Technical Report

UCAM-CL-TR-899
ISSN 1476-2986

Number 899

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Deep embodiment: grounding semantics in perceptual modalities

Douwe Kiela

February 2017

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted July 2016 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Darwin College.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

Multi-modal distributional semantic models address the fact that text-based semantic models, which represent word meanings as a distribution over other words, suffer from the grounding problem. This thesis advances the field of multi-modal semantics in two directions. First, it shows that transferred convolutional neural network representations outperform the traditional bag of visual words method for obtaining visual features. It is then shown that these representations may be applied successfully to various natural language processing tasks. Second, it performs the first ever experiments with grounding in the non-visual modalities of auditory and olfactory perception using raw data. Deep learning, a natural fit for deriving grounded representations, is used to obtain the highest-quality representations compared to more traditional approaches. Multi-modal representation learning leads to improvements over language-only models in a variety of tasks. If we want to move towards human-level artificial intelligence, we will need to build multi-modal models that represent the full complexity of human meaning, including its grounding in our various perceptual modalities.

Acknowledgements

I would like to express my enormous gratitude to Stephen Clark, my supervisor, teacher and friend over the past years. His expertise, support and guidance were invaluable, and he always managed to strike the right balance between keeping my feet on the ground and letting me explore crazy ideas. Special thanks to my examiners, Anna Korhonen and Marco Baroni, for their insightful thoughts and excellent feedback on this thesis.

My collaborators, Laura Rimell, Ivan Vulić, Anita Verő and Ekaterina Shutova have been amazing to work with. I am also grateful to Léon Bottou, for being a great inspiration and mentor. Felix Hill and Luana Bulat have been instrumental to this thesis, and have kept me motivated, sane and intellectually stimulated throughout my PhD.

Enormous thanks to my friends, in particular Erik and Willy Lu in Cambridge and Edwin, Bob, Gijs, Will, Roman and Tijs elsewhere, for keeping me going. Eternal gratitude to Alice, Ted, Heleen and Peter for their caring and encouragement. Finally, this thesis would not have been possible without Kat, whose love, enthusiasm and boundless support never cease to amaze me.

Contents

Part I

Introduction

chapter 1

Introduction

Meaning has been called the "holy grail" of many scientific subjects: not only of linguistics, but also of philosophy, psychology and neuroscience (Jackendoff, 2002). Artificial Intelligence (AI) is very much a part of that list: without meaning, achieving a level of intelligence similar to humans seems impossible. Embodiment theories in cognitive science hold that human semantic representation depends on sensori-motor experience, or in other words, that human meanings are grounded in perception of the physical world. Despite this, AI in general and natural language processing (NLP) in particular, have focused mostly on tasks that involve a single modality — solely language, in the case of NLP. If we want to move towards human-level artificial intelligence, we will need to build multi-modal models that represent the full complexity of human meaning, including its grounding in perceptually rich environments.

Such theoretical considerations have given rise to the field of multi-modal semantics, which aims to construct models that can account for the fact that meaning is grounded. Grounding has been found to boost performance in various natural language processing tasks, indicating that the theoretical motivations in fact lead to practical improvements. A natural way for investigating such grounding of meaning is through neural networks, which have become popular for representing natural language and which have led to great improvements in AI in recent years due to the availability of large amounts of data and cheap computational power (LeCun et al., 2015).

In this thesis, multi-modal semantics is extended in two general directions. First, deep learning methods are used to improve visual grounding, showing that transferred features from convolutional neural networks perform much better than the existing approaches in multi-modal semantics. These transferred features are then applied to other NLP tasks to show their general applicability. Second, grounding is taken beyond visual perception, into the previously unexplored territory of auditory grounding, where we first introduce a novel approach called bag of audio words and then show that deep learning can improve on this method also. Along similar lines, a bag of chemical compounds model is introduced for achieving olfactory (smell) grounding, which is of particular interest because olfaction is the most primitive sensory modality, making smells difficult to capture in words. Broadly speaking, this thesis has two general aims: to show that deep learning yields better representations than previously used methods; and to show that grounding need not be limited to the visual modality.

1.1 Thesis outline

This thesis is structured as follows. Existing literature is reviewed in Part I. The thesis then consists of two parts that discuss the results. The first part concerns grounding in the visual modality. It is shown that visual and multi-modal representations can be improved through deep learning, showing that convolutional neural networks outperform the traditional bag of visual words approach. Various neural network architectures are explored in a systematic study of architectures and data sources, which shows that the improvements over the traditional method extend to different neural network architectures. It also shows that search engines can be used to return relevant images and that these images work as well as human annotated image resources. These novel visual representations are then shown to be of use in two important natural language processing tasks: lexical entailment and bilingual lexicon induction.

The second part takes multi-modal semantics into unexplored territory, beyond the visual modality. A similar approach to bag of visual words, called bag of audio words, is introduced. It is shown that deep learning improves representations in this case as well. A deep convolutional neural network that is similar to the one used for improving visual grounding is trained on an auditory recognition task. The transferred representations from this network outperform the bag of audio words approach. Lastly, as a proof of concept, grounding is performed in the olfactory modality through a novel approach called bag of chemical compounds. In that case, data is considerably more sparse, making a deep learning approach less feasible.

The thesis finishes with a discussion of the future of multi-modal semantics and proposes full virtual embodiment through video games as an area where a concentrated effort in multi-modal semantics may lead to large improvements on the path towards achieving general artificial intelligence. This is followed by a conclusion that summarizes the results obtained in this thesis. The appendix introduces a multi-modal feature extraction toolkit to facilitate further research in multi-modal semantics.

1.2 Published work

All experiments in this thesis have been performed by its author. Léon Bottou helped with transferring features from the original convolutional neural network; Luana Bulat assisted in annotating two datasets for perceptual relevance; Anita Verő helped with experimenting with different types of convolutional networks. Parts of this thesis have been published in the following papers:

- D. Kiela and L. Bottou (2014). Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. Proceedings of EMNLP, Doha, Qatar.

- D. Kiela, L. Rimell, I. Vulić and S. Clark (2015). Exploiting Image Generality for Lexical Entailment Detection. Proceedings of ACL, Beijing, China.

- D. Kiela, L. Bulat and S. Clark (2015). Grounding Semantics in Olfactory Perception. Proceedings of ACL, Beijing, China.

- D. Kiela and S. Clark (2015). Multi- and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception. Proceedings of EMNLP, Lisbon, Portugal.
- D. Kiela, I. Vulić and S. Clark (2015). Visual Bilingual Lexicon Induction with Transferred ConvNet Features. Proceedings of EMNLP, Lisbon, Portugal.
- D. Kiela (2016). MMFEAT: A Toolkit for Extracting Multi-Modal Features. Proceedings of ACL: System Demonstrations, Berlin, Germany.
- D. Kiela, A.L. Verő and S. Clark (2016). Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. Proceedings of EMNLP, Austin, TX.
- D. Kiela, L. Bulat, A.L. Verő and S. Clark (2016). Virtual Embodiment: A Scalable Long-term Strategy for Artificial Intelligence Research. NIPS Workshop on Machine Intelligence (MAIN), Barcelona, Spain.

In addition, several papers have been published, some as second author, that relate to the topical matter of the thesis but that are not directly included in its contents:

- D. Kiela and S. Clark (2014). A Systematic Study of Semantic Vector Space Model Parameters. Proceedings of EACL, Second Workshop on Continuous Vector Space Models and their Compositionality (CVSC), Gothenburg, Sweden.
- D. Kiela, F. Hill (joint first authors), A. Korhonen and S. Clark (2014). Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More. Proceedings of ACL, Baltimore, MA.
- D. Kiela, F. Hill and S. Clark (2015). Specializing Word Embeddings for Similarity or Relatedness. Proceedings of EMNLP, Lisbon, Portugal.
- I. Vulić, D. Kiela, S. Clark and M.F. Moens (2016). Multi-Modal Representations for Improved Bilingual Lexicon Learning. Proceedings of ACL, Berlin, Germany.
- L. Bulat, D. Kiela and S. Clark (2016). Vision and Feature Norms: Improving Automatic Feature Norm Learning through Cross-modal Maps. Proceedings of NAACL-HLT, San Diego, CA.
- E. Shutova, D. Kiela and J. Maillard (2016). Black Holes and White Rabbits: Metaphor Identification with Visual Features. Proceedings of NAACL-HLT, San Diego, CA.
- A.J. Anderson, D. Kiela, S. Clark and M. Poesio (2016). Visually Grounded and Textual Semantic Models Differentially Decode Brain Activity Associated with Concrete and Abstract Nouns. Transactions of the Association for Computational Linguistics (TACL)

CHAPTER 2

Background

Information processing in the brain can be roughly described to occur on three levels: perceptual input, conceptual representation and symbolic reasoning (Gazzaniga, 1995). Modelling the latter has a long history in AI and sprang from its "good old fashioned" roots (Haugeland, 1985), while the former has been advanced greatly through the application of pattern recognition to perceptual input (see e.g. LeCun et al., 2015). Understanding the middle level is arguably more of an open problem: how is it that perceptual input leads to conceptual representations that can be processed and reasoned with?

2.1 Distributional semantics

Much of the recent success of natural language processing and machine learning depends on improved data representation (Bengio et al., 2013). Distributional semantic models (Turney and Pantel, 2010; Erk, 2012; Clark, 2015) constitute one of the key ways in which data, in particular lexical data, is represented in natural language processing. Distributional semantics relies on the distributional hypothesis (Harris, 1954; Firth, 1957), which postulates that words that appear in similar contexts tend to have similar meanings.

Exploiting the fact that contextual information can be used to approximate word meaning has a long history in cognitive science (Miller and Charles, 1991) and computational linguistics (Manning and Schütze, 1999). Traditional models construct a vector space by counting co-occurrences between target words and their contexts. The criterion for determining co-occurrences varies greatly for different methods, ranging from occurrence in the same document (Landauer and Dumais, 1997) to occurrence in the same window (Lund and Burgess, 1996) to occurrence along a number of arcs in a dependency graph (Padó and Lapata, 2007). Every word denotes a point in the vector space and, following the distributional hypothesis, points that are close to each other in the space have a similar meaning. Such models have become known as vector space models.

Distributional vector space models have been successfully applied to many important problems in artificial intelligence having to do with language, including information retrieval (Salton et al., 1975), text classification (see Sebastiani, 2002), question answering (Tellex et al., 2003), information extraction (Paşca et al., 2006), semantic role labelling (Pennacchiotti et al., 2008), word sense discrimination (Schütze, 1998), word sense disambiguation (Padó and Lapata, 2007), word clustering and thesaurus construction (Grefenstette, 1994; Lin, 1998), metaphor detection (Shutova et al., 2012), selection preference

modelling (Erk, 2007), bilingual lexicon induction (Rapp, 1995), phrasal similarity and composition (Mitchell and Lapata, 2010), compositionality detection (Baldwin et al., 2003), lexical entailment (Weeds et al., 2004) and even in the neurosciences (Mitchell et al., 2008; Murphy et al., 2012).

Neural language models (Bengio et al., 2006) have become an alternative to vector space models as a method for obtaining semantic representations. It has been argued that neural language models are fundamentally different from vector space models: instead of explicitly counting co-occurrences, they implicitly represent co-occurrence information. That is, they treat the problem of representation as a supervised learning[1] task (Baroni et al., 2014). Employing this method yields distributed representations (Hinton et al., 1986), often called "embeddings". The approach has been differentiated from distributional semantics by calling it *distributed* semantics instead (e.g. Hermann and Blunsom, 2014).

Although a practical distinction can thus be made between "count" and "predict" models (Baroni et al., 2014), the boundaries between distributional and distributed semantics are not as clear-cut as one might think. Analogies between vector space models and embedding methods have been found, showing that many embedding methods are in fact implicit approximations of matrix factorization over weighted matrices (Levy and Goldberg, 2014; Pennington et al., 2014). It has been argued that much of the improved performance of embedding methods stems from the choice of hyperparameters (Levy et al., 2015), rather than some intrinsic property of supervised distributional models that makes them better than traditional distributional methods. In fact, they both learn similar things and rely on the same underlying assumption: the distributional hypothesis. This raises the question whether we should explicitly distinguish between the two types of models, since they are both distributional in essence. Currently popular distributed methods do seem to have at least one advantage, in that predicting contexts is computationally more efficient and requires less memory, since counting and factorizing are done in the same step.

In what follows, the two types of models and their parameters are discussed in more detail, insofar as is relevant to this thesis.

2.1.1 Vector space models

It has long been known that raw co-occurrence counts do not work well for constructing vector spaces (Baroni et al., 2014). The past decades of research in distributional methods have explored many methods for improving raw vector space models. These improvements generally address one or more of several related problems that raw co-occurrence count vector space models are known to suffer from:

- Frequency effects: If a word w only occurs once, it does not constitute an informative context for distinguishing between words; likewise, if a word w occurs disproportionately often but is not informative (e.g. *the*), it will dominate the vector space, which may be detrimental to representation quality.

[1]Supervised learning, in this case, simply means that the task involves predicting a label—in this case the next word or a context word. Confusingly, typical neural language models are sometimes called unsupervised because they do not require any human annotation and can be learned from corpora without any additional annotation.

- Noise: Large corpora are used as approximations of the "true distribution" of words, but (obviously) never perfectly capture that true distribution, which introduces noise.

- Sparsity: Words in a language tend to be distributed according to Zipf's law (Baayen and Lieber, 1996), which implies that relatively few words are used very often, while most are used only rarely. This leads to sparsity, since many words will not co-occur with each other often, if at all.

There have been several studies that closely examine the parameters for optimal vector space model performance (e.g. Bullinaria and Levy, 2007; Baroni and Lenci, 2010; Bullinaria and Levy, 2012; Kiela and Clark, 2014; Lapesa and Evert, 2014). Often the first step is to introduce frequency thresholds that cut off highly frequent or very infrequent words (sometimes through "stop lists", which consist of highly frequent words with less semantic content that are to be excluded). If we remove infrequent words, however, we decrease coverage and make models susceptible to what has been called the "rare word problem" (Luong et al., 2014).

A prominent approach that addresses some of these problems is to introduce a **weighting scheme** that modifies the raw count matrix. The simplest weighting scheme is **normalization**, where we divide components (i.e., raw co-occurrence counts) by the norm of either the row or the column (which is often called *scaling* in machine learning literature, to contrast it with the more standard row-wise normalization). Specifically, if $\mathbf{M}_w$ is the weighted matrix, $\|x\|$ is the L_2 norm of the vector x and $\circ$ is the Hadamard product, then

$$\mathbf{M}_w = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \circ \begin{pmatrix} \frac{1}{\|a_1\|} & \frac{1}{\|a_1\|} & \cdots & \frac{1}{\|a_1\|} \\ \frac{1}{\|a_2\|} & \frac{1}{\|a_2\|} & \cdots & \frac{1}{\|a_2\|} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\|a_m\|} & \frac{1}{\|a_m\|} & \cdots & \frac{1}{\|a_m\|} \end{pmatrix} \tag{2.1}$$

Normalization corrects for some frequency effects, but not all. Hence, researchers often apply more sophisticated weighting schemes. Many varieties exist, all modifying the co-occurrence distribution in one form or another (Curran, 2004). A well-known example is *tf-idf* (Spärck Jones, 1972).

A popular weighting scheme is pointwise mutual information (PMI) (Church and Hanks, 1990), which measures the degree of statistical dependence between two variables. It is defined as the log ratio between a word w and context c's joint probability and the product of the marginals:

$$PMI(w, c) = \log \frac{P(w, c)}{P(w)P(c)} \tag{2.2}$$

In practice, we infer the joint and marginal probabilities, $\hat{P}(\cdot, \cdot)$ and $\hat{P}(\cdot)$, directly from the raw co-occurrence matrix, hence:

$$PMI(w, c) = \log \frac{\hat{P}(w, c)}{\hat{P}(w)\hat{P}(c)} = \log \frac{\#(w, c)/n}{\#w/n \times \#c/n} = \log \frac{\#(w, c) \times n}{\#w \times \#c} \tag{2.3}$$

Or written out more explicitly:

$$\mathbf{M}_w = \begin{pmatrix} a_{1,1} \times N & a_{1,2} \times N & \cdots & a_{1,n} \times N \\ a_{2,1} \times N & a_{2,2} \times N & \cdots & a_{2,n} \times N \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} \times N & a_{m,2} \times N & \cdots & a_{m,n} \times N \end{pmatrix} \circ \begin{pmatrix} \frac{1}{\sum_i^m a_{i,1} \sum_j^n a_{1,j}} & \frac{1}{\sum_i^m a_{i,2} \sum_j^n a_{1,j}} & \cdots & \frac{1}{\sum_i^m a_{i,n} \sum_j^n a_{1,j}} \\ \frac{1}{\sum_i^m a_{i,1} \sum_j^n a_{2,j}} & \frac{1}{\sum_i^m a_{i,2} \sum_j^n a_{2,j}} & \cdots & \frac{1}{\sum_i^m a_{i,n} \sum_j^n a_{2,j}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sum_i^m a_{i,1} \sum_j^n a_{m,j}} & \frac{1}{\sum_i^m a_{i,2} \sum_j^n a_{m,j}} & \cdots & \frac{1}{\sum_i^m a_{i,n} \sum_j^n a_{m,j}} \end{pmatrix} \tag{2.4}$$

A cut-off can be introduced by further specifying that the PMI can never be below zero (i.e. if w and c are frequent but co-occur relatively few times, the cases where the product of the marginals is larger than the joint probability are set to zero), that is, *positive* PMI:

$$PPMI(w, c) = \max(0, PMI(w, c)) \tag{2.5}$$

Empirical studies have found PMI and PPMI to work well compared to other weighting schemes on a variety of tasks (Bullinaria and Levy, 2012; Kiela and Clark, 2014; Levy and Goldberg, 2014).

Another method for improving vector space quality is applying **dimensionality reduction**. A dimensionality reduction technique is usually applied after the weighting scheme, and has the benefit that it removes noise and reduces sparsity by turning the original space into a lower-dimensional dense vector space, which may yield computational benefits as well. The most popular approach is single value decomposition (SVD), which computes a three-way factorization $M = U\Sigma V^T$ where $\Sigma \in \mathbb{R}^{n \times m}$ is a diagonal matrix of ranked singular values and $U \in \mathbb{R}^{n \times n}$ and $V^T \in \mathbb{R}^{m \times m}$ are orthonormal matrices (i.e. $U^T U = I$ and $V^T V = I$). The dimensionality can be reduced by taking the first r singular values and corresponding orthonormal basis vectors. A well-known example of a vector space model that applies SVD is latent semantic analysis (LSA) (Deerwester et al., 1990; Landauer and Dumais, 1997).

Finally, after having manipulated the vector space through normalization, weighting and dimensionality reduction, we can compute the similarity between two words through a **similarity function** that outputs a scalar similarity score. While many similarity functions have been suggested over the years (see e.g. Weeds et al., 2004), the most popular method is cosine similarity (Deerwester et al., 1990), as derived from the Euclidean dot product:

$$cos(a, b) = \frac{a \cdot b}{\|a\| \|b\|} \tag{2.6}$$

2.1.2 Distributed semantics

Distributed representations learned by neural language models (Bengio et al., 2006) are an alternative to vector space model representations. For historical reasons, such distributed representations have become known as *embeddings*. Word embeddings have almost become synonymous with supervised (discriminative) models, as opposed to explicitly count-based vector space models. This distinction is probably not completely

accurate: a dimensionality-reduced vector space model (e.g. Lebret and Collobert, 2014) or a (generative) topic model-based representation (e.g. Arora et al., 2015) have also been referred to as embeddings. In this section we review supervised distributional, or distributed, semantic models.

Supervised distributional models are, with a few exceptions, discriminative models[2] that learn to predict properties of a target word's context or vice versa (e.g., words occuring in a context, or words occuring in global and local context). There are many such models available (Collobert and Weston, 2008; Mnih and Hinton, 2009; Turian et al., 2010; Huang et al., 2012; Pennington et al., 2014). The contributions of Mikolov et al. (2013a,b,c) have become very popular in the natural language processing community due to their being included in the *word2vec*[3] toolkit, which introduces two models that we will now discuss in more detail.

The continuous bag of words (CBOW) model (Mikolov et al., 2013a) learns to predict a target word in the middle of a symmetric window based on the sum of the vector representations of the context words in the window. Let u be the dot product of a target word embedding and the average of the context word embeddings:

$$u(w_t|v_1,\ldots,v_C) = v'_{w_t} \cdot \frac{1}{C}\mathbf{W} \cdot (v_1 + v_2 + \ldots + v_C) \tag{2.7}$$

where C is the context window size, $\mathbf{W}$ is a weight matrix and v'_{w_t} is the t-th row of another weight matrix **W'** of equal size to $\mathbf{W}$ (one can think of these as two separate lookup tables, one serving as the target word and the other as the context, where the objective is to predict one from the other). The probability of a word occurring in a context is given by the softmax function:

$$p(w_t|v_1,...,v_n) = \frac{\exp u(w_t|v_1,\ldots,v_{C_t})}{\sum_{j'=1}^{V} \exp u(w_{j'}|v'_1,\ldots,v'_{C_{j'}})} \tag{2.8}$$

and the objective function of the network is to maximize the log probability of the softmax:

$$\frac{1}{T}\sum_{t=1}^{T} \log p(w_t|w_{t-c},...,w_{t-1},w_{t+1},...,w_{t+c}) \tag{2.9}$$

where T is the size of the corpus. That is, the CBOW model predicts the target word from the averaged context representation.

The skip-gram (SG) model (Mikolov et al., 2013b) learns to predict the words that can occur in the context of a target word. Its objective function is as follows:

$$\frac{1}{T}\sum_{t=1}^{T} \sum_{-c\leq j\leq c, c\neq 0} \log p(w_{t+j}|w_t) \tag{2.10}$$

where T is the size of the corpus and the probability is calculated again using the softmax function:

$$p(w_{t+j}|w_t) = \frac{\exp(v'_{w_i} \cdot v_{w_t})}{\sum_{w=1}^{W} \exp(v'_w \cdot v_{w_t})} \tag{2.11}$$

[2]That is, they learn $P(x|y)$ instead of $P(x,y)$, or in absence of a probabilistic interpretation, they learn decision boundaries as opposed to the distribution.

[3]https://code.google.com/archive/p/word2vec/

That is, while the CBOW model first averages the context representations, the SG model averages a series of individual target-context predictions.

In practice, the softmax is not computed directly, but via an approximation such as a hierarchical softmax (Morin and Bengio, 2005) or negative sampling (Mnih and Teh, 2012; Mikolov et al., 2013b). One of the reasons for the popularity of these particular models, in addition to their easy accessibility in *word2vec*, is that these models are able to capture interesting linguistic regularities (Mikolov et al., 2013c), such as analogical reasoning via simple linear manipulations of the space[4]. Extensions of these models include GloVe (Pennington et al., 2014), which is another well-known embedding model that learns embeddings exhibiting linear substructures of the space in alignment to linguistic regularities, which it explains through interpreting embedding models as implicit matrix factorization. In fact, it has been shown that the objective function of skip-gram with negative sampling can be rewritten as the PMI function in Equation 2.3, with a shift of $-\log k$ (Levy and Goldberg, 2014).

In short, empirical and theoretical results (Pennington et al., 2014; Levy and Goldberg, 2014; Li et al., 2015) suggest that supervised distributional models are approximations of matrix factorization applied to PMI-weighted vector space models. It has been argued that much of the improvements resulting from the "embeddings revolution" are actually due to new hyperparameters and subsampling methods (Levy et al., 2015). These findings corroborate the earlier observation that distributional and distributed models learn similar representations in different ways: where one method explicitly learns representations from the distribution of words over contexts, the other implicitly learns approximate representations of the same (shifted) distribution, which would argue in favor of grouping both methods together under the nomer of distributional semantics. The main advantage of the implicit approach, then, is that it can be computationally more efficient.

2.1.3 Evaluations

One of the benefits of learning word representations is that they can be applied in a variety of tasks, without having to learn task-specific representations each time. Word representations are especially useful when data is limited. In order to be able to make direct comparisons between models that learn representations, distributional semantic models tend to be evaluated on the same datasets. Intrinsic evaluations measure how well representations can capture human judgments of similarity or relatedness between words. We can interpret this from a transfer learning perspective: the more accurately representations reflect human judgments, and hence human meaning representation, the more transferable they are to other tasks that rely on accurate natural language understanding.

In other words, the underlying assumption for evaluating representations on intrinsic datasets is that intrinsically high-quality representations are more likely to lead to higher performance in downstream evaluations whose performance relies on meaning representation. Intrinsic evaluations are contrasted with, and seen as separate from, extrinsic (i.e., downstream) evaluations, which measure the performance that a model achieves on a task itself, rather than directly measuring representational quality.

Intrinsic representation quality is usually evaluated by calculating the Spearman ρ_s rank correlation between the gold-standard human similarity judgments and the similarity scores computed by a model for the same word pairs. In the past, Pearson correlation

[4]Famously, *king-man+woman* $\approx$ *queen*, for instance.

has also been used for measuring performance, but upon closer inspection its applicability is questionable: humans find it much harder to attach a numerical score to a pairwise comparison like *cat-dog*, rather than having to judge whether that comparison is more similar than *cat-television* (i.e., ranking is more natural than scoring). In this section, since they are used throughout this thesis, some of the main instrinsic evaluations are introduced. Any extrinsinc evaluations are discussed in the relevant chapters themselves.

Rubenstein and Goodenough (1965) were the first to empirically corroborate the distributional hypothesis using human similarity ratings, examining "the hypothesis that the proportion of words common to the contexts of word A and to the contexts of word B is a function of the degree to which A and B are similar in meaning" (Rubenstein and Goodenough, 1965, p. 627). They constructed a dataset of 65 word pairs and obtained similarity ratings by first asking annotators to rank all comparisons according to similarity of meaning, and then assigning a score between 0 and 4. Indeed, they found evidence that "a pair of words is highly synonymous if their contexts show a relatively great amount of overlap" (Rubenstein and Goodenough, 1965, p. 633). Miller and Charles (1991)'s study of the distributional hypothesis in terms of substitutability involved re-annotating a subset of 30 word pairs of the R&G dataset in order to determine whether a new group of subjects would agree with the original semantic similarity ratings. These new annotations, on the same word pairs, have become an evaluation dataset (M&C) in their own right.

Perhaps the most popular evaluation gold standard for semantic similarity, until recently, was WordSim-353 (WS353) (Finkelstein et al., 2002). WS353 consists of human ratings for a set of 353 word pairs on a 10-point similarity scale. It has the benefit that it is considerably larger than R&G and M&C. Despite its popularity, WS353 has been criticized for a number of reasons. One of the main issues is that word pairs were annotated without making any explicit distinction between similarity and relatedness (Agirre et al., 2009). This distinction is important because it has potential repercussions for downstream performance: if the learned representations capture mere relatedness, they are not suitable for constructing a thesaurus of synonyms, because highly related but dissimilar terms like *car* and *petrol* will receive high similarity scores under the model. Conversely, for a task such as text classification, we are much more interested in the relatedness of words, especially if the classes are distinguished by topic: knowing that a *dog* and a *cat* are both animals is more informative of the semantic content of text than knowing that *canine* and *feline* are their respective synonyms. Another problem with WS353 is that it contains proper names specific to the time around which it was created, such as *Arafat* and *Maradona*. More importantly, inter-annotator agreement is low, and 353 is a relatively small number of word pairs. Several datasets have been created to alleviate some of these problems.

The MEN dataset (Bruni et al., 2012) consists of 3000 comparisons of randomly selected words that occur at least 50 times as tags in the ESP game dataset (von Ahn and Dabbish, 2004). It was constructed specifically for the study of grounded semantic models, and consequently contains more concrete words (i.e., words with physical referents) than some other datasets. It consists of comparisons between 751 individual lexical items. Each pair is scored on a semantic relatedness (as opposed to similarity) scale through the online Amazon Mechanical Turk crowdsourcing platform. Another way to measure the capability of a model to capture relatedness is via norming studies such as the University of South Florida (USF) assocation norms (Nelson et al., 2004). Association norms are ob-

tained by presenting subjects with a cue word and asking them to name associated words in response. For instance, the cue *rice* might be associated with *white, food, wedding*, et cetera. The USF association norms provide a set of associated words and the frequencies with which they were produced, which can be used to compute a probability distribution over associated words per cue.

SimLex-999 (Hill et al., 2015) was constructed specifically to address the similarity-versus-relatedness problem that WS353 suffers from. Agirre et al. (2009) had tried to address this issue through splitting WS353 into similarity and relatedness-specific subsets (WordSim and WordRel, respectively), but did not re-annotate the new subsets, whose scores consequently still did not distinguish between similarity and relatedness. Hill et al. (2015) cite three main issues with WS353: 1) many dissimilar word pairs receive a high rating; 2) no associated but dissimilar concepts receive low ratings; and 3) it has low inter-annotator agreement, that has already been surpassed by various distributional semantic models. SimLex-999 focuses explicitly on what it calls "genuine similarity": only genuinely similar word pairs, e.g. *car-automobile*, receive high scores, while dissimilar but related words, e.g. *coffee-mug*, receive low scores.

There are many other instrinsic evaluations available and the construction of such datasets is an active area of research; see e.g. Faruqui and Dyer (2014) for a more comprehensive list. While intrinsic evaluations are often used to evaluate the performance of semantic representations, it is important to note that such representations are good for much more than simply mirroring human similarity ratings. The representations are used in many engineering applications and constitute core components of many natural language processing pipelines and are worth studying on their own as psychological representations of meaning (Lenci, 2008).

2.2 Grounded distributional semantics

A key observation for understanding how conceptual representations bridge the gap between perceptual input and symbolic reasoning, is that concepts are *grounded* in physical reality and sensorimotor experience through perception (Louwerse, 2008). The fact that distributional semantic models represent the meaning of a word as a distribution over others implies that they suffer from the *grounding problem* (Harnad, 1990; Perfetti, 1998; Barsalou, 1999). Indeed, it has been found that text-based distributional models capture linguistic properties of word meaning, but often fail to capture concrete aspects, such as the fact that bananas tend to be yellow (Baroni and Lenci, 2008; Andrews et al., 2009; Riordan and Jones, 2011). There has been a surge of recent work on perceptually grounded semantic models that try to mitigate this problem, which have outperformed state-of-the-art text-based methods on a variety of natural language processing tasks.

2.2.1 The grounding problem

The grounding problem has a long history in the philosophy of meaning, arguably going all the way back to Plato and Aristotle. Its most well-known incarnation within the context of Artificial Intelligence is, indubitably, the Chinese room thought experiment by Searle (1980): Imagine a non-Chinese speaker locked inside a room with nothing but a big book of rules that dictate how to manipulate sequences of Chinese input symbols in order to generate perfectly grammatical sequences of Chinese output symbols. From

the perspective of an outside observer, if input sequences yield perfectly sound output sequences, does the person in the room speak and understand Chinese? With the problem phrased like this, even if it seems to the outside observer that the person in the room is fluent in Chinese, we are inclined to answer in the negative: the person inside the room is following a set of rules for symbol manipulation without understanding the meaning of the symbols themselves or the compositional meaning of symbol sequences. The Chinese room argument has spawned an enormous amount of philosophical literature (Cole, 2015). For our current purposes, the most important implication of this argument is that it leads to what has become known as the symbol grounding problem (Harnad, 1990): how can you know the meaning of a symbol if it is defined only through other symbols? The implied circularity in this question is problematic, since it pre-empts symbols having non-symbolic referents, which provides a solely solipsistic account of the meaning of symbols where symbols gain content only by virtue of other symbols, ad infinitum.

One might argue that the Chinese room, as a system, in fact does speak and understand Chinese. This has been called the systems reply, which is closely aligned with a connectionist interpretation of the same principle, sometimes called the "brain simulator" reply, which argues that the room as a system simulates the brain. Another, arguably even more convincing, explanation for the problem has become known as the robot reply (Cole, 2015), which concedes that the person (i.e., the symbol computer) in the room does not understand Chinese, but that this does not imply that the symbol computer cannot know meaning. The robot reply argues that symbol meaning may be understood through experiencing physical reality—seeing, making, tasting, or hearing others speak of, a concept. That is, it argues that suitable causal connections with the world can provide content to internal symbols, which implies that meaning is grounded in physical reality through agents being *embodied* in the world.

2.2.2 Embodiment

Complementary to these philosophical considerations, one of the main motivations for building perceptually grounded models lies in human concept acquisition. There is a lot of evidence that human semantic knowledge is grounded in the perceptual system and sensorimotor experience (Glenberg and Kaschak, 2002; Barsalou, 2008). For example, language acquisition in young children is heavily dependent on their direct environment (Jones et al., 1991; Landau et al., 1998) and children learn concrete, perceptual, nouns first (Bornstein et al., 2004). The question of grounding is, empirically speaking, heavily intertwined with the notion of embodiment in cognitive science—the hypothesis that cognitive processes of all kinds are rooted in perception and action (Meteyard and Vigliocco, 2008).

Theories of semantic representation in cognitive psychology and neuroscience put varying degrees of emphasis on the presence or absence of sensory and motor information in word meaning (Clark, 1999; Wilson, 2002). To distinguish between different underlying assumptions, the theories can be described as a continuum of degrees of embodiment (Meteyard et al., 2012):

- Unembodiment (Quillan, 1966; Newell, 1980). Semantic representation is fully independent of sensory and motor information: semantic information is symbolic and semantic processing does not require perception.

- Secondary embodiment (Mahon and Caramazza, 2008). Semantic content is grounded by interaction via secondary activation with sensory and motor information, but semantic representations are modality invariant (amodal). The semantic system is independent of but directly associated with sensory and motor information.

- Weak embodiment (Farah and McClelland, 1991; Simmons and Barsalou, 2003). Semantic representations are at least partly constituted by sensori-motor information. Sensory and motor information has a representational role and activation of semantic content will be able to influence processing in primary cortical areas, and vice versa.

- Strong embodiment (Barsalou, 1999; Glenberg and Kaschak, 2002; Gallese and Lakoff, 2005). Semantic representations are completely dependent on sensory and motor systems. Sensory and motor systems represent semantic content during 'simulation' (Gallese, 2007), directly modulating semantic processing.

There are interesting parallels between this continuum of embodiment in cognitive psychology and neuroscience on the one hand and theories of meaning in philosophy and linguistics on the other. Unembodiment is closely associated with the classic cognitive theory of symbolic computation (Newell and Simon, 1976), while connectionism, which posits that cognition happens in a parallel (neural) network in a distributed fashion (Rumelhart et al., 1986; Smolensky, 1988), has traditionally been viewed in cognitive science as fundamentally incompatible with unembodied theories of meaning (Meteyard et al., 2012).

It could be said that the meaning of *meaning* has two sides (Evans, 2015): language and concept representation. A theory of semantics has to account for how the two are connected (in the brain, or otherwise). Symbolic and unembodied theories do not account for this aspect of semantics at all—it has been argued that it was seen as too difficult of a problem, so it was simply explained away (Evans, 2015). Conversely, strong embodiment occupies the other side of the spectrum, essentially explaining away language understanding as low-level perceptual processing. There is broad agreement in the cognitive sciences that semantic representations interact with sensori-motor information. What exactly constitutes true semantic representation—and whether sensori-motor information is necessary and sufficient, rather than secondary—remains a matter of debate (Meteyard et al., 2012).

In summary, there exists a large body of work in the cognitive sciences that supports at least some form of embodiment, i.e. some degree of grounding, in semantic representation. How exactly that grounding is done, and to what extent meaning representations rely on embodiment, remains an open question.

2.2.3 Perceptual representations

In part to address the grounding problem, as well as from a general curiosity about the possibilities of applying machine learning to the problem of connecting natural language processing and perception (Mooney, 2008), new types of models have emerged in recent years that combine corpus-derived textual data (e.g. as described in the previous section) with perceptual data, in order to derive grounded representations.

Perceptually grounded models learn semantic representations from both textual and perceptual input. One method for obtaining perceptual representations is to rely on direct

human semantic knowledge, in the shape of feature or association norms (e.g. Nelson et al., 2004; McRae et al., 2005), which have been used successfully in a range of multi-modal models (Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013; Hill and Korhonen, 2014; Kievit-Kylar, 2014; Bulat et al., 2016). However, norms are elicited from human annotators and as a consequence are limited in coverage and relatively expensive to obtain. An alternative approach, that does not suffer from these limitations, is to make use of raw data as the source of perceptual information: images, for example.

2.2.3.1 Bag of visual words

A popular approach has been to collect images associated with a concept, and then lay out each image as a set of keypoints on a dense grid, where each keypoint is represented by a robust local feature descriptor such as SIFT (Lowe, 2004). These local descriptors are subsequently clustered, across concepts, into a set of "visual words" using a standard clustering algorithm such as k-means and then quantized into vector representations by comparing the descriptors with the centroids. This approach has become known as "bag of visual words" (BoVW) (Sivic and Zisserman, 2003) and has been used extensively in computer vision, as well as by models aiming to perform grounding in perceptual information.

More precisely, let $\mathcal{I}$ be a set of images. The process then comprises the following steps:

1. Keypoint identification. For each image $i_w \in \mathcal{I}$, identify keypoints $k_i \in \mathcal{K}(i_w)$ for which features will be obtained. Example identification methods include laying keypoints out as a dense grid, applying segmentation or automatically identifying points of interest.

2. Feature description. For each keypoint k_i, obtain a feature descriptor $f(k_i)$, using e.g. SIFT (Lowe, 2004). If the identification mechanism in step (1) consists of a dense grid, this is known as DSIFT (dense SIFT). The feature description mechanism yields a set $\mathcal{F}(i_w) = \{f(k_i) \mid k_i \in \mathcal{K}(i_w)\}$ of local feature descriptors per image.

3. Visual word generation. Using a clustering algorithm such as k-means, obtain a set of centroids $\mathcal{C}$ for the set of all local feature descriptors for all images, i.e., $\{f \mid f \in \cup_{i_w \in \mathcal{I}} \mathcal{F}(i_w)\}$, or a randomly sampled subset thereof.

4. Visual word assignment. Let $g(f_{k_i}) = idx(\text{argmin}_c\, d(f_{k_i}, c))$ for $c \in \mathcal{C}$, where idx returns the index of the centroid in $\mathcal{C}$ and d is some distance function. That is, let g return the index of the closest centroid to a given local feature descriptor. Let r_i be the image representation vector indexed by i.

   Then $r_i = \sum \mathbf{1}_{g(f_{k_i})=i}$ for $1 \leq i \leq |\mathcal{C}|$, where $\mathbf{1}$ is an indicator function, i.e., we count the number of occurrences of each cluster assignment.

2.2.3.2 Sources of visual perceptual input

The easy availability of images on the World Wide Web makes the visual modality the modality of choice for perceptual grounding. Images can be obtained from a variety of sources. Ideally, one would jointly learn grounded representations from parallel multi-modal data, such as text containing images, but such data is hard to obtain, has limited

coverage and can be noisy or biased (e.g. news article images are unlikely to be very descriptive, while Wikipedia articles are more likely to be illustrative, etc.). Hence, image representations are often learned independently. This has the major advantage that we can learn representations independently, from much larger sources of textual and/or image data. Raw image data is cheap, plentiful, easy to obtain and has much better coverage (Baroni, 2016).

ImageNet (Deng et al., 2009) is a large ontology of images developed for a variety of computer vision applications. It serves as a benchmarking standard for various image processing and computer vision tasks, including the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015). ImageNet is constructed along the same hierarchical structure as WordNet (Miller, 1995), by attaching images to the synset (synonym set).

The ESP Game dataset (von Ahn and Dabbish, 2004) is a dataset containing 100,000 images labeled through a so-called "game with a purpose". Two players are matched online and must independently and within a time limit agree on an appropriate word label for a randomly selected image. Once a word has been mentioned a certain number of times in the game, the word becomes a taboo word and cannot be used as a label anymore. ESP contains 20,515 unique tags.

Search engines that allow image search may also be used, such as Google Images[5] (e.g. Bergsma and Goebel, 2011) or Bing[6], or image upload websites such as Flickr[7] or Tumblr[8]. It has been shown that images from Google yield higher quality representations than comparable resources such as Flickr and are competitive with "hand prepared datasets" (Fergus et al., 2005), meaning that an annotator manually created datasets for given items and inspected how well these matched Google's search results.

Although it is beyond the scope of this thesis, there has also been work on grounding language in video data (see e.g. Gupta et al., 2009; Regneri et al., 2013; Yu et al., 2015, and references therein), as well as in robotics (Cangelosi and Riga, 2006) and even cognitive ecology (Hutchins, 1995).

2.2.3.3 Aggregation

Commonly, a set of images associated with a certain word is retrieved from a perceptual data source. A method such as bag of visual words, or another method, can subsequently be applied in order to obtain an **image representation** for each of the associated images. Depending on the task (or more specifically, the similary function), it is often useful to aggregate these image representations into a single **visual representation** for the given word. In other words, for a set of images $\mathcal{I}_w$ for a word w, we apply some aggregation function f to obtain a visual representation v:

$$v_w = f(i_w^1, i_w^2, ..., i_w^n) \tag{2.12}$$

where $i_w^i \in \mathcal{I}_w$ are the image representations. Examples of aggregation functions are summing, averaging, or taking the pointwise maximum.

[5]https://images.google.com/
[6]https://www.bing.com/images
[7]https://www.flickr.com/
[8]https://www.tumblr.com/

2.2.4 Multi-modal models

The first multi-modal distributional semantic model to use image data was introduced by Feng and Lapata (2010). They propose a generative model based on latent Dirichlet allocation (LDA) (Blei et al., 2003) that learns latent multi-modal dimensions from a large collection of news articles that contain both text and accompanying images. That is, their model represents documents as a mix of textual co-occurrences (from standard bag of words) and visual words (from bag of visual words). The resultant representations perform better on the WordSim353 and USF association norm intrinsic evaluations when visual information is taken into account.

A related approach is that of Leong and Mihalcea (2011b), who construct a vector space model of textual contexts and visual words and subsequently apply LSA to reduce the number of dimensions and extract latent representations. Leong and Mihalcea (2011a), instead, keep the textual and visual spaces as individual spaces and apply similarity functions separately, yielding a textual and a visual similarity score. These similarity scores are subsequently combined, by taking the sum or the harmonic mean (the F1-score (Rijsbergen, 1979)) of the modality-specific similarity scores.

Bruni et al. (2011) introduced a third way for combining textual and visual spaces: instead of learning a latent space or combining the individual per-space scores, they concatenate vectors from the two spaces into a single multi-modal space. Both spaces are normalized to ensure that components from each modality have equal weight. Bruni et al. (2012) use a similar method, but do *weighted* concatenation:

$$F = \alpha \times v_{text} \parallel (1 - \alpha) \times v_{vis} \tag{2.13}$$

where $\parallel$ denotes concatenation. In addition to using BoVW features, they also use LAB (Fairchild, 2005) features, which explicitly encode color information. They find that the usage of visual information is particularly useful for modelling the meaning of words with visual correlates, such as color terms, even in tasks that involve non-literal usages of color terms.

The above approaches have several aspects in common. All of them use BoVW with automatic keypoint detection algorithms (difference-of-Gaussians for Feng and Lapata (2010), SIFT for the others). In all cases except the method of Feng and Lapata (2010), which aggregrates only implicitly through LDA, the aggregation function that takes image representations and constructs concept-level visual representations consists of simply summing up the features. Bruni et al. (2014) instead use a dense grid for keypoints, laid out at different scales, called PHOW (Bosch et al., 2007), which is computationally more efficient than having to first identify points of interest before obtaining local feature descriptors (Nowak et al., 2006). They introduce a generalized multi-modal framework, which takes a textual and visual vector space and performs what they call "latent multi-modal mixing", followed by a splitting step where the two modalities are separated, culminating in a multi-modal similarity estimation. In essence, this framework closely mirrors text-based distributional semantic models, in that it constructs a vector space (in this case a multi-modal one), applies weighting and dimensionality reduction, and computes a similarity function over the resultant latent space.

Silberer et al. (2013) take a different approach from the above BoVW-based methods: they train a set of visual attribute classifiers (see e.g. Ferrari and Zisserman, 2007; Farhadi et al., 2009) and integrate the classifier predictions with text-based distributional semantic models. A separate classifier is learned for each feature, thus yielding a vector

$(s^1(i_w), s^2(i_w), ..., s^n(i_w))$ of scores s^i for n attributes for an image i_w related to word w. Image representations are aggregated into a visual representation by taking the mean. Silberer and Lapata (2014) follow up on this work by training a stacked autoencoder (Bengio et al., 2007) to induce a multi-modal semantic representation from textual and visual representations, which constitutes the first deep learning approach to the integration of multi-modal features. Essentially, instead of using e.g. SVD for dimensionality reduction, they use neural networks for reducing the dimensionality of the data (Hinton and Salakhutdinov, 2006).

An interesting hybrid approach between using raw image data, and norms (association as well as feature norms) as discussed in Section 2.2.3, is Roller and Schulte im Walde (2013). They use SURF (Bay et al., 2008) instead of SIFT, plus a GIST (Oliva and Torralba, 2001) descriptor that gives a global representation of an image, which they combine with textual and norm-based features. Their aggregation method is a variant of LDA, where document representations are tri-modal, consisting of text-based, image-based and norm-based (what they call "cognitive") components.

Lazaridou et al. (2015b) observe that constructing textual and visual spaces separately and then merging them is very different from how humans learn about concepts. Furthermore, these approaches are arguably founded on an underlying assumption that textual and visual information are available for each and every concept. To alleviate these issues, they introduce a multi-modal skip-gram model that modifies the skip-gram objective of Equation 2.10 to also incorporate visual information, if available:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, c\neq 0}\log p(w_{t+j}|w_t) + \mathcal{L}_{vision}(w_t) \tag{2.14}$$

where $\mathcal{L}_{vision}(w_t)$ is an additional objective, that is, one of two vision-based skip-gram objectives: either a max-margin criterion that maximizes similarity between textual and visual representations, or by mapping from visual representations to textual space via a cross-modal mapping (Lazaridou et al., 2014, see also Section 2.2.6) and treating the mapped representation as an additional context.

With the exception of Lazaridou et al. (2015b)'s model, which learns distributed representations, all multi-modal models outlined in this section use vector space models for obtaining textual semantic representations.

2.2.5 Fusion

Multi-modal models typically comprise three steps: 1) aggregation of uni-modal representations; 2) *fusion* or mixing of representations; and 3) the application of fused representations. As the previous section has shown, there are a variety of ways in which information from the textual and perceptual modality can be fused (also called "combined", "mixed" or "integrated" in the literature). In order to get a clearer view of the types of multi-modal models one can apply, we divide multi-modal fusion methods into three distinct types: *early*, *middle* and *late* fusion. The three types are distinguished by whether, how and at what stage each of these three steps are applied. Bruni et al. (2014) discuss a similar division based on image analysis methods for information retrieval, but do not distinguish between early and middle fusion (they call both early fusion). See Table 2.1 for an overview.

	Early	**Middle**	**Late**
Aggregation	Joint objective	Sum, mean, max, etc.	Sum, mean, max, etc.
Mixing	Latent representation	Concatenation (normalized), dimensionality reduction, etc.	Keep separate
Similarity	Compute score	Compute score	Combine scores
Examples	Feng and Lapata (2010), Roller and Schulte im Walde (2013), Lazaridou et al. (2015b)	Leong and Mihalcea (2011b), Bruni et al. (2011), Bruni et al. (2012) Silberer et al. (2013), Silberer and Lapata (2014)	Leong and Mihalcea (2011a)

Table 2.1: A categorization of multi-modal fusion methods.

2.2.5.1 Early fusion

In early fusion, uni-modal representations are typically implicitly aggregated. Examples of early fusion are Feng and Lapata (2010), Roller and Schulte im Walde (2013), Srivastava and Salakhutdinov (2014) and Lazaridou et al. (2015b). Instead of first explicitly computing concept-level modality-specific spaces, these methods perform fusion as a part of the learning objective. That is, although such methods do not necessarily require textual and visual data to be extracted from the same corpus, they learn multi-modal representations *jointly*, through some objective that incorporates information from both modalities. More formally, such models learn some posterior distribution, or a factorization thereof, based on both modalities, i.e. $P(\cdot \mid f^1_{text}, f^2_{text}, \ldots, f^m_{text}, f^1_{vis}, f^2_{vis}, \ldots, f^n_{vis})$, where f^i_{mod} is a modal feature.

2.2.5.2 Middle fusion

Whereas early fusion requires a joint training objective that takes into account both modalities, middle fusion allows for individual training objectives and non-overlapping, i.e., independent, training data. Similarity between two multi-modal representations is calculated as follows:

$$sim(u, v) = g(f(u^l, u^a), f(v^l, v^a))$$

where g is some similarity function, u^l and v^l are textual representations, and u^a and v^a are perceptual representations. A common formulation for $f(x, y)$ is $\alpha\, x \parallel (1 - \alpha)y$, where $\parallel$ is concatenation. Usually, but not necessarily, the uni-modal representations are normalized. If $\alpha = 0.5$, f becomes concatenation without any modality-specific weighting. After having created a single multi-modal space, weighting or dimensionality reduction may be applied. Most multi-modal models fall in this category.

2.2.5.3 Late fusion

Late fusion can be seen as the converse of middle fusion, in that the similarity function is computed first before the similarity scores are combined:

$$sim(u, v) = h(g(u^l, v^l), g(u^a, v^a))$$

where g is some similarity function and h is a way of combining similarities, often a weighted average: $h(x, y) = \alpha\, x + (1 - \alpha)y$; and we use cosine similarity for g. Since cosine similarity is the normalized dot-product, middle and late fusion are equivalent if $\alpha = 0.5$ and the uni-modal representations are normalized. Leong and Mihalcea (2011a)'s combination of uni-modal similarity scores is an example of this type of fusion.

2.2.5.4 Polymodal fusion and cognitive plausibility

It is easy to see how this categorization straightforwardly extends beyond the bi-modal case, which we might call *polymodal* to distinguish it from the case where we use a single perceptual modality, as is usually done in multi-modal semantics. The more modalities we introduce, the more parameters will be involved to govern the exact fusion strategy. Cognitively speaking, late fusion seems somewhat unlikely: the brain does not compute a scalar function (e.g. similarity) per modality to solve a specific task by combining the scores. Middle fusion methods are probably too simplistic, given that human concept acquisition is situational and often relies on joint stimuli, i.e., we often get input from multiple modalities at the same time. Early fusion, however, is probably too rigid: we can be very familiar with the sound of a violin without ever having seen one or knowing anything about violins, or know what a lavender field in Southern France looks like without ever having smelled lavender—that is, joint learning is not always necessary. The most cognitively plausible fusion method, thus, is probably a combination of early and middle fusion, which allows for learning uni-modal representations independently but which also allows for combining said representations into an overall multi-modal one that takes all modalities into account, possibly in varying degrees. The fact that humans are suspectible to modal priming (e.g. that comparing "blue" to "green" makes us focus on the visual modality, rather than the auditory "blue" note in music) (Vallet et al., 2010) suggests that modal representations are at least to some degree separate in the brain.

It is important to note that an answer to these questions directly relates to the discussion about embodiment in cognitive science, discussed in Section 2.2.2. That is, early fusion can be thought of as tending more toward *secondary embodiment*, where semantic representations are amodal but directly associated with sensory and motor information, while middle fusion, in that view, more closely resembles *weak embodiment*, where sensorimotor information is represented separately for each modality and the semantic content of the representations influences processing.

2.2.6 Cross-modal semantics

While the above focuses on fusing information from textual and perceptual modalities into multi-modal representations, which we might call *representational* grounding, the grounding problem can also be addressed by designating the referent for a concept, which can be seen as a specific case of grounding, namely *referential* grounding. Such a distinction has deep roots in the theory of meaning: it closely mirrors e.g. Frege's sense

and reference distinction (Frege, 1892) or Peirce's three levels of meaning—icon, referent and interpretant (i.e., representation, "whose relation to their objects is an imputed character") (Peirce, 1936; Atkin, 2013). Grounding meanings in this referential manner is sometimes called cross-modal semantics, and its goal is to map across modalities, from textual representations into perceptual ones, and vice versa, in order to establish a link between words and the things they denote in the physical world (Baroni, 2016).

Consider for instance the statement *There is a dog in the room* (Lazaridou et al., 2014). A purely text-based system might understand the meanings of *dog* and *in* and *room*, but in order to know whether a thing in the room is a dog, it will need to know what dogs look like and not mistake it for a cat. In other words, it cannot establish the truth or falsity of the statement because it is solipsistic and not linked to reality. We can think of this as a special case of the grounding problem, and its solution is to learn a cross-modal map between the textual space and the visual space, in order to identify the referents of words. Both Frome et al. (2013) and Socher et al. (2014) take this approach mapping from vision to text. Lazaridou et al. (2014) also learn the mapping from text to vision: if a human reads the sentence *There is a cute hairy wampimuk sitting by the tree*, they will have a good idea of what a "wampimuk" will look like even if they have never seen it (or indeed, if it doesn't exist at all). Lazaridou et al. (2015a) extend cross-modal mapping beyond nouns to adjectives and adjective-noun pairs. Cross-modal mappings are usually evaluated through leave-one-out experiments, where the objective is to map from one space to the other and retrieve the correct concept, without ever having seen it before. This is a special case of zero-shot learning (Palatucci et al., 2009), where the goal is to learn a classifier $f : X \rightarrow Y$ that predicts novel values of Y that were not in the training set.

Cross-modal semantics is a recent idea and there are many open questions. A natural question to ask is why the cross-modal mapping is performed from the linguistic to the visual modality and vice versa, instead of from multi-modal space to image space or vice versa—since the goal is to establish the referent of a language token denoting a concept, it seems unfair not to include the entire concept representation but only its linguistic subset. This is partially addressed by Lazaridou et al. (2015b), who learn a mapping from multi-modal space to visual space, but still map to the mean representation (i.e., the visual representation) of concepts, instead of referring to individual instances or tokens of visual entities (i.e., image representations). A similar approach is described in Bulat et al. (2016), who learn such a mapping using feature norms instead of multi-modal representations. It would make much more sense, in the case of referential grounding, to learn to map to image space, or even better, a segment of an image, to pick out the exact referent.

2.3 Language and vision

Language and vision are two of the core pillars of artificial intelligence research. Since both often rely on related machine learning techniques, there has long been an interest in cross-pollination between the fields of computer vision and computational linguistics (see e.g. Mooney, 2008) and recent years have seen a sharp increase in such works. There are many tasks that require a combination of linguistic and visual information and there is a wide variety of data available on the World Wide Web that incorporates both visual and textual information, in the form of images with tags or captions, news articles with images, diagrams or maps explaining processes, slides in conjuction with lectures, videos

with subtitles, multi-modal social media posts, and so on. Such tasks have become increasingly important in both communities, as exemplified in the emergence of "language and vision" tracks at premier conferences in both fields and dedicated workshops focusing on the topic. A core problem has become automatically providing descriptions of images—i.e., captions—which unifies the respective goals of understanding images and analyzing and generating language of the two fields; see Bernardi et al. (2016) for an excellent survey. More recently, visual question answering (Antol et al., 2015), the task of answering questions about visual scenes or scenarios, has also gained attention as an important AI task.

There are many examples of language information being used in computer vision tasks (e.g. Barnard et al., 2003; Berg et al., 2010; Frome et al., 2013). Conversely, computer vision techniques have successfully been applied to various natural language processing tasks. For instance, Bergsma and Goebel (2011) show how visual information can be exploited for predicting selection preferences, while Bergsma and Van Durme (2011) used image representations for identifying words in different languages with the same meaning. The numerous applications at the intersection between language and vision are beyond the scope of this thesis, but the fact that this is a rapidly growing area illustrates the importance of multi-modal research.

2.4 Deep learning

In recent years natural language processing has benefited enormously from the improved representations that were obtained using neural networks, these days often referred to as deep learning[9]. The impact of deep learning on other fields, in particular speech recognition and computer vision, but also drug discovery and genomics, has possibly been even greater (LeCun et al., 2015). The main reason driving the success of deep learning is the availability of enormous amounts of data and just as importantly, computational power, often in the form of GPUs.

2.4.1 Deep learning and grounding

Grounding and connectionism, as the study of neural networks has also been called, have something of a shared history. With the initial advent of connectionism and its subsequent popularity in the philosophy of mind (Van Gelder, 1991), it was claimed that the grounding problem could be solved with neural networks (Harnad, 1993). Indeed, connectionism was popularized within the cognitive science community in part as a way of solving the grounding problem (Christiansen and Chater, 2001). Although this view has several shortcomings (discussed in e.g. Christiansen and Chater, 1992), there is some merit to this claim: in particular, without going into the full philosophical details, the processing of a neural network is systematically (and causally) determined by its inputs, according to the semantic content of its distributed representations. As such, if a network is provided, at least partially, with inputs from physical reality, the network is grounded. It has been argued that connectionism is not grounded because input representations are arbitrary

[9]In fact, in the narrow definition of deep learning, it requires that neural networks are deep–i.e., consist of more than a single layer. Lately, the term has come to more generally refer to any type of learning that involves neural networks. We will use the term here in the latter sense. Most of the networks applied in this thesis, but not all of them, are in fact deep.

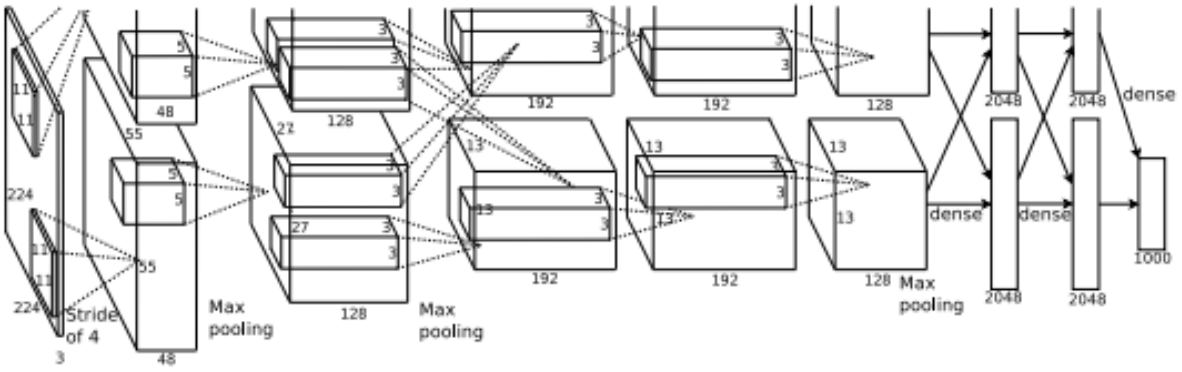

Figure 2.1: Illustration of a convolutional neural network (from Krizhevsky et al., 2012, p. 5), for the ImageNet image recognition task.

reflections of physical reality (Christiansen and Chater, 1992): while this was true of most connectionist models of that era, it certainly is not true of deep learning models making use of direct perceptual input, e.g. in the form of raw image data. In cognitive science, the interest in connectionism has largely been superseded by dynamical systems theory (noting that connectionist networks are a special case of dynamical systems), which has also been studied in an embodied cognition context (Hotton and Yoshimi, 2011). With the reawakening of deep learning, which has so far focused mainly on "engineering" oriented applications, and the increased focus on representation learning (Bengio et al., 2013), it makes sense to revive the interest in grounding with neural networks, or "deep embodiment", if we make use of deep learning applications. There have been recent preliminary studies in applying deep learning to grounding (e.g. Monner and Reggia, 2011), and this thesis is another exponent of that idea.

2.4.2 Convolutional neural networks

Convolutional neural networks (CNNs) (LeCun et al., 1998) are inspired by biological visual processing (Fukushima, 1980). A CNN is characterized by a network architecture that aims for a degree of shift, scale and distortion invariance, through local receptive fields, shared weights or weight replication (i.e. feature maps), and spatial or temporal sub-sampling (i.e. pooling) (LeCun et al., 1998, p. 6). Although CNNs have also been successfully applied to NLP tasks and some of the first word embeddings were obtained by CNNs (Collobert and Weston, 2008), their biggest impact has been in the computer vision community. According to LeCun et al. (2015), CNNs were largely forsaken until they were successfully applied in the ImageNet competition (Russakovsky et al., 2015) by Krizhevsky et al. (2012)[10]. The usage of CNNs led to almost halving the error rates of the best competing approaches. They are now the dominant approach for almost all recognition and detection tasks in the computer vision community, approaching or even exceeding human performance on some tasks (e.g. Weyand et al., 2016).

[10]This view has been disputed, see e.g. this blogpost by Jürgen Schmidhuber: `http://people.idsia.ch/~juergen/deep-learning-conspiracy.html`

The network by Alex Krizhevsky (2012), sometimes also called AlexNet, introduces the following network architecture (see Figure 2.1): first, there are five convolutional layers, followed by three fully-connected layers, where the final layer is fed into a softmax which produces a distribution over the class labels (in this case ImageNet labels). All layers apply rectified linear units (ReLUs) (Nair and Hinton, 2010) and use dropout for regularization (Hinton et al., 2012). The network is trained to maximize the multinomial logistic regression objective:

$$J(\theta) = -\sum_{i=1}^{D}\sum_{k=1}^{K} \mathbf{1}\{y^{(i)} = k\} \log \frac{\exp(\theta^{(k)\top} x^{(i)})}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x^{(i)})} \tag{2.15}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, $y^{(i)}$ is the output, $x^{(i)}$ is the input and training is performed on D examples with K classes.

2.4.3 Transfer learning

Based in part on the recent successes of deep learning and convolutional neural networks, the technique of deep transfer learning has gained attention in the computer vision community. First, a deep convolutional neural network is trained on a large label dataset, such as ImageNet. The convolutional layers are then used as mid-level feature extractors on a variety of computer vision tasks in which features from a trained network are transferred to a different task (Oquab et al., 2014; Girshick et al., 2014; Zeiler and Fergus, 2014; Donahue et al., 2014). Although transferring convolutional network features is not a new idea (Driancourt and Bottou, 1990), the simultaneous availability of large datasets and cheap GPU co-processors has contributed to the achievement of considerable performance gains on a variety computer vision benchmarks. Such "off the shelf" CNN features have led to improvements over traditional descriptors, such as SIFT (Lowe, 2004) and HOG (Dalal and Triggs, 2005), in a wide variety of taks (Razavian et al., 2014). The current thesis applies this idea across modalities, and transfers CNN features to computational semantics tasks that require natural language understanding.

2.5 Discussion

In this chapter, we reviewed some of the core components of the subject matter of this thesis: distributional models for language, the grounding of such models, and the rise of deep learning. We have seen how neural networks have led to more sophisticated representations in NLP and how deep learning has transformed computer vision, due to a combination of data availability and increased computational power. Representation learning is becoming ever more important, and the idea that such representations need to be grounded is gaining traction in the NLP community. This work finds itself at the confluence of these developments.

As noted, connectionism and grounding have been closely related in discussions in cognitive science, artificial intelligence and the philosophy of mind. It seems that neural networks are particularly well-suited for grounding, on account of their handling and representing of raw perceptual data—an old idea which this thesis hopes to revive. Combining this with linguistic information leads to grounded representations of a high quality,

which further lead to improvements both in applied natural language processing tasks, as well as to possibilities for examining empirical questions in cognitive science.

In what follows, we build on this background, and aim to perform grounding using deep learning techniques, specifically focusing on how features can be transferred. In particular, we will focus on the quality of uni-modal perceptual representations and how an increased quality of representations leads to better multi-modal models even with relatively simple fusion techniques (specifically, middle fusion through concatenation). This focus is deliberate: a plethora of fusion methods have been suggested, and the question easily merits a thesis of its own right. Furthermore, given the discussion of the cognitive plausibility of fusion techniques above, this thesis subscribes to a weak embodiment view, assuming that modal content is semantically represented individually and fused at a later stage in further processing depending on the problem the brain/system is trying to solve.

The Chinese room argument, of which the grounding problem is an exponent, was cleverly designed to illustrate how important meaning, specifically grounded meaning, is to the qualitative nature of human consciousness. The fact that meaning is so essential to human intelligence should be taken as an indicator of the importance of meaning—and its study, in the shape of semantics—to the great endeavor of AI.

Neural networks, which have become popular for representing natural language and which have led to substantial improvements in computer vision, constitute a natural way for investigating the grounding of meaning. Many core natural language processing tasks are uni-modal, focusing on linguistic input, despite the fact that human language understanding is grounded in perceptually rich environments. If we want to move towards human-level artificial intelligence, we will need to build multi-modal models that represent the full complexity of human meaning, including its grounding in our various perceptual modalities.

Part II

Visual grounding

CHAPTER 3

Improving visual grounding with CNNs

As we have seen in the previous section, the bag of visual words (BoVW) method has been superseded in computer vision by deep convolutional neural networks (CNNs) (LeCun et al., 1998; Krizhevsky et al., 2012). Transfer learning techniques have gained considerable traction in computer vision, especially with regard to deep learning (Razavian et al., 2014). CNNs now hold the state-of-the-art in almost every computer vision task (LeCun et al., 2015). This chapter reports on results obtained by using CNN-extracted features in multi-modal distributional semantic models. While all previous multi-modal models used BoVW, we here report on the first effort to apply CNN features. Through combining such features with supervised distributional semantic models, the work described herein was the first approach to multi-modal distributional semantics that exclusively relies on deep learning in both its linguistic and visual components.

3.1 Model

Figure 3.1 illustrates how the proposed system computes multi-modal semantic representations. For a word w, a set of relevant image representations $\mathcal{I}_w$ are obtained and aggregated into a visual representation r_w^v, which is subsequently combined, through some middle fusion function f, with a linguistic skip-gram representation r_w^l to get a multi-modal representation: $r_w^{mm} = f(r_w^v, r_w^l)$.

Instead of using BoVW features to obtain image representations, as in previous work in multi-modal semantics, the model extracts (i.e. transfers) the pre-softmax layer from a trained convolutional neural network as the image representation. Specifically, a convolutional neural network similar to the one defined by Krizhevsky et al. (2012) is trained on a large-scale image recognition task, such as ILSVRC2012 (Russakovsky et al., 2015). We then freeze the trained parameters, remove the last network layer, and use the remaining final layer as a filter to compute a feature vector on arbitrary input images (we also experiment with earlier layers). For a given concept, we obtain images associated with words, labels or tags representing that particular concept. Each image is fed through the neural network, yielding an image representation. In other words, for each image, we do a forward pass through the network and take the pre-softmax layer as the image representation.

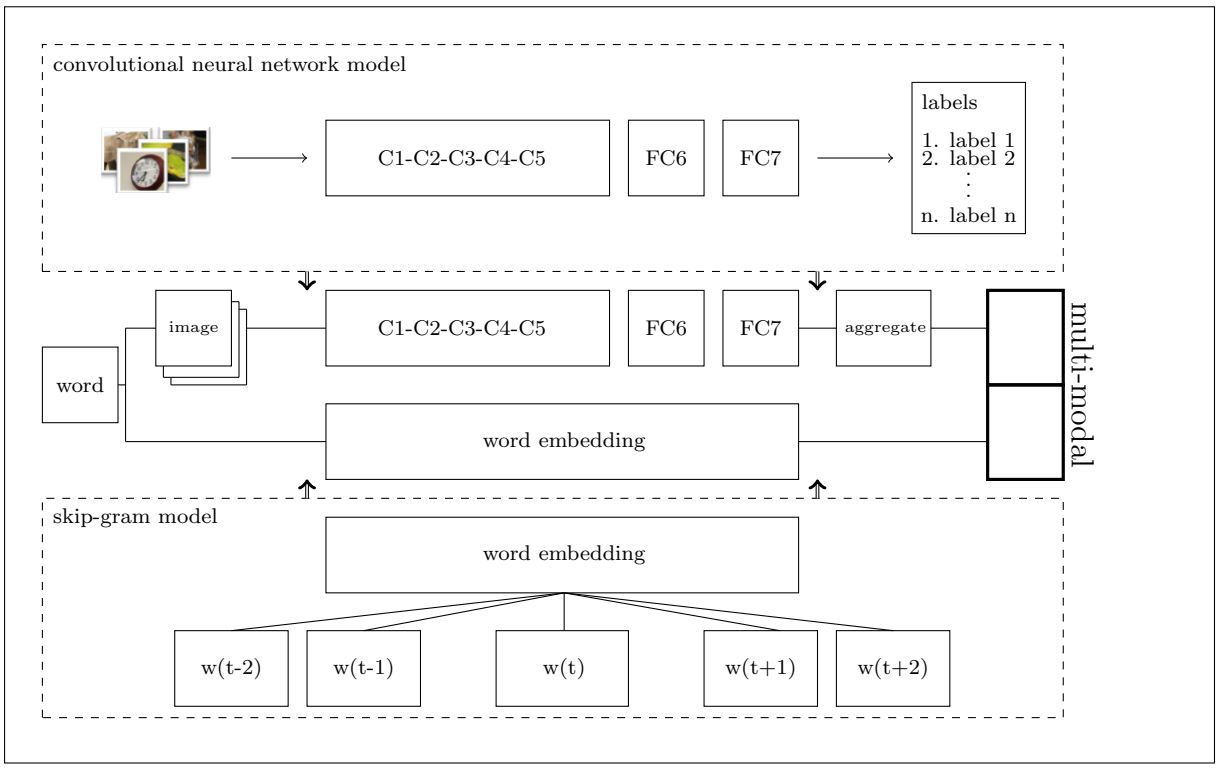

Figure 3.1: Computing multi-modal representations.

Two ways for aggregating image representations as visual representations are considered: taking the mean, or taking the elementwise maximum[1]. Following Bruni et al. (2014), we construct multi-modal semantic representations by concatenating the centered and L_2-normalized linguistic and perceptual feature vector representations r_w^l and r_w^v:

$$r_w^{mm} = \alpha \times r_w^l \;||\; (1-\alpha) \times r_w^v \tag{3.1}$$

where $||$ denotes the concatenation operator and α is an optional tuning parameter (without tuning, it is set to 0.5).

Two sets of experiments are performed with this model. First, we describe experiments with a neural network specifically designed for transfer learning due to Oquab et al. (2014), using two human-annotated image datasets, ImageNet and the ESP Game dataset. The transferred features are compared with the BoVW method to show that these novel representations lead to significant improvements. Second, since the network from the first experiment was not available for subsequent work in this thesis[2], we show that the same results hold for different adaptations, modifications or enhancements of the Krizhevsky et al. (2012) network. We also show that images from search engines such as Google

[1]This approach makes sense because modern convolutional neural networks tend to use rectified linear units. Computing a rectified function such as $x = max(0, x)$ implies setting negatives to zero, which means that extracted representations are relatively sparse (about 22% non-zero coefficients in typical CNNs, about 50% in a randomly initialized network). Taking the maximum over dense vectors would lead to a lot of noise, but since these representations are very sparse, and hence have less overlap in their components, this is less of a problem.

[2]Initial experiments were conducted during an internship at Microsoft Research, New York. As a result, both the visual and linguistic representations used in those experiments were not available for the remainder of the thesis.

Figure 3.2: Examples of *dog* in the ESP Game dataset.

Figure 3.3: Examples of *golden retriever* in ImageNet.

and Bing may be used without sacrificing any performance over ImageNet or the ESP Game dataset. The findings indicate that the method extends to different convolutional network architectures and to different sources of perceptual input. This latter finding corroborates results obtained by Fergus et al. (2005), who show that images from search engines yield representations competitive with directly human-annotated datasets. Using images from search engines rather than human-annotated datasets has the advantage that we can have much better coverage, while search engines return high quality images and support multiple languages.

3.2 Improving visual representations

In the current experiment, images are obtained either from ImageNet or the ESP Game dataset. Image representations are extracted by transfering the pre-softmax layer from the network of Oquab et al. (2014)[3]. This network was designed to show how image representations learned with CNNs on large-scale annotated datasets can efficiently be transferred to other computer vision tasks when training data is limited. The resultant image representations are aggregated into overall visual representations by taking the mean or elementwise maximum, and fused with skip-gram representations as described in the previous section.

Representation quality is evaluated using the WordSim353 (Finkelstein et al., 2002) and MEN (Bruni et al., 2012) datasets, as described in Section 2.1.3. The main purpose of the current experiment is to examine whether CNN representations outperform BoVW representations.

3.2.1 Image sources, selection and processing

Experiments were conducted using two distinct sources of images to compute the visual representations: ImageNet and the ESP Game dataset (see also Section 2.2.3.2). The ImageNet dataset (Deng et al., 2009) is a large-scale ontology of images organized according to the hierarchy of WordNet (Miller, 1995). The dataset was constructed by manually re-labelling candidate images collected using web searches for each WordNet synset. The images tend to be of high quality with the designated object roughly centered in the image. The copy of ImageNet used in this experiment contains about 12.5 million images organized in 22K synsets, which implies that ImageNet covers only a small fraction of the 117K synsets that exist in WordNet.

The ESP Game dataset (von Ahn and Dabbish, 2004) was collected as a "game with a purpose", in which two players must independently and rapidly agree on a correct word label for randomly selected images. Once a word label has been used sufficiently frequently for a given image, that word is added to the image's tags. This dataset contains 100K images, but with every image having on average 14 tags, that amounts to a coverage of 20,515 words. Since players are encouraged to produce as many terms per image as they can think of, the dataset's increased coverage is at the expense of accuracy in the word-to-image mapping: a dog in a field with a house in the background might be a *golden retriever* in ImageNet and could have tags *dog, golden retriever, grass, field, house, door* in the ESP Dataset. In other words, images in the ESP dataset do not make a distinction between objects in the foreground and in the background, or between the relative size of the objects (tags for images are provided in a random order, so the top tag is not necessarily the best one).

Figures 3.2 and 3.3 show typical examples of images belonging to these datasets. Both datasets have attractive properties. On the one hand, ImageNet has higher quality images with better labels and a more natural annotation strategy. On the other hand, the ESP dataset has better coverage on the MEN task, which was specifically designed to be covered by the ESP dataset.

3.2.1.1 Image selection

Since ImageNet follows the WordNet hierarchy, selecting images at all nodes below all subtrees of a node's senses is not feasible: for high-level concepts such as *entity*, *object* or *animal*, we would have to include almost all images in the dataset. Doing so is both computationally expensive and unlikely to improve the results. For this reason, we randomly sample up to N distinct images from the subtree associated with each concept. When this returns less than N images, we attempt to increase coverage by sampling images from the subtree of the concept's hypernym instead. In order to allow for a fair comparison, we apply the same method of sampling up to N on the ESP Game dataset. In this experiment, $N = 1,000$. We used the WordNet lemmatizer from NLTK (Bird et al., 2009) to lemmatize tags and concept words so as to further improve the dataset's coverage.

3.2.1.2 Image processing

The ImageNet images were preprocessed as described by Krizhevsky et al. (2012). The largest centered square contained in each image is resampled to form a 256×256 image.

[3] http://www.di.ens.fr/willow/research/cnn/

The CNN input is then formed by cropping 16 pixels off each border and subtracting 128 to the image components. The ESP Game images were preprocessed slightly differently because we do not expect the objects to be centered. Each image was rescaled to fit inside a 224×224 rectangle. The CNN input is then formed by centering this image into the input field with zero-padding and subtracting 128 from the image components (i.e. centering RGB values at zero).

Each image is fed through the first seven layers of the convolutional network defined by Krizhevsky et al. (2012) and adapted by Oquab et al. (2014). This network takes 224×224 pixel RGB images and applies five successive convolutional layers followed by three fully connected layers. Its eighth and last layer produces a vector of 1512 scores associated with 1000 categories of the ILSVRC-2012 challenge and the 512 additional categories selected by Oquab et al. (2014). This network was trained using about 1.6 million ImageNet images associated with these 1512 categories. We then freeze the trained parameters, remove the last network layer, and use the remaining seventh layer as a filter to compute a 6144-dimensional feature vector on arbitrary images.

BoVW features are obtained by computing DSIFT descriptors using VLFeat (Vedaldi and Fulkerson, 2008) for both datasets. These descriptors are subsequently clustered using mini-batch k-means (Sculley, 2010) with 100 clusters (i.e., $k = 100$, which was found to work best in initial experiments). That is, each image is represented by a bag of clusters (visual words) quantized as a 100-dimensional feature vector. In the BoVW case, we do not experiment with different aggregation methods and simply take the mean, as is done in previous work in multi-modal semantics.

3.2.2 Linguistic representations

For our linguistic representations we extract 100-dimensional continuous vector representations using the log-linear skip-gram model of Mikolov et al. (2013a) trained on a corpus consisting of the 400M word Text8 corpus of Wikipedia text[4] together with the 100M word British National Corpus (Leech et al., 1994). The skip-gram model learns high quality semantic representations based on the distributional properties of words in text, and outperforms standard distributional models on a variety of semantic similarity and relatedness tasks. Better performance has been reported for the linguistic component than what we achieve, including by standard distributional models e.g. by Bruni et al. (2014), but we are primarily interested here in the relative improvement compared to uni-modal linguistic representations, rather than in obtaining state-of-the-art performance on a given task.

3.2.3 Evaluation

Since multi-modal representations rely on different modalities, it often occurs that data relevant to a given word is only available for one of, or a subset of, the modalities. That is, the coverage per modality for a given word may vary, depending on the source corpora. Consequently, multi-modal semantic models are often evaluated on subsets of the full datasets. Standard text-based distributional semantic models can also have coverage issues, but this is much more pertinent in the visual modality, given that relevant high-quality images may be hard to find or non-existent for abstract nouns (e.g. "democracy"),

[4] http://mattmahoney.net/dc/textdata.html

	ImageNet				ESPGame			
	MEN	MEN*	W353	W353*	MEN	MEN*	W353	W353*
LINGUISTIC	0.64	0.62	0.57	0.51	0.64	0.62	0.57	0.51
BOVW	-	0.40	-	0.30	0.17	0.35	-	0.38
CNN-MEAN	-	0.64	-	0.32	0.51	0.58	-	0.44
CNN-MAX	-	0.63	-	0.30	0.20	0.57	-	0.56
MM-BOVW	0.64	0.64	0.58	0.55	0.64	0.63	0.58	0.52
MM-MEAN	**0.70**	**0.72**	0.59	0.56	**0.71**	0.69	0.59	0.55
MM-MAX	0.67	0.71	**0.60**	**0.57**	0.65	**0.70**	**0.60**	**0.61**

Table 3.1: Results with the Oquab et al. (2014) network.

as well as for many verbs (e.g. "finding" or "being"), adjectives (e.g. "specific" or "precise") and adverbs (e.g. "really" or "very").

This problem makes comparisons difficult: multi-modal representations are often evaluated on an unspecified subset of datasets, making it impossible to directly compare the reported scores. In this experiment, scores are reported on the full WordSim353 (W353) dataset by setting the visual representation r_w^v to zero for concepts without images. We also report scores on the subset of pairs for which both concepts have both ImageNet and ESP Game images available. The MEN dataset was constructed in such a way that only frequent words with at least 50 images in the ESP Game dataset were included in the evaluation pairs. It is much larger than WordSim353, with 3000 words pairs consisting of 751 individual words. Although MEN was constructed so as to have at least a minimum amount of images available in the ESP Game dataset for each concept, this is not the case for ImageNet. Hence, similarly to WordSim353, we also evaluate on a subset for which images are available in both datasets. In both cases, the covered subset is marked by an asterisk (*), so the covered subset of MEN is MEN* and that of W353 is W353*.

3.2.4 Results

Model performance is measured in terms of their Spearman ρ correlation with human similarity and relatedness ratings. The similarity score between the representations associated with a pair of words is calculated using the cosine similarity. We evaluate uni-modal linguistic, uni-modal visual and multi-modal representations. Scores are reported for BoVW, as well as mean-aggregated (CNN-Mean) and max-aggregated (CNN-Max) visual representations extracted from the CNN, if available. For all datasets we report the scores obtained by multi-modal representations. Since we have full coverage with the ESP Game dataset on MEN, we are able to report visual representation scores for the entire dataset as well. The results can be seen in Table 3.1.

3.2.4.1 Representation quality

In all cases, CNN-extracted visual representations perform better or as good as BoVW representations. This confirms the motivation outlined above: by applying state-of-the-art approaches from computer vision to multi-modal semantics, we obtain a signficant

performance increase over standard multi-modal models. Higher-quality perceptual input leads to better-performing multi-modal representations. In all cases multi-modal models with CNNs outperform multi-modal models with BOVW, occasionally by a substantial margin. In all cases, multi-modal representations outperform purely linguistic vectors that were obtained using a state-of-the-art supervised distributed representation learning approach. This re-affirms the importance of multi-modal representations for distributional semantics.

3.2.4.2 The contribution of images

Since the ESP Game images come with a multitude of word labels, one may wonder whether a performance increase of multi-modal models based on that dataset comes from the images themselves, or from overlapping word labels. It might also be possible that similar concepts are more likely to occur in the same image, which encodes relatedness information without necessarily taking the image data itself into account. For instance, Hill and Korhonen (2014) use ESP Game tags in multi-modal models as their perceptual input. In short, it is natural to ask whether the performance gain is due to image data or due to word label associations. We conclusively show that the image data matters in two ways: (a) using a different dataset (ImageNet) with the same method also results in a performance boost, and (b) using higher-quality image features (i.e., CNN over BovW) on the ESP game images increases the performance boost without changing the association between word labels.

3.2.4.3 Image datasets

Another factor that could have a large impact on performance is the source image dataset. Although the scores for the visual representation in some cases differ, performance of multi-modal representations remains close for both image datasets. This implies that our method is robust over different datasets. It also suggests that it is beneficial to train on high-quality datasets like ImageNet and to subsequently generate embeddings for other sets of images like the ESP Game dataset that are more noisy but have better coverage.

3.2.4.4 Error analysis

One way to qualitatively evaluate the results is to look at differences between scores. Table 3.2 shows the top 5 worst scoring word pairs for the two datasets using CNN-Mean multi-modal vectors. The MEN words *potatoes* and *tomato* probably have low quality ImageNet-derived representations, because they occur often in the bottom pairs for that dataset. The MEN words *dessert*, *bread* and *fruit* occur in the bottom 5 for both image datasets, which implies that their linguistic representations are probably not very good. For WordSim353, the bottom pairs on ImageNet could be said to be similarity mistakes, while the ESP Game dataset contains more relatedness mistakes (*king* and *queen* would evaluate similarity, while *stock* and *market* would evaluate relatedness). It is difficult to say anything conclusive about this discrepancy, but it is clearly something worth exploring further.

W353-Relevant

ImageNet				ESP Game			
word1	word2	sys	gold	word1	word2	sys	gold
cell	phone	0.27	0.78	law	lawyer	0.33	0.84
discovery	space	0.10	0.63	monk	slave	0.58	0.09
closet	clothes	0.22	0.80	gem	jewel	0.41	0.90
king	queen	0.26	0.86	stock	market	0.33	0.81
wood	forest	0.13	0.77	planet	space	0.32	0.79

MEN-Relevant

ImageNet				ESP Game			
word1	word2	sys	gold	word1	word2	sys	gold
bread	potatoes	0.88	0.34	bread	dessert	0.78	0.24
fruit	potatoes	0.80	0.26	jacket	shirt	0.89	0.34
dessert	sandwich	0.76	0.23	fruit	nuts	0.88	0.33
pepper	tomato	0.79	0.27	dinner	lunch	0.93	0.37
dessert	tomato	0.66	0.14	dessert	soup	0.81	0.23

Table 3.2: The top 5 best and top 5 worst scoring pairs with respect to the gold standard.

3.3 Comparing architectures and data sources

The alternative to obtaining images from ImageNet and the ESP Game dataset is to get them from an image search engine such as Google Images or Bing Images. This alleviates some of the problems that we had with selecting relevant images in ImageNet, such as the fact that there are no images for "dog" but only for its synsets (e.g., "golden retriever"), meaning that we would have to sample from the potentially noisy set of hyponyms. It also does not suffer from some of the problems the ESP Game dataset has, such as the fact that it is unclear whether a tag was placed because it has something to do with what occurs in the picture, as opposed to something both humans inferred from the picture, or some small thing that occurs in the background. Search engines return ranked results, meaning that they are more likely to return the best first, as opposed to both ImageNet and the ESP Game dataset, which make no such distinction. Furthermore, provided the search engine functions well, we can retrieve images for pretty much any search term—dramatically increasing coverage. There are two obvious downside to using search engines: we are essentially letting them do the job of selecting relevant images for us, and they make use of proprietary algorithms the details of which are not accessible.

While the previous section shows that one particular type of neural network performs better than BoVW, it is unclear whether this is a result of the particular network architecture, or whether it is a more general finding that extends to different architectures. In order to examine this, we compare the AlexNet representations to other convolutional neural network representations.

In this set of experiments, we examine the following questions:

- How important is the source of images? Is there a difference between search engines and manually annotated data sources?
- Does the improved performance over bag of visual words extend to different convolutional network architectures, or is it specific to Krizhevsky's AlexNet? Do others work even better?

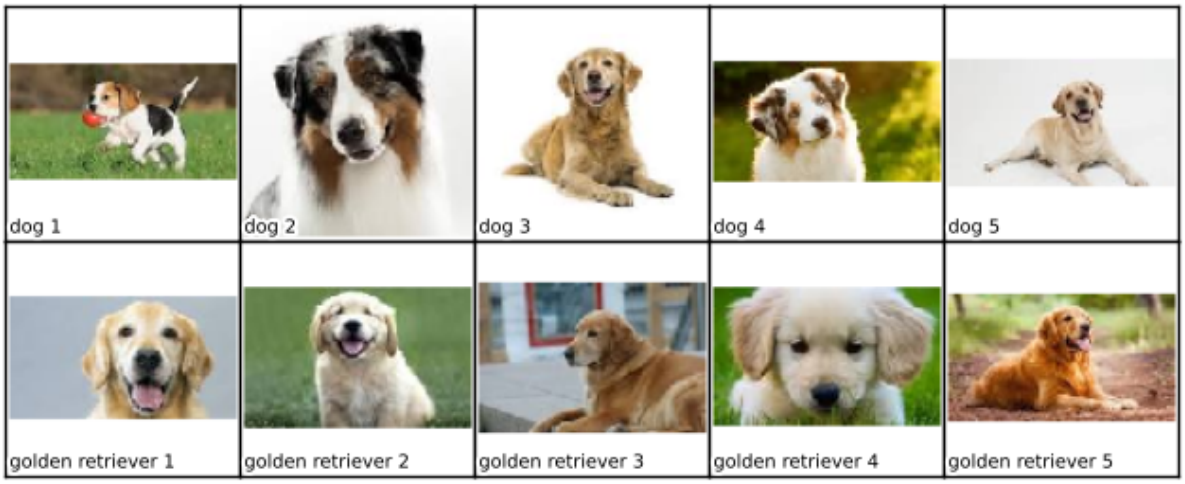

Figure 3.4: Examples of *dog* and *golden retriever* from Google Images.

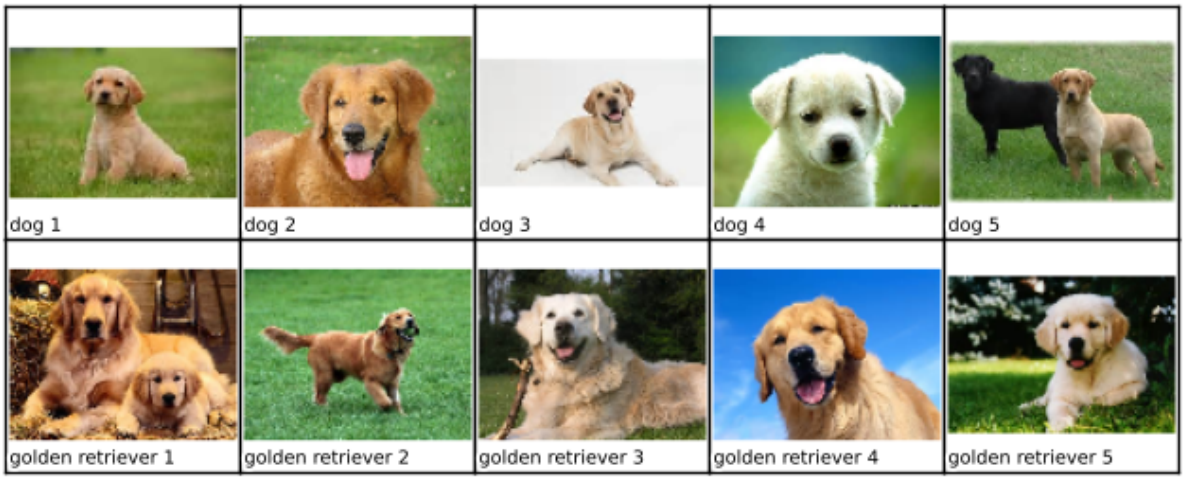

Figure 3.5: Examples of *dog* and *golden retriever* from Bing Images.

- Do these findings extend to languages other than English?

3.3.1 Evaluation

In this experiment we evaluate on MEN (Bruni et al., 2012) and SimLex-999 (Hill et al., 2015). The reason we no longer use WordSim353 in these experiments is because of its inadequacies. SimLex-999 was designed in part to address the issues with that dataset, as discussed in the preceding chapter, so it is a better choice of dataset. As in the previous experiment, we do not necessarily have full coverage. We thus report results on the maximally covered subset per data source, as well as the common covered subset. That is, while in the previous experiment we set non-existent representations to zero to compute full coverage, here we evaluate performance simply on the subset of images that we have coverage for. The reason for making the distinction along slightly different lines is that search engines have full coverage. Hence, visual representations based on Google would never need to be set to zero, while the same is obviously not the case for ImageNet or the ESP Game dataset, which makes for an unfair comparison. Like before, the subset

	MEN (3000)	SimLex (999)
Google	3000	999
Bing	3000	999
ImageNet	1326	373
ESPGame	2927	833
Common subset	1310	360

Table 3.3: Coverage on MEN and SimLex for our data sources.

	AlexNet	**GoogLeNet**	**VGGNet**
ILSVRC winner	2012	2014	2015
Number of layers	7	22	19
Number of parameters	~60 million	~6.7 million	~144 million
Receptive field size	11×11	3×3	$1 \times 1, 3 \times 3, 5 \times 5$
Fully connected layers	Yes	No	Yes

Table 3.4: Network architectures. Layer counts only include layers with parameters.

that has common coverage is marked with an asterisk (*). In other words, MEN is known as MEN* and SimLex as SimLex*, for the subsets of pairs where images exist in all data sources (Google, Bing, ESP Game and ImageNet). Coverage numbers are reported in Table 3.3.

3.3.2 CNN implementations

We obtain image representations for three different convolutional network architectures: AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015) and VGGNet (Simonyan and Zisserman, 2015). Image representations are turned into an overall word-level visual representation by either taking the mean or the elementwise maximum of the relevant image representations. All three networks are trained to maximize the multinomial logistic regression objective on the ImageNet classification task using mini-batch gradient descent with momentum (Equation 2.15). In this section, we describe the network architectures and their properties.

AlexNet In this case, we actually use the CaffeNet reference model, which is almost identical to AlexNet (described in the previous chapter), with the difference that it is not trained with relighting data-augmentation, and that the order of pooling and normalization layers is switched (in CaffeNet, pooling is done before normalization, instead of the other way around). Performance of CaffeNet is very similar to AlexNet, with some modifications to make it more stable.

	Google	Bing	ImageNet	ESP Game
Type	Search	Search	Database	Tagged
Annotation	Automatic	Automatic	Human	Game
Coverage	Unlimited	Unlimited	Limited	Limited
Multi-lingual	Yes	Yes	No	No
Sorted	Yes	Yes	No	No
Tag specificity	Unknown	Unknown	Specific	Loose

Table 3.5: Sources of image data.

GoogLeNet The ILSVRC 2014 ImageNet classification challenge-winning Goog-LeNet (Szegedy et al., 2015) uses "inception modules" as a network-in-network method (Lin et al., 2013) for enhancing model discriminability for local patches within the receptive field. It uses much smaller receptive fields and explicitly focuses on efficiency: while it is much deeper than AlexNet, it has fewer parameters. Its architecture consists of two convolutional layers, followed by inception layers that culminate into an average pooling layer that feeds into the softmax decision. That is, it has no fully connected layers. Dropout (Srivastava et al., 2014) is only applied on the final layer. All connections use rectifiers.

VGGNet The ILSVRC 2015 ImageNet classification challenge was won by VGGNet (Simonyan and Zisserman, 2015). Like GoogLeNet, it is much deeper than AlexNet and uses smaller receptive fields. It has many more parameters than the other networks. The architecture is similar to AlexNet, in that it consists of a series of convolutional layers followed by the fully connected ones, except that it has more convolutional layers. All layers are rectified and dropout is applied to the first two fully connected layers.

These networks were selected because they are very well-known in the computer vision community. They exhibit interesting qualitative differences in terms of their depth (i.e., the number of layers), the number of parameters, regularization methods and the use of fully connected layers. They have all been winning network architectures in the ILSVRC ImageNet classification challenges. See Table 3.4 for a summary of some of the differences.

3.3.3 Linguistic representations

Instead of the relatively small 100-dimensional vector representations used in the previous experiment, we use the much higher-quality 300-dimensional continuous skip-gram vectors Mikolov et al. (2013b) trained on a larger corpus consisting of a recent dump of English Wikipedia combined with newswire text. A shellscript by Mikolov et al. (2013b)[5] was used to obtain this corpus.

[5]The demo-train-big-model-v1.sh script from http://word2vec.googlecode.com.

	Arch.	AlexNet				GoogLeNet				VGGNet			
	Agg.	Mean		Max		Mean		Max		Mean		Max	
Source	Type/Eval	SL	MEN	SL	MEN	SL	MEN	SL	MEN	SL	MEN	SL	MEN
Wikipedia	Text	.310	.682	.310	.682	.310	.682	.310	.682	.310	.682	.310	.682
Google	Visual	.340	.503	.334	.513	.358	.495	.367	.501	.342	.512	.332	.494
	MM	.380	.711	.370	.719	.379	.711	.365	.716	.380	.714	.365	.716
Bing	Visual	.325	.567	.316	.554	.310	.526	.303	.520	.304	.551	.289	.507
	MM	.373	.727	.360	.725	.364	.723	.350	.724	.361	.727	.349	.719
ImageNet	Visual	.313	.561	.313	.561	.341	.540	.411	.603	.404	.584	.401	.578
	MM	.362	.713	.362	.713	.373	.719	.401	.731	.427	.727	.412	.723
ESPGame	Visual	.018	.448	.026	.376	.063	.487	.050	.434	.125	.506	.106	.451
	MM	.208	.686	.187	.672	.243	.700	.246	.696	.269	.708	.260	.698

Table 3.6: Performance on covered datasets.

3.3.4 Image search engines

We experiment with two search engines: Google Images[6] and Bing Images[7]. These two image search engines are widely known to be state-of-the-art image retrieval systems. Examples of images they return for *dog* and *golden retriever* can be found in Figures 3.4 and 3.5. This work is the first to see whether multi-modal semantics can be performed using search engine results instead of datasets such as ImageNet and the ESP Game dataset which explicitly rely on human annotators (either directly, or through a "game with a purpose"). See Table 3.5 for a comparison of the data sources.

The results obtained by these models are not directly comparable to the results in the previous experiment: they use a different network architecture and different linguistic representations. The primary objective of this experiment is not to compare these network architectures to the one of Oquab et al. (2014), but rather to show that the same idea extends to other network architectures and to different sources of images, with different linguistic representations.

3.3.5 Selecting and processing images

Selecting images for Google and Bing is done through their respective APIs: we query for the desired word and obtain the top 10 images. In the case of ImageNet and the ESP Game dataset, images are not ranked and vary greatly in number of tags: for some words there is only a single image, while others have thousands. For ImageNet, like before, if a word has no associated images for any of its hyponyms, we apply the same method to the hyponyms of its hypernyms—in other words, we go up one level in the hierarchy and see if that yields any relevant images in the subtree. We subsequently randomly sample 100 images associated with the word and obtain semi-ranked (i.e., the ones ranked most typical/closest to the mean by this method) results by selecting the 10 images closest to the median representation as the relevant image representations. We use the same method for the ESP Game dataset. In all cases, images are resized and center-cropped to ensure that they are the correct size for the given network architecture. Note that these images are randomly sampled again for this experiment and have a different sampling strategy, and so are different from the images used in the previous experiment.

[6]https://images.google.com/

[7]https://www.bing.com/images

3.3.6 Results

Table 3.6 shows the results. The first row repeats for each architecture the results for the text-based linguistic representations that were obtained from Wikipedia. For each of the three architectures, we evaluate on SimLex (SL) and MEN, using either the mean (Mean) or elementwise maximum (Max) method for aggregating image representations into visual ones. For each data source, we report results for the visual representations, as well as for the multi-modal representations that fuse the visual and textual ones together. As we can see, performance across architectures is stable: we have had to report results up to three decimal points to show the difference in performance in some cases.

First of all, note that the visual representations from Google, Bing and ImageNet outperform linguistic representations on SimLex, except for Bing with VGGNet. In all cases, multi-modal representations outperform linguistic ones, on both datasets, except for ESPGame and MEN. There is not a huge difference between mean or max aggregation, although the former works slightly better on SimLex and the latter on MEN. Google, Bing and ImageNet obtain the best results, ESPGame is not very good. ESP's bad performance is somewhat surprising, especially given that it performed reasonably well in the previous experiment. The reason for this is most likely the different sampling method: taking the mean over a thousand images, apparently, works much better than only selecting ten images. The highest score obtained on SimLex is 0.427 by multi-modal VGGNet on ImageNet with mean aggregation. It is interesting to see that VGGNet scores particularly highly when using ImageNet, which might indicate that it is specialized on this dataset more than the other two networks (after all, it was trained on a subset of ImageNet). The highest score obtained on MEN is 0.731 by GoogLeNet, again using ImageNet, with max aggregation. In fact, the best visual representations are from ImageNet. Although this shows the strength of that data source, this should not be overstated: Google and Bing were within 0.02 of its MEN score, and within 0.06 for SimLex. Similarly, results were very close across architectures, with sometimes very small differences. These numbers indicate the robustness of the approach: we find that multi-modal representation learning yields better performance across the board: for different network architectures, different data sources and different aggregation methods. If computational efficiency or memory usage are issues, then GoogLeNet or AlexNet are the best choices. If we have the right coverage, then ImageNet will probably get us the best results, especially if we can use VGGNet. However, coverage is often the main issue, in which case search engines like Google and Bing yield images that are of similarly high quality, that obtain almost identical performance to the manually annotated ImageNet.

We can make similar comparisons, but while making sure that we are looking at the same common subsets of the datasets. This particularly supports comparisons across the different data sources. Results on the common covered subset can be found in Table 3.7. Some of our findings are very similar: performance does not vary greatly amongst network architectures, nor does the aggregation method have a big impact. The main observation is that ImageNet performance has dropped relative to Google and Bing, except when using VGGNet. This is interesting, because it once again shows how VGGNet and ImageNet combine remarkably well. Performance for ESP goes up a small amount. This indicates that Google and Bing were unfairly punished by having better coverage: better coverage means having images for more abstract concepts that have less clear images; if we take these concepts out, performance increases and rises above that of ImageNet. It appears that Bing performs slightly better than Google.

	Arch.	AlexNet				GoogLeNet				VGGNet			
	Agg.	Mean		Max		Mean		Max		Mean		Max	
Source	Type/Eval	S*	MN*	S*	MN*	S*	MN*	S*	MN*	S*	MN*	S*	MN*
Wikipedia	Text	.310	.682	.310	.682	.310	.682	.310	.682	.310	.682	.310	.682
Google	Visual	.406	.549	.402	.552	.420	.570	.434	.579	.430	.576	.406	.560
	MM	.366	.691	.344	.693	.366	.701	.342	.699	.378	.701	.341	.693
Bing	Visual	.431	.613	.425	.601	.410	.612	.414	.603	.400	.611	.398	.569
	MM	.384	.715	.355	.708	.374	.725	.343	.712	.363	.720	.340	.705
ImageNet	Visual	.316	.560	.316	.560	.347	.538	.423	.600	.412	.581	.413	.574
	MM	.348	.711	.348	.711	.364	.717	.394	.729	.418	.724	.405	.721
ESPGame	Visual	.037	.431	.039	.347	.104	.501	.125	.438	.188	.514	.125	.460
	MM	.179	.666	.147	.651	.224	.692	.226	.683	.268	.697	.222	.688

Table 3.7: Performance on the common covered subsets of the datasets (S* = SimLex*, MN* = MEN*).

		CNN-Mean		CNN-Max	
	Layer(s)	SimLex	MEN	SimLex	MEN
Visual	P5	0.316	0.463	0.315	0.416
	FC6	0.333	0.499	0.312	0.515
	FC7	0.340	0.503	0.334	0.513
	P5+FC6	0.332	0.502	0.328	0.508
	FC6+FC7	0.339	0.507	0.328	0.525
	P5+FC6+FC7	0.339	0.508	0.335	0.523
Multi-modal	P5	0.367	0.712	0.354	0.710
	FC6	0.378	0.715	0.359	0.721
	FC7	0.380	0.711	0.370	0.719
	P5+FC6	0.374	0.719	0.359	0.721
	FC6+FC7	0.380	0.715	0.366	0.722
	P5+FC6+FC7	0.376	0.718	0.364	0.723

Table 3.8: Google Images dataset results on different layers of an AlexNet.

3.3.6.1 CNN layers

In each of the previous experiments, we have only used the pre-softmax layer, FC7. It has been found, however, that other layers in the network also have good properties for usage in transfer learning (Girshick et al., 2014; Yosinski et al., 2014). Here, we experiment with transfering other layers than FC7, using images from Google and the AlexNet architecture. Either individual layers were used, or layers were combined with other CNN layers by concatenating the normalized layers. See Table 3.8 for the results. Scores are given with three decimals to show that there are in fact differences between the layers, but that these are very small. These findings clearly show that using different layers, or adding additional information by concatenating layers, does not change performance much. FC6 and FC7 are the best-performing individual layers. Concatenating layers does not appear to add much new information into the representation.

		SimLex-EN	SimLex-IT
Wikipedia	Linguistic	.310	.179
Google	Visual	.340	.231
	Multi-modal	.380	.231
Bing	Visual	.325	.212
	Multi-modal	.373	.227

Table 3.9: Performance on two languages for SimLex.

3.3.6.2 Multi-lingual applicability

Although there are some indicators that visual representation learning extends to other languages, particularly in the case of bilingual lexicon learning (Bergsma and Van Durme, 2011), this has not been shown directly on the same set of human similarity and relatedness judgments. Unfortunately, there are no non-English versions of ImageNet and the ESP Game dataset. However, Google and Bing are available in different languages, which allows us to examine this question more closely: do the same findings hold for other languages?

We compare results on the original English SimLex and on the Italian version of SimLex, due to Leviant and Reichart (2015). There is no non-English version of MEN available. Linguistic representations are trained on recent dumps of the English and Italian Wikipedia. Images are obtained from Google and Bing by explicitly specifying the language, setting it either to English or to Italian. Since we know that performance across architectures is very similar, we use AlexNet representations.

The results can be found in Table 3.9. We find that the same pattern as before emerges: in all cases, visual and multi-modal representations outperform linguistic ones. For all three types of representations (linguistic, visual and multi-modal), the Italian version of SimLex appeared more difficult. Although somewhat preliminary, these results are a good indicator that multi-modal semantics can just as fruitfully be applied to languages other than English.

3.4 Conclusion

This chapter presented a novel approach to visually grounded multi-modal semantics. Instead of the traditional BoVW method, deep convolutional neural network-extracted features were used. Such features obtained high results on well-known and widely-used semantic relatedness benchmarks, with increased performance both in the separate visual representations and in the combined multi-modal representations.

These results indicate that such multi-modal representations outperform both linguistic and standard bag of visual words multi-modal representations, and furthermore, that the approach is robust and that CNN-extracted features from separate image datasets can succesfully be applied to semantic relatedness. In addition to improving multi-modal representations, these findings indicate that the source of this improvement is due to image data and is not simply a result of word label associations, which was shown by

obtaining performance improvements on two different image datasets, and by obtaining higher performance with higher-quality image features on the ESP game images, without changing the association between word labels.

Furthermore, we showed that performance is robust across different neural network architectures and that mean and maximum aggregation may be used with similar performance. Concatenation yields good results, but has the limitation that it always requires the visual modality to be present, in contrast with early fusion. It appears, however, that images for abstract concepts are at least somewhat meaningful in multi-modal word representations, given how the many abstract concepts included in the evaluation datasets did not lead to detrimental performance. In addition, it was also found that image search engines may be used, and that these in fact yield higher-quality images than even ImageNet, a very carefully human-annotated image dataset. Image search engines provide excellent coverage and are multi-lingual. The fact that they return high-quality images of similar or even better quality than ImageNet opens up all kinds of areas for research, some of which will be explored in this thesis.

CHAPTER 4

Applications of CNN representations

The previous chapter showed that CNN-extracted feature representations perform very well on instrinsic evaluations of representational quality, as measured by correlation with human similarity and relatedness ratings. In this chapter, we explore the possibilities for exploiting the qualities of these visual representations for other tasks in natural language processing, namely, lexical entailment and bilingual lexicon induction. These applications matter for two reasons: first, they show that extra-linguistic information can be useful beyond learning higher-quality grounded representations that perform better at mirroring human similarity ratings; and second, entailment and translation are two core problems for natural language processing.

4.1 Lexical entailment

Automatic detection of lexical entailment—determining whether one lexical item logically entails another lexical item—is useful for a number of NLP tasks, including search query expansion (Shekarpour et al., 2013), recognizing textual entailment (Garrette et al., 2011), metaphor detection (Mohler et al., 2013), and text generation (Biran and McKeown, 2013). Given two semantically related words, a key aspect of detecting lexical entailment, or the hyponym-hypernym relation, is the *generality* of the hypernym compared to the hyponym. For example, *bird* is more general than *eagle*, having a broader intension and a larger extension. This property has led to the introduction of lexical entailment measures that compare the entropy of distributional word representations, under the assumption that a more general term has a higher-entropy distribution (Herbelot and Ganesalingam, 2013; Santus et al., 2014).

The hypothesis in this chapter is that visual representations can be particularly useful for lexical entailment detection. The intuition is that the set of images returned for *animal* will consist of pictures of different kinds of animals, the set of images for *bird* will consist of pictures of different birds, while the set for *owl* will mostly consist only of images of owls, as can be seen in Figure 4.1.

There have been approaches, using the linguistic modality, that are similar to what is proposed here. The most closely related works are by Herbelot and Ganesalingam (2013) and Santus et al. (2014), both of whom use unsupervised distributional generality mea-

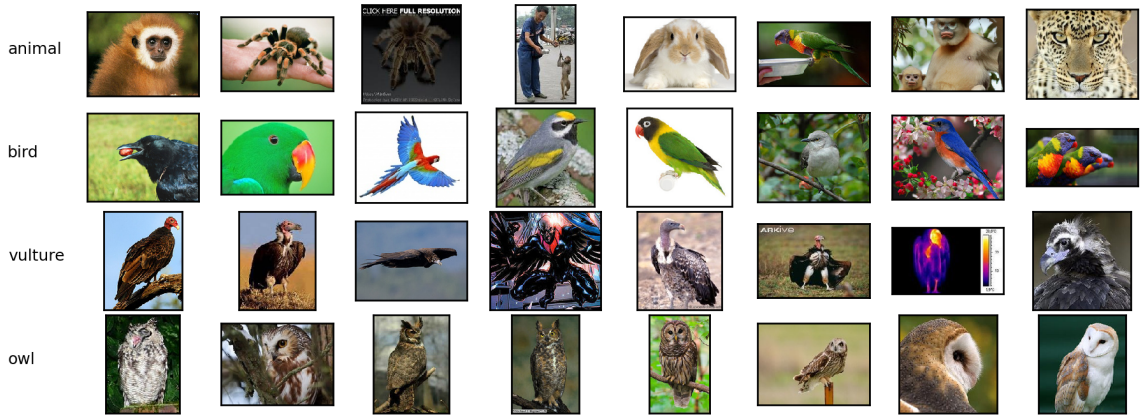

Figure 4.1: Example of how *vulture* and *owl* are less dispersed concepts than *bird* and *animal*, according to images returned by Google image search.

sures to identify the hypernym in a hyponym/hypernym pair. Herbelot and Ganesalingam (2013) use Kullback-Leibler (KL) divergence to compare the probability distribution of context words given a term, to the background probability distribution of context words. More specific terms are hypothesised to have more informative distributions and therefore higher KL divergence. Santus et al. (2014) use the median entropy of the probability distributions associated with the 50 top-weighted context words for a term as a measure of information content. More specific terms are hypothesised to occur with context words having lower-entropy distributions, again because the contexts are more informative. Like the currently proposed method, these measures focus on term generality and must be combined with a semantic similarity measure to distinguish hyponym/hypernym pairs from e.g. meronym/holonym pairs or pairs of unrelated words in which one is more general.

A number of weakly-supervised methods for hypernym detection have been proposed as well, which compare the top-weighted contexts of two terms (Weeds et al., 2004; Clarke, 2009; Kotlerman et al., 2010; Lenci and Benotto, 2012; Rei and Briscoe, 2014) or measure the semantic coherence of non-overlapping contexts (Rimell, 2014). Good results have also been achieved with fully-supervised classifiers using as features the concatenated term vectors and/or some of the weakly supervised measures (Baroni et al., 2012; Rimell, 2014; Weeds et al., 2014), or the difference between the term vectors and a set of reference vectors (Turney and Mohammad, 2015). Fu et al. (2014) learn a set of linear projections of neural word embeddings onto their hypernyms.

The intuition that visual representations may be useful for detecting lexical entailment has also been exploited by Deselaers and Ferrari (2011), who find that concepts and categories with narrower intensions and smaller extensions tend to have less visual variability on ImageNet. This notion is extended to Google image search results in this thesis, and applied to the lexical entailment task.

In what follows, three different vision-based methods are introduced for measuring term generality on the semantic tasks of hypernym detection and hypernym directionality.

4.1.1 Approach

We use two standard evaluations for lexical entailment: hypernym directionality, where the task is to predict which of two words is the hypernym; and hypernym detection, where the task is to predict whether two words are in a hypernym-hyponym relation

BLESS	turtle—animal	1
WBLESS	owl—creature	1
	owl—vulture	0
	animal—owl	0
BIBLESS	owl—creature	1
	owl—vulture	0
	animal—owl	-1

Table 4.1: Examples for evaluation datasets.

(Weeds et al., 2014; Santus et al., 2014). We also introduce a third, more challenging, evaluation that combines detection and directionality.

For the directionality experiment, we evaluate on the hypernym subset of the well-known BLESS dataset (Baroni and Lenci, 2011), which consists of 1337 hyponym-hypernym pairs. In this case, it is known that the words are in an entailment relation and the task is to predict the directionality of the relation. BLESS data is always presented with the hyponym first, so we report how often our measures predict that the second term is more general than the first.

For the detection experiment, we evaluate on the BLESS-based dataset of Weeds et al. (2014), which consists of 1168 word pairs and which we call WBLESS. In this dataset, the positive examples are hyponym-hypernym pairs. The negative examples include pairs in the reversed hypernym-hyponym order, as well as holonym-meronym pairs, co-hyponyms, and randomly matched nouns. Accuracy on WBLESS reflects the ability to distinguish hypernymy from other relations, but does not require detection of directionality, since reversed pairs are grouped with the other negatives.

For the combined experiment, we assign reversed hyponym-hypernym pairs a value of -1 instead of 0. We call this more challenging dataset BIBLESS. Examples of pairs in the respective datasets can be found in Table 4.1.

Images for the words in the evaluation datasets are obtained from Google Images, and for each image layer FC7 of AlexNet is extracted as the image representation. Thus, this approach is an instance of deep transfer learning as well; that is, a deep learning representation trained on one task (image classification) is used to make predictions on a different task (conceptual generality).

4.1.1.1 Generality measures

This section introduces three measures that can be used to calculate the generality of a set of images. The image *dispersion* d of a concept word w is defined as the average pairwise cosine distance between all image representations $\{r_w^{img(1)}, ..., r_w^{img(n)}\}$ of the set of images returned for w:

$$d(w) = \frac{1}{n(n-1)} \sum_{j \leq n} \sum_{k \leq n, k \neq j} 1 - cos(r_w^{img(j)}, r_w^{img(k)}) \tag{4.1}$$

This measure was originally introduced to account for the fact that perceptual information is more relevant for e.g. *elephant* than it is for *happiness*. It acts as a substitute for the concreteness of a word and can be used to regulate how much perceptual information should be included in a multi-modal model (Kiela et al., 2014, not included in this thesis).

A second measure follows Deselaers and Ferrari (2011), who take a similar approach but instead of calculating the pairwise distance, calculate the distance to the *centroid* μ of $\{r_w^{img(1)} \ ... \ r_w^{img(n)}\}$:

$$c(w) = \frac{1}{n} \sum_{1 \leq k \leq n} 1 - cos(r^{img(k)}, \mu) \tag{4.2}$$

For the third measure we follow Lazaridou et al. (2015b), who try different ways of modulating the inclusion of perceptual input in their multi-modal skip-gram model, and find that the *entropy* of the centroid vector μ works well (where $p(\mu_j) = \frac{\mu_j}{\| \mu \|}$ and m is the vector length):

$$H(w) = -\sum_{j=1}^{m} p(\mu_j) \log_2(p(\mu_j)) \tag{4.3}$$

4.1.1.2 Hypernym detection and directionality

The directionality of a hyponym-hypernym pair is calculated with a measure f using the following formula for a word pair (p, q). Since even co-hyponyms will not have identical values for f, we introduce a threshold α which sets a minimum difference in generality for hypernym identification:

$$s(p, q) = 1 - \frac{f(p) + \alpha}{f(q)} \tag{4.4}$$

In other words, $s(p, q) > 0$ iff $f(q) > f(p) + \alpha$, i.e. if the second word (q) is (sufficiently) more general. To avoid false positives where one word is more general but the pair is not semantically related, we introduce a second threshold θ which sets f to zero if the two concepts have low cosine similarity. This leads to the following formula:

$$s_\theta(p, q) = \begin{cases} 1 - \frac{f(p)+\alpha}{f(q)} & \text{if } \cos(\mu_p, \mu_q) \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

Various methods for obtaining the mean vector representations for cosine (hereafter μ_c) in Equation (4.5) were tested. It was found that multi-modal representations worked best. That is, for a word w with image representations $\{r_w^{img(1)} \ ... \ r_w^{img(n)}\}$, we set $\mu_c = r_w^l \parallel \frac{1}{n} \sum_i^n r_w^{img(i)}$, after normalizing both representations. Results are also reported for a visual-only mean μ_c, which performed slightly worse.

For BLESS, we know the words in a pair stand in an entailment relation, so we set $\alpha = \theta = 0$ and evaluate whether $s(p, q) > 0$, indicating that q is a hypernym of p. For WBLESS, we set $\alpha = 0.02$ and $\theta = 0.2$ without tuning, and evaluate whether $s_\theta(p, q) > 0$ (hypernym relation) or $s_\theta(p, q) \leq 0$ (no hypernym relation). For BIBLESS, we set $\alpha = 0.02$ and $\theta = 0.25$ without tuning, and evaluate whether $s_\theta(p, q) > 0$ (hyponym-hypernym), $s(p, q) = 0$ (no relation), or $s(p, q) \leq 0$ (hypernym-hyponym).

4.1.2 Results

The results can be found in Table 4.2. The proposed generality methods are compared with a frequency baseline, setting $f(p) = \text{freq}(p)$ in Equation 4.4 and using the fre-

	BLESS	WBLESS	BIBLESS
Frequency	0.58	0.57	0.39
WeedsPrec	0.63	—	—
WeedsSVM	—	0.75	—
WeedsUnSup	—	0.58	—
SLQS	0.87	—	—
Dispersion	0.88	0.75 (0.74)	0.57 (0.55)
Centroid	0.87	0.74 (0.74)	0.57 (0.54)
Entropy	0.83	0.71 (0.71)	0.56 (0.53)

Table 4.2: Accuracy. For WBLESS and BIBLESS we report results for multi-modal μ_c, with visual-only μ_c in brackets.

quency scores from Turney et al. (2011). Frequency has proven to be a surprisingly challenging baseline for hypernym directionality (Herbelot and Ganesalingam, 2013; Weeds et al., 2014). In addition, we compare to the reported results of Santus et al. (2014) for WeedsPrec (Weeds et al., 2004), an early lexