and Meteyard et al. (2012) reviewed the current debate and evidence from neuroscientific and neuropsychological studies and both came to similar conclusions: (1) Neither disembodied nor fully embodied accounts of meaning are supported by existing data. (2) Semantic memory consists of both modality-specific and supramodal representations, the latter created in high level convergence zones which combine information from multiple modalities. These findings and the evidence that linguistic and perceptual data are interdependent (Louwerse, 2008; Riordan and Jones, 2011; Andrews et al., in press) have led to research that aims at integrating both data sources in single semantic models. An overview of this research area is given in the next section.

2.3 Models integrating distributional and embodied accounts of meaning

Andrews et al. (2009) acknowledge the importance of both *experiential data*, derived from our perception and interaction with the world and extralinguistic in nature, and intralinguistic *distributional data*, representing the statistical distribution of words across spoken and written language, for learning semantic representations. The authors further point out that learning from both experiential and distributional data is more than just a simple combination of two independent representations as it allows the discovery of correlations between the two data sources and speculate that the final representations in human brain are based on these intercorrelations. Using speaker-generated feature norms as experiential data, Andrews et al. learn topic models first on the two data sources independently and then jointly over both experiential and distributional data. The joint model learns couplings of feature clusters, i.e. distributions over features akin to topics in topic models, with discourse topics extracted from a text corpus. Having evaluated the models on a series of tasks, the authors conclude that semantic representations learned from multiple data sources in combination are more realistic than those obtained from either source alone.

Johns and Jones (2012) focus on exploiting the redundancies between language and perception and present a model capable of bootstrapping perceptual representations from a small set of seed representations. The model relies on two components. The first component is a distributional semantic model derived from text. Perceptual representations, in this case feature norms from McRae et al. (2005), form the second component. A complete representation of a concept is given by concatenating these two representations. However, as the feature norms are available only for a limited set of concepts, most of the concepts do not have their perceptual representation assigned. The proposed method induces these representations through an idealized perceptual simulation based on past experiences with other, related concepts (in this case the seed set of concepts). For example, the perceptual representation for a *table* can be

induced as a weighted combination of representations for objects such as a *chair*, *wood*, and a *room*. The weights are given by semantic similarity scores between *table* and the individual terms based on the distributional semantic model. The induction step is repeated twice and all concepts in the model have their perceptual representations assigned. After studying the effect of lexicon and corpus size, the authors demonstrate that the method is successful in inducing meaningful representations and evaluate them on tasks such as priming modelling and semantic similarity.

Silberer and Lapata (2012) similarly use feature norms from McRae et al. (2005) as proxies to sensory-motor information and examine three different methods of combining them with a distributional semantic model extracted from the British National Corpus (BNC). The first approach based on topic models is identical to the one proposed in Andrews et al. (2009). The second model, already used in Johns and Jones (2012), is a simple concatenation of feature norm vectors and distributional representations. The third model combines the two modalities using Canonical Correlation Analysis (CCA) (Hardoon et al., 2004). CCA is a statistical technique used in this case for dimensionality reduction across two semantic spaces – perceptual and textual. It exploits the linear relationship between two different representations of the same concept. All models were evaluated on word association, semantic similarity, and perceptual inference tasks in order to answer the following three questions: (1) Are multimodal semantic spaces better at modelling behavioural data? (2) What is the best technique to combine textual and perceptual information? (3) Is it possible to infer perceptual features of concepts for which they are missing? The results show the following: All three models show performance gains when both modalities are included as opposed to only one. The model of Andrews et al. (2009) performed best at semantic similarity tasks, followed by the CCA model. The model of Johns and Jones (2012), on the other hand, was the best at inducing perceptual features.

The issues with using speaker-generated feature norms have already been alluded to in the introductory chapter. Apart from the fact that they are generally available only for small numbers of concepts, it is also questionable whether they always represent our everyday sensory-motor experience. Some feature norm sets combine perceptual information with other kinds of knowledge, such as the feature *requires_landlord* for the apartment concept in the popular set of McRae et al. (2005) (Figure 2.2). Moreover, the features might be skewed towards certain kinds of experience which are easy to recall for human subjects that are generating them. It is therefore important to explore other and more direct methods of acquiring perceptual information. The models discussed below employ computer vision methods to generate representations of images in terms of visual words[3] and use these representations as a surrogate for visual information.

Feng and Lapata (2010) induce topic models jointly over visual and textual words

[3]In this approach, a common vocabulary of "visual words" is extracted from an image collection and these words are subsequently used to create discrete image representations. They will be discussed in more detail in Chapter 3.

Apartment:

- found in high rises
- has elevators
- is furnished
- requires a landlord
- has tenants

Tangerine:

- has vitamin C
- grows in Florida
- grows on trees
- is citrus
- grows in warm climates

Figure 2.2: Many features included in the feature norm set McRae et al. (2005) are not directly related to perceptual experience.

extracted from a dataset of articles paired with images illustrating some parts of its content. The underlying assumption is that there is a shared set of topics that generated both the images and the text. The model is evaluated on two tasks: semantic similarity and word associations, with the similarity between two words measured by the extent to which they share the same topics. The model that incorporated both text and images outperformed the one using only text.

Bruni et al. (2011, 2012a,b) enrich distributional semantic models based solely on text with visual features from labelled images. The authors first collect co-occurrence counts of concepts with visual words across a large image collection and create visual semantic models. Subsequently, they show that visual and text-based models contain complementary information and discuss different ways to combine visual and text-based models. Their combined semantic model is evaluated on several tasks and outperforms a pure textual model on some of them, e.g. on semantic similarity. This approach served as the basis for the VSEM toolkit and will be described in more detail in the following chapter.

Chapter 3

VSEM: An open library for visual semantics representation

3.1 Introduction

In this chapter, we will describe a typical computer vision pipeline for representing images. We will also show how it is implemented in the VSEM toolkit. In this method, an image is regarded as a document and described by general features kept in a dictionary. Therefore, just as a document can be described by the bag-of-words method, an image can be described by a bag of visual words.

The basic pipeline of image representation is as follows. Firrst, interesting local patches of an image are found. These are subsequently described by descriptor vectors and mapped to their respective visual words from a pre-made visual dictionary. A visual word can be regarded as a cluster of similar descriptor vectors. In this fashion, the whole image can be described by a visual word histogram. To arrive at a concept representation, histograms of images tagged with the same concept are aggregated.

3.2 Pipeline for visual representation

The computer vision pipeline for extracting representations of images can be divided into two main steps: (1) vocabulary creation, and (2) image representation. First, a common vocabulary of visual words is created by clustering lower level image features from a training set. Having created the vocabulary, the system can proceed to representing images in terms of bag-of-visual-words histograms using the following steps:

- extraction of local image features,
- mapping of local features to higher-level visual words contained in the vocabulary,

- creation of bag-of-visual-words histograms, based on the mapping obtained in the previous step, and
- spatial binning.

This pipeline is illustrated in Figure 3.1.

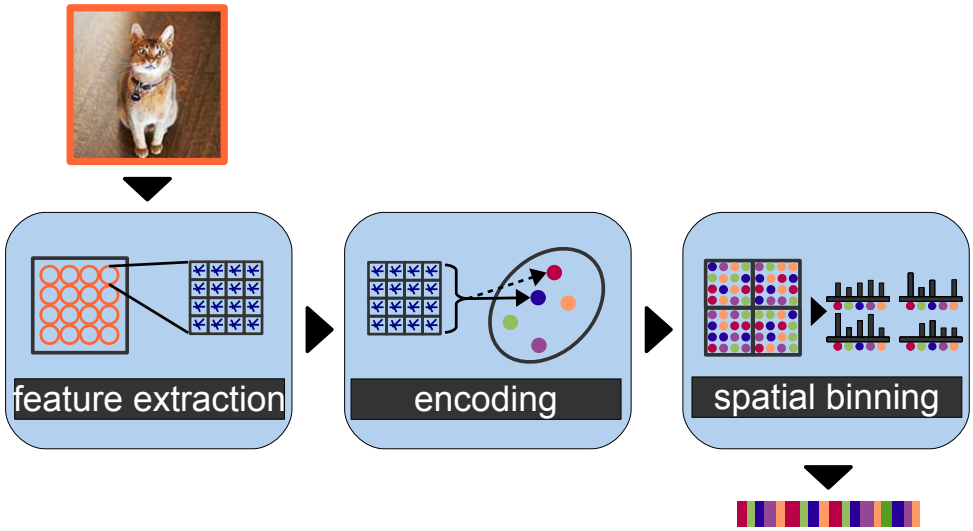

Figure 3.1: An example of a bag-of-visual-words image representation pipeline. First, local features are extracted from an image. Each is mapped to a visual word from a predefined vocabulary. Spatial binning is performed as the last step (Grauman and Leibe, 2011).

However, we are not interested only in producing representations of images. Our ultimate goal is to create visual representation of concepts. Therefore, we need at least one more step:

- aggregation of visual words on a per-concept basis in order to obtain the co-occurrence counts for each concept.

At this point, we have arrived at conceptual representations that are visual analogues to semantic models extracted from text. Just as with the textual co-occurrence models, we can refine the counts by applying the following two steps:

- transformation of counts into association scores, and
- dimensionality reduction.

3.2.1 Feature Extraction

Local image features are low-level features that encode information from small, representative regions of images, also known as "keypoints". It is important for the process of extracting and encoding this information to be invariant to common image transformations such as translation, rotation and scaling (Grauman and Leibe, 2011) so that similar information is extracted from images of the same object under different conditions. These invariant features are important for image matching. VSEM uses an implementation of the widely-used Scale Invariant Feature Transform method (SIFT) (Lowe, 2004), which transforms an image into a set of 128-dimensional vectors called **descriptors**.

3.2.2 Creating a Vocabulary of Visual Words

In this step, the set of descriptors from all images is clustered using the standard k-means clustering algorithm into k clusters[1] (represented by coloured dots in Figure 3.2). Each cluster is regarded as a distinct **visual word** and represents a set of similar visual features encountered across all images in the dataset.

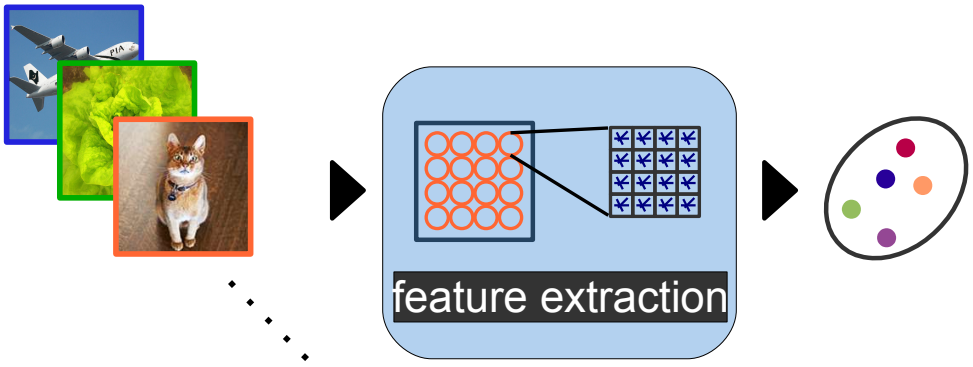

Figure 3.2: A common vocabulary of visual words is created by clustering continuous SIFT descriptors extracted from a collection of images.

3.2.3 Encoding

At this point of the pipeline, we have the following two components: (1) a set of descriptors for each image, and (2) a common vocabulary of visual words. During the encoding step, each descriptor is mapped onto a visual word and the image is represented by a bag-of-visual-words (BoVW) feature vector, where each feature

[1]Generally, only a random subset of descriptors is used by the clustering algorithm due to memory considerations. VSEM uses a random subset of 1 million descriptors.

represents the number of descriptors that were mapped to the respective visual word. We are therefore making the transition from a continuous representation of an image using SIFT descriptors to a discrete representation with visual words.

The most common encoding strategies are: (1) *hard quantization*, which maps a descriptor to a cluster whose centroid is closest, measured by Euclidean distance, and (2) *Fisher encoding*, which exploits the average first and second order differences between the SIFT descriptors and the centres of a pre-trained Gaussian mixture model (Perronnin et al., 2010; Chatfield et al., 2011).

3.2.4 Spatial Binning and Localization

The bag-of-words model by itself does not contain any information regarding the position of individual words, both in texts and in images. Spatial information can be introduced into the model by using the technique called spatial binning (Lazebnik et al., 2006), during which an image is divided into several regions and the encoding step is done for each region separately. As a result, every region is represented by its own feature vector and these are then concatenated to create the so called **spatial histogram**.

Furthermore, it is possible to map descriptors from keypoints on the surface of the object separately from those of the background, in case this annotation is available. The system then produces two representations, one of the object and one of the background, which can be combined or used separately in the later stages.

3.2.5 Aggregation

At this point, we have bag-of-words representations for all images. As each concept is represented by several images, the next step is to pool individual image representations to create a single conceptual representation, as illustrated in Figure 3.3. In VSEM, this is achieved by summing up the individual images vectors.

3.2.6 Transformations

Just as in semantic models created from text corpora, some visual words are more informative about concepts than others. In order to distinguish between interesting co-occurrences from those that are due to chance, we can employ various weighting schemes to adjust the raw co-occurrence counts. VSEM implements two types of mutual information association score: local (LMI) and pointwise (PMI) (Evert, 2005). The dimensionality of the final semantic space can be reduced using singular value decomposition.

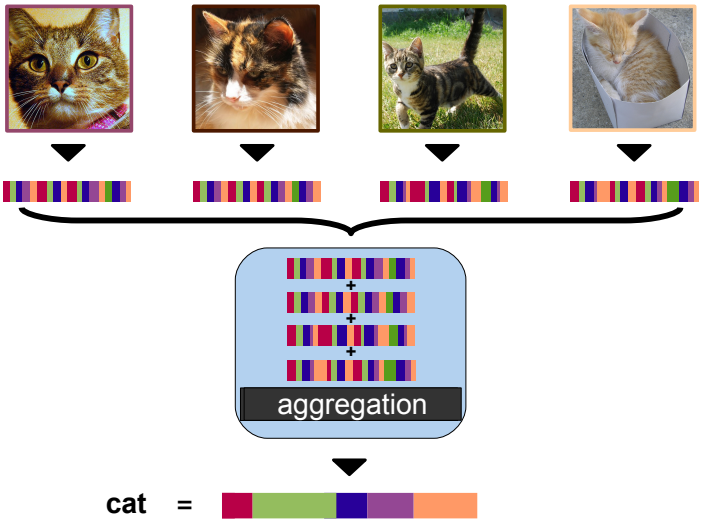

Figure 3.3: To create a single representation, feature vectors from all images tagged with a given concept are pooled (Bruni et al., 2013).

3.3 Implementation

VSEM is implemented in Matlab and is divided into four main packages. Their description follows.

`dataset` The dataset package implements several standard dataset formats. It loads an image collection from its location and checks for its consistency with the given format.

`vision` This package implements the whole visual pipeline described in the previous section. It has three sub-packages, `features`, `vocabulary`, and `histograms`. The extraction of visual features is built upon the state-of-the-art VLFeat toolkit (Vedaldi and Fulkerson, 2010).

`concepts` This package contains the semantic space functionality, including aggregation of feature vectors from individual images and matrix transformations.

`benchmarks` Contains various evaluation benchmarks that are described in Chapter 4.

The toolkit comes with demos that illustrate individual steps. We will now describe an example implementation of the full pipeline. The following three variables will be used throughout the pipeline:

- `imagePaths` – paths to all images in the dataset,
- `annotations` – image labels and possibly extra annotation, such as object localization,
- `conceptList` – list of all objects.

We can either populate them on our own or use the functionalities of the `VsemDataset` object, which can import several standard dataset formats:

```
dataset = datasets.VsemDataset(imagesPath, 'annotationFolder',...
    annotationPath);
```

The dataset object has now prepared the whole dataset and we can extract the three variables that are needed in the rest of the pipeline:

```
annotatedImages = dataset.getAnnotatedImages();
imagePaths = annotatedImages.imageData(:,1);
annotations = annotatedImages.imageData(:,2);
conceptList = annotatedImages.conceptList;
clear annotatedImages;
```

The next step is the vocabulary creation. The following lines of code cluster the image descriptors using k-means clustering algorithm and return a vocabulary that is required for the hard quantization encoding method:

```
featureExtractor = vision.features.PhowFeatureExtractor();
KmeansVocabulary = vision.vocabulary.KmeansVocabulary('voc_size',...
    vocabularySize);
vocabulary = KmeansVocabulary.trainVocabulary(imagePaths,...
    featureExtractor);
```

Extracting a vocabulary for Fisher encoding is analogous:

```
featureExtractor = vision.features.PhowFeatureExtractor();
GMMVocabulary = vision.vocabulary.GMMVocabulary('voc_size',...
    vocabularySize);
vocabulary = GMMVocabulary.trainVocabulary(imagePaths,...
    featureExtractor);
```

The configuration of the encoding step and of spatial binning is governed by the histogram extractor object:

```
histogramExtractor = ...
    vision.histograms.bovwhistograms.VsemHistogramExtractor(...
    featureExtractor, vocabulary, 'localization', localization,...
    'quad_divs', squareDivisions, 'horiz_divs', horizontalDivisions);
```

The extraction of image representations and their aggregation into a single semantic space is achieved with the following lines:

```
conceptExtractor = concepts.extractor.VsemConceptsExtractor();
conceptSpace = conceptExtractor.extractConcepts(histogramExtractor, ...
    imagePaths, annotations, conceptList);
```

The raw co-occurrence counts can be tranformed using local mutual information:

```
conceptSpace = conceptSpace.reweight();
```

The last step is to assess the quality of the final semantic model against a standard semantic relatedness benchmark:

```
[score, pValue] = benchmarks.runBenchmark(conceptSpace, 'menFull');
```

Further details are given in Appendix 1 and on the website of the toolkit:

`http://clic.cimec.unitn.it/vsem/`

Chapter 4

Evaluation

4.1 Tasks

Semantic models created by a visual pipeline similar to that of VSEM have already been evaluated previously in Bruni et al. (2011) and Bruni et al. (2012a). The authors primarily focused on producing and evaluating multimodal semantic spaces that combined models extracted separately from text and images. Bruni et al. (2011) test the multimodal model on semantic relatedness and concept categorization tasks and show that enhancing distributional semantic models with features from images leads to interesting qualitative differences in performance. However, they underline that their results should be considered as a proof of concept only and further experimenting is needed. In Bruni et al. (2012a), the authors extend the set of tests by two more tasks: (1) object colour guessing, and (2) distinguishing between literal and non-literal uses of colour terms. Their results show that distributional semantic models based on text can be outperformed by models extracted from images on tasks in which visual information is important.

In our evaluation, we would like to focus on models extracted from images only in order to investigate the following questions:

- How does the performance of visual semantic models on a semantic relatedness task depend on the selected number of visual words?
- What do the concept neighbourhoods look like?
- How do label co-occurrences affect these results?

10 most common tags	white, black, blue, man, red, woman, green, hair, girl, gray
10 of the least common tags	aarrgghh, ability, acknowledge, botox, heartburn, leaflet, narration, peperoni, perception, zoidberg

Table 4.1: Examples of the most and the least common tags in the ESP Dataset.

1,440,550	total number of label-image associations
100,000	total number of pictures
20,515	number of labels
9,455	number of labels used three times or more
8,512	number of labels used once only
2,548	number of labels used twice
70.22	average number of images per label
42	maximum number of labels per image
14.41	average number of labels per image
5	minimum number of labels per image

Table 4.2: Statistics of the ESP Dataset

4.2 Experimental Setup

4.2.1 Image dataset

For our experiments, we will use the ESP Game Dataset[1]. It is a large collection of images with English labels that were collected through the ESP Game (von Ahn and Dabbish, 2004), in which two human players partnered online simultaneously suggest labels for a randomly selected image and are required to rapidly agree on a common label (some examples of the produced labels are given in Table 4.1). When and if both players suggest the same label, it is added to the set of labels for that image. This is an effective and simple method for labelling large amounts of images and has proved to produce reliable labels, as further attested by our experiments. Basic statistics of the ESP Dataset are given in Table 4.2.

We have selected this dataset for three main reasons: (1) It has been used before and therefore it simplifies comparison between systems. (2) It covers a wide range of images and concepts. (3) It is publicly available. One of its main drawbacks, however, is that the quality of pictures is sometimes low and the pictures often aren't typical representatives of their labels.

We have reduced the total number of concept labels in ESP to 1236 and use the total of 99,971 images.

[1]The dataset is available for download at http://www.cs.cmu.edu/~biglou/resources/

Word pair		Score
type	kind	8.97
president	medal	3.00
war	troops	8.13
stock	CD	1.31
physics	chemistry	7.35
bishop	rabbi	6.69
drink	ear	1.31
drink	mouth	5.96
drink	eat	6.87

Table 4.3: Examples of word pairs and their similarity judgments from the WordSim353 dataset.

4.2.2 Visual pipeline

We will use the VSEM toolkit for our experiments with the following settings:

- Descriptors: SIFT descriptors with the gray colourscale settings.
- Dictionary: k-means dictionary.
- Encoding: Hard quantization.
- Spatial binning: 2 square divisions, 3 horizontal divisions, giving rise to a feature vector eight times the size of the number of visual words.

4.2.3 Benchmarks

In order to assess the quality of semantic models created from images, we will examine the correlation of semantic relatedness estimates produced by the model (using the cosine similarity measure as described in Chapter 2) with human-assigned similarity judgments. Specifically, we will use the WordSimilarity-353 Test Collection[2] (Finkelstein et al., 2002) and the MEN Test Collection[3] (Bruni et al., 2012a), both of which are included in the VSEM toolkit:

> **WordSim353** is a collection of 353 word pairs with similarity scores assigned by 29 subjects. The relatedness of words was estimated on a scale from 0 (completely unrelated) to 10 (very much related or identical). The collection includes all 30 word pairs from the work by Miller and Charles (1991) discussed in Chapter 2. WordSim353 is a common collection and has been used by other

[2]http://www.cs.technion.ac.il/∼gabr/resources/data/wordsim353/wordsim353.html
[3]http://clic.cimec.unitn.it/∼elia.bruni/MEN.html

Word pair		Score
metro	train	41
country	work	19
happy	post	14
ceramic	ocean	8
cute	dirty	11
frozen	shop	13
paint	sing	7
airplane	guitar	7
dessert	wine	23

Table 4.4: Examples of word pairs and their relatedness judgments from the MEN dataset.

multimodal semantic models, for instance in Silberer and Lapata (2012) and Bruni et al. (2012a). One of its drawbacks is that it consists mostly of noun-noun word pairs. Examples of some of the pairs and their similarity judgments are shown in Table 4.3.

MEN was introduced in Bruni et al. (2012a) and consists of 3,000 word pairs, randomly selected from words that occur at least 700 times in the ukWaC and Wackypedia corpora[4] and at least 50 times as tags in the ESP dataset. During the dataset creation, human subjects were presented with two pairs of words and were asked to judge which pair is more semantically correlated. Each pair was rated in this way against 50 other randomly selected comparison pairs and the scores are therefore on a 0 to 50 scale, representing the number of times the given pair was judged more similar than its comparison pair. In contrast to WordSim353, the MEN dataset contains a substantial percentage of verb-noun and noun-adjective pairs. Examples from this dataset are shown in Table 4.4.

Our dataset reaches a coverage of 270 word pairs out of 353 on WordSim353 and 2927 word pairs out of 3000 on MEN.

4.2.4 Similarity and correlation measures

In order to measure the similarity between concepts, we will use the cosine similarity measure introduced in Chapter 2. For determining the correlation between similarity estimates from semantic models and human judgments, we will employ the Spearman's rank correlation coefficient. The values of two variables X_i and Y_i, which represent the model estimates and human judgments for n word pairs in the

[4]http://wacky.sslmit.unibo.it/

No. of visual words	WordSim353		MEN	
	Counts	LMI	Counts	LMI
5	0.3027	0.2772	0.3084	0.3114
10	0.2987	0.2931	**0.3166**	0.3699
50	**0.3203**	0.2976	0.2968	0.3939
100	0.3162	0.3100	0.2741	**0.3976**
500	0.2993	0.3118	0.2315	0.3913
1,000	0.2943	0.3092	0.2163	0.3802
2,500	0.2863	**0.3134**	0.1986	0.3729
10,000	0.2704	0.3104	0.1740	0.3513
20,000	0.2601	0.3116	0.1658	0.3424

Table 4.5: The effect of the number of visual words on Spearman correlation of models on WordSim353 and MEN datasets. All scores significantly different from zero on a $p < 0.001$ level.

benchmark, are converted to ranks x_i and y_i, from which the correlation is computed as:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

4.3 Results

4.3.1 Number of visual words

The effect of the number of visual words on the Spearman correlation scores of the models on WordSim353 and MEN datasets is reported in Table 4.5. When considering only raw co-occurrence matrices, the models perform best when the number of visual words is limited to dozens. As the number of visual words grows, the matrices become sparser (also due to the fact that the image dataset is limited) and the cosine similarity measure does not seem to capture the similarities between concepts well enough any more. However, this can be partly alleviated by the use of association scores. It is clear from the results that local mutual information improves the final performance significantly, especially when considering models with large numbers of visual words (e.g. 20,000). The overall importance of applying local mutual information is rather limited in the case of the WordSim353 dataset, which is probably due to its smaller size. However, its role increases significantly for the MEN dataset, where it accounts for the significant jump in correlation from 0.3166 to 0.3976.

It can be assumed that the optimal number of visual words depends largely on the size of the image dataset, the association score used and on the type of spatial binning. In our experimental setup and considering only the larger MEN dataset, the number of visual words between 50 and 1000 produce optimal results.

Image	Labels
1	air, airplane, cloud, engine, flight, flying, grey, sky, white
2	building, city, cloud, dirt, old, panorama, road, ruins, street, window
3	art, brown, floor, paitning, picture, room, seat, wall, white, wood
4	baby, boy, boys, brother, kids, red, stripes
5	blue, face, man, movie, star

Table 4.6: Labelsets from the ESP dataset.

Label-label	Counts	0.5983
	LMI	0.7466
Label-image	Counts	0.6686
	LMI	0.6718

Table 4.7: Spearman correlation of label-label and label-image semantic models on the MEN dataset.

4.3.2 Effect of label co-occurrences

In the ESP dataset, images are usually labelled with multiple concepts. Some of these label sets are illustrated in Table 4.6. By examining the table, it is clear that the system could learn meaning similarities only by exploiting label co-occurrences. If two concepts, e.g. *lighthouse* and *tower*, co-occurred often in the image collection, their visual word feature representations would be aggregated from mostly identical images. The two concepts would therefore be judged as very similar by the cosine correlation measure. However, this would not be due to their shared visual features, but only due to their co-occurrence. We therefore wanted to: (1) examine this co-occurrence effect, and (2) try to eliminate it and examine the differences.

In order to see how much information is contained in label co-occurrences, we have created two co-occurrence matrices: label-label and label-image (akin to the HAL and LSA models of distributional meaning considering co-occurring labels as context). These models are very easy to create, as they completely bypass the visual pipeline. We then assessed these semantic models against the MEN dataset. The results are reported in Table 4.7 and they lead us to two conclusions: (1) Clearly, semantic models based on image labels are very good and interesting information can be gained by exploiting them. (2) The ESP Game and its rules have succeeded in providing label sets that are meaningful and without much noise.

Judging by the results of label co-occurrence matrices, it would be possible to criticize semantic models extracted from images for exploiting label co-occurrences more than actual visual features and that the correlations with human relatedness judgments are mostly due to this effect. In order to show the opposite, we have created a reduced label set for the ESP collection. Whenever a pair of words that

No. of visual words	WordSim353		MEN	
	Counts	LMI	Counts	LMI
5	0.2882	0.2646	0.3071	0.3107
10	0.2696	0.2485	**0.3148**	0.3681
50	**0.2923**	0.2786	0.2953	0.3934
100	0.2837	0.2938	0.2723	**0.3979**
500	0.2681	0.2984	0.2288	0.3924
1,000	0.2574	0.2948	0.2131	0.3816
2,500	0.2475	**0.2988**	0.1949	0.3744
10,000	0.2313	0.2875	0.1699	0.3527

Table 4.8: Spearman correlation of semantic models built from the ESP image collection using reduced label sets on MEN and WordSim353. All scores significantly different from zero on a $p < 0.001$ level.

was present in either WordSim353 or MEN datasets occurred together as labels for one single image, we have randomly removed one. In this way, no two directly compared words appear together as labels for one image. Using this reduced collection, we have recreated the semantic models and assessed their performance on MEN and WordSim353. The results, which are reported in Table 4.8, show that the effect of the label set reduction is not strong. Furthermore, Bruni et al. (2012a) show that semantic models from images perform just as good as or better than the two models extracted solely from labels on tasks requiring distinguishing literal and non-literal uses of colour terms.

4.3.3 Concept neighbourhoods

The resulting semantic space can also be evaluated qualitatively by examining the *neighbourhoods* of concepts. The neighbourhood of a given concept consists of concepts that are judged by the model to be most similar to it. These structures are discussed in detail in Andrews et al. (2009), where the authors show qualitative differences between semantic models from text and from feature norms. Examples of the set of nearest neighbours that our system produces is given in Table 4.9. Unfortunately, it is impossible to eliminate the effect of label co-occurrences by using the ESP collection, as the number of images per concept would have to be greatly reduced and the results would therefore become unreliable. It will be nevertheless important to examine concept neighbourhoods produced by a semantic model extracted from a large enough image collection where each image is assigned one label only.

Concept	Nearest neighbours
cheerleaders	ceremony, pepper, peacock, tribe, parade, dolls, decoration, team
skyline	castle, sky, town, tower, church, bridge, frozen, roof
bishop	sting, flamingo, chopper, space, dark, moon, glow, night
tree	rock, park, statue, sculpture, palm, lion, stone, grass
animal	green, dirty, winter, brown, bird, dead, dog, sleep, grey
bucket	basket, pot, summer, patter, nuts, painting, run, cow, fun
chisel	journey, mammals, screwdriver, titanium, profit, diamond, stich
sailing	temple, crane, lighthouse, ancient, pier, tower, castle, monument

Table 4.9: Nearest neighbours of concepts.

4.3.4 Result summary

We have evaluated the semantic models extracted from images using the VSEM toolkit on WordSim353 and MEN datasets, showing positive correlation with human relatedness judgments. Furthermore, we tested the performance of the systems in relation to the number of visual words, which is a fundamental parameter, and showed that systems with the number of visual words in the order of hundreds produce optimal results. However, this is most certainly conditioned by the number of concepts and the size of the dataset (100,000 images) and should be studied more using larger datasets. To assess the role of label co-occurrences, we produced semantic models using labels as context and showed that their performance is very high. However, we have also shown that the effect of this co-occurrence on the resulting performance of visual semantic models is rather limited.

Chapter 5

Future Work

5.1 Visual Attributes

There is a discrepancy between semantic models extracted from text, which use high level features such as words and documents as features in semantic representations, and visual models produced by VSEM, which use much lower level visual words. Silberer et al. (2013) have recently explored the possibilities of using visual attributes, such as *is round* and *has stripes*, in semantic representation. The advantage of using visual attributes is that they are higher-level features than visual words and are comparable with the features generated by humans in norming studies. Just as visual words, visual attributes were initially introduced to help with object recognition (Farhadi et al., 2009).

Silberer et al. present a large collection of 688,000 images from the ImageNet dataset (Deng et al., 2009) labelled with the same concepts as those used in McRae et al. (2005). The concepts are represented in terms of 412 attributes, examples of which are provided in Table 5.1. The authors show that these automatically derived visual attributes improve the performance of distributional models on word association tasks.

5.2 Learning Representations with Autoencoders

5.2.1 Autoencoders

Autoencoders are a type of neural networks whose input and output layers are of the same size, but their hidden layers are considerably smaller. It is possible to use autoencoders for dimensionality reduction by training them to reconstruct their input and taking the activation vector of their central layer as a low-dimensional code of the input.

Attribute Category	No. of attributes	Examples
Colour patterns	25	is red, has stripes
Diet	35	eats nuts, eats grass
Shape size	16	is small, is chubby
Parts	125	has legs, has wheels
Botany, anatomy	25, 78	has seeds, has fur
(In)animate behaviour	55	flies, waddles, pecks
Texture (material)	36	made of metal, is shiny
Structure	3	2 pieces, has pleats

Table 5.1: Visual attributes used in Silberer et al. (2013)

Hinton and Salakhutdinov (2006) successfully used deep autoencoders to produce representations for handwritten digits and documents. Their novel training procedure is based on greedy, layer-wise training with Restricted Boltzmann Machines (RBM), which is completely unsupervised, followed by fine tuning with backpropagation. The RBM is a two-layer network with stochastic, binary visible and hidden units. Every visible unit is connected to every hidden unit, but there are no interactions between units of the same layer. An overview of deep architectures is given in Bengio (2009) and Bengio et al. (2013); the effects of the unsupervised pre-training phase are examined in Erhan et al. (2010).

5.2.2 Evaluation

We have implemented the deep autoencoder learning algorithm with layer-wise RBM pre-training based on methods described in Hinton and Salakhutdinov (2006) and Hinton (2010). We then trained autoencoders of various forms, used them to reduce the dimensionality of visual semantic spaces, and evaluated these reduced spaces on WordSim353 and MEN datasets.

We illustrate the performance of dimensionality reduction using autoencoders on the following setup. A visual semantic space with 4000 dimensions was used. The raw co-occurrence counts were transformed using local mutual information and each concept representation vector was normalized to unit length. We trained a 4000-2000-1000-500-250-125 autoencoder, using batches of 4 concepts and a total of 100 learning epochs. We report the correlation scores on the MEN dataset using activations from individual layers and compare them to scores obtained by a semantic space reduced using Principal Component Analysis (PCA) in Table 5.2. Although the performance of models reduced with autoencoders is lower than those reduced wtih PCA, we believe that we have not yet reached the full potential of the method and consider these results as preliminary.

No. of dimensions	Autoencoder	PCA
2000	0.2599	0.3913
1000	0.1898	0.3913
500	0.1743	0.3916
250	0.1712	0.3919
125	0.1708	0.3918

Table 5.2: Comparison of Spearman correlation scores on MEN achieved by semantic spaces whose dimensionality was reduced using autoencoders and PCA. All results significant on the $p < 0.005$ level.

5.2.3 Supramodal Representations

The availability of concept representations coming from multiple modalities brings about the question of how to use these different representations. One possibility is to keep these representations separate and use each for tasks that they are best suited to. However, having a concept is probably more than just having separate modality-specific representations. As it has been already discussed in Chapter 2, the evidence coming from numerous neuroscientific and neuropsychological studies suggests that our semantic memory consists of both modality-specific and supramodal representations, and these supramodal representations are created in high level convergence zones which combine information from multiple modalities (Binder and Desai, 2011; Meteyard et al., 2012).

Several different approaches to multimodal fusion and therefore to the discovery of correlations between multiple modalities have already been described in literature. The simplest approach is a plain concatenation of individual feature vectors, employed for instance by Johns and Jones (2012). Bruni et al. (2012b) concatenate the two modalities and subsequently project the representations onto a space of lower dimensionality using Singular Value Decomposition. Feng and Lapata (2010) and Andrews et al. (2009) both employ topic models and learn topics that combine distributional and perceptual features. Silberer and Lapata (2012) compare fusion methods based on topic models, concatenation and canonincal correlation analysis. These systems have been discussed in more detail in Chapter 2.

Ngiam et al. (2011) proposed a novel method of learning features over both audio and visual information based on deep networks and applied it in the context of speech perception. Humans are known to integrate both speech audio and lip movements when perceiving speech because each modality provides different types of information, as exemplified for instance by the McGurk effect. However, the correlation between the two modalities does not manifest itself at the "low-level" of audio waveforms and pixels, but rather at the "mid-level" of phonemes and visemes[1]. For this reason,

[1]Visemes are the visual analogues of phonemes, defined by lip poses and motions employed when producing a speech sound. However, there is not a one-to-one mapping between phonemes

multilayer networks could be a very good fit for discovering these correlations, as they can be used first to create higher-level representations, and second to search for intercorrelations among these higher-level representations.

We are encouraged by the success of the multimodal fusion method described by Ngiam et al. (2011) and believe that a similar story of correlation at a higher level can be told for text and vision. Therefore, we would like to pursue this line of research in the future. On the other hand, due to the high number of parameters, training autoencoders and deep networks in general is known to be difficult.

5.3 Predicting Human Brain Activity

Mitchell et al. (2008) presented a computational model capable of predicting functional magnetic resonance imaging (fMRI) neural activation associated with the meaning of nouns. The idea behind the model is rather simple and consists of learning a mapping between a distributional semantic space extracted from a large corpora and a neural semantic space of fMRI activations and subsequently using this mapping to predict neural activity. The availability of neural semantic models has expanded the range of possible tasks on which automatically extracted semantic models can be tested.

Devereux et al. (2010) and Murphy et al. (2012) review several different models and obtain prediction accuracies similar to those published in Mitchell et al. (2008). However, both studies limit themselves to models extracted from text. We believe that integrating visual models could be beneficial to the prediction task as all the nouns in the set are highly imageable (compare Table 5.3) and the subjects were presented images of the nouns while in the scanner. Therefore, we would like to investigate different ways of combining both textual and visual models to improve the prediction.

As all words from the set are included in the ESP dataset, we can use the semantic models derived in Chapter 4 and test their accuracy in prediction. We reimplemented the pipeline proposed in Mitchell et al. (2008) in the following way: (1) The mapping between the visual and neural semantic space is learned through multiple linear regression. (2) Pearson correlation is used for matching predicted and actual neural activations, as this was reported to produce better results (Devereux et al., 2010). (3) Only the 500 most stable neurons are used for matching. (4) The visual semantic space is constructed with 500 visual words and then reduced using PCA to 25 dimensions. The results are reported in Table 5.4. Most of the results are not significantly different from the random baseline, except for subject 7. However, this was just a preliminary test and we would like to further investigate this area, possibly using supramodal representations produced by deep neural networks.

---

and visemes, as several phonemes might correspond to one viseme.

Category	Words
animals	bear, cat, cow, dog, horse
body parts	arm, eye, foot, hand, leg
buildings	apartment, barn, church, house, igloo
building parts	arch, chimney, closet, door, window
clothing	coat, dress, pants, shirt, skirt
furniture	bed, chair, desk, dresser, table
insects	ant, bee, beetle, butterfly, fly
kitchen utensils	bottle, cup, glass, knife, spoon
man made objects	bell, key, refrigerator, telephone, watch
tools	chisel, hammer, pliers, saw, screwdriver
vegetables	carrot, celery, corn, lettuce, tomato
vehicles	airplane, bicycle, car, train, truck

Table 5.3: The 60 nouns using in Mitchell et al. (2008)

Method	P1	P2	P3	P4	P5	P6	P7	P8	P9
Mitchell et al. 2008	0.84	0.84	0.77	0.81	0.79	0.67	0.72	0.63	0.68
Visual model	0.53	0.58	0.43	0.43	0.46	0.46	0.64	0.46	0.54

Table 5.4: Prediction accuracy for the the original model and our visual model.

Chapter 6

Conclusion

In this work, we presented a framework for extracting visual information from images associated with one or more concepts. Our goal was to ground semantic spaces extracted from text in perception. This has recently become an area of active research, driven also in part by the evidence from neuroscientific and neuropsychological studies which support the hypothesis that semantic memory consists of both modality-specific and supramodal representations. The main motivation for this work was to simplify the process of acquiring perceptual data, as we consider feature norms – the commonest method in use – to be only very indirect proxies for sensory-motor information and mostly inappropriate for use in large scale projects.

The whole process of extracting visual representations of concepts is implemented in the Visual Semantics toolbox (VSEM), which was presented in this work and made publicly available. We have evaluated the visual semantic models created by the toolkit on the MEN and WordSim353 datasets, showing positive correlation with human relatedness judgements.

The VSEM toolkit and the experiments presented here constitute only a small step in exploring the possibilities of semantic representation using visual and other multimodal information. One of the next steps should be the evaluation of visual semantic models on an image collection in which each image is assigned one label only. A different research problem is the creation of supramodal representations from individual modality-specific representations. We have presented some initial experiments using autoencoders; however, more work needs to be done in the future.

Bibliography

Mark Andrews, Gabriella Vigliocco, and David Vinson. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3): 463–498, 2009.

Mark Andrews, Stefan Frank, and Gabriella Vigliocco. Reconciling embodied and distributional accounts of meaning in language. *Topics in Cognitive Science*, in press.

Marco Baroni and Alessandro Lenci. Concepts and properties in word spaces. *Rivista di Linguistica*, 20(1):55–88, 2008.

Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, December 2010.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2): 222–254, 2010.

Lawrence W. Barsalou. Grounded cognition. *Annual Review of Psychology*, 59:617–645, 2008.

Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Jeffrey R. Binder and Rutvik H. Desai. The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11):527–536, 2011.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Elia Bruni, Giang Binh Tran, and Marco Baroni. Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 22–32. Association for Computational Linguistics, 2011.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea, July 2012a. Association for Computational Linguistics.

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1219–1228. ACM, November 2012b.

Elia Bruni, Ulisse Bordignon, Adam Liska, Jasper Uijlings, and Irina Sergienya. Vsem: An open library for visual semantics representation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of the British Machine Vision Conference*, pages 76.1–76.12, August 2011.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Miami, Florida, 2009.

Barry Devereux, Colin Kelly, and Anna Korhonen. Using fmri activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 70–78, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learnining Research*, 11:625–660, March 2010.

Stefan Evert. *The statistics of word cooccurrences*. PhD thesis, Dissertation, Stuttgart University, 2005.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785, Miami Beach, Florida, 2009.

Yansong Feng and Mirella Lapata. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99. Association for Computational Linguistics, 2010.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, January 2002.

John R. Firth. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell Publishers, Oxford, UK, 1957.

Vittorio Gallese and George Lakoff. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 22(3-4):455–479, 2005.

Kristen Grauman and Bastian Leibe. *Visual Object Recognition*. Number 11 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114(2):211–244, 2007.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Stevan Harnad. The symbol grounding problem. *Physica D*, 42(1):335–346, 1990.

Zellig S. Harris. Distributional structure. *Word*, 10(2–3):146–162, 1954. Reprinted in Harris (1970), pp. 775–794.

Zellig S. Harris. *Papers in Structural and Transformational Linguistics*. D. Reidel Publishing Company, Dordrecht, Holland, 1970.

Geoffrey E. Hinton. A practical guide to training restricted boltzmann machines. Technical Report UTML TR 2010-003, Department of Computer Science, University of Toronto, Toronto, Canada, August 2010.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

Brendan T. Johns and Michael N. Jones. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4(1):103–120, 2012.

Markus Kiefer and Friedemann Pulvermüller. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*, 48 (7):805–825, 2012.

Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

Max M. Louwerse. Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, 15(4):838–844, 2008.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28 (2):203–208, 1996.

Bradford Z. Mahon and Alfonso Caramazza. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology – Paris*, 102(1):59–70, 2008.

Bradford Z. Mahon and Alfonso Caramazza. Concepts and categories: A cognitive neuropsychological perspective. *Annual Review of Psychology*, 60:27–51, 2009.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press, Cambridge, UK, 2008.

Ken McRae and Michael Jones. Semantic memory. In Daniel Reisberg, editor, *The Oxford Handbook of Cognitive Psychology*. Oxford University Press, Oxford, 2012.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments, & Computers*, 37(4):547–559, 2005.

Lotte Meteyard, Sara Rodriguez Cuadrado, Bahador Bahrami, and Gabriella Vigliocco. Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, 48(7):788 – 804, 2012. Language and the Motor System.

George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

Brian Murphy, Partha Talukdar, and Tom Mitchell. Selecting corpus-semantic models for neurolinguistic decoding. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 114–123, Montréal, Canada, June 2012. Association for Computational Linguistics.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, Bellevue, WA, USA, June 2011.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The measurement of meaning*. University of Illinois Press, 1957.

Sebastian Padó and Mirella Lapata. Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 126–135, Sapporo, Japan, 2003.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV'10, pages 143–156, Berlin, Heidelberg, 2010. Springer-Verlag.

Brian Riordan and Michael N. Jones. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345, 2011.

Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communication of the ACM*, 8(10):627–633, October 1965.

Hinrich Schütze. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 787–796. IEEE Computer Society Press, 1992.

Carina Silberer and Mirella Lapata. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433. Association for Computational Linguistics, 2012.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

Andrea Vedaldi and Brian Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA, 2010. ACM.

Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. Association for Computing Machinery (ACM).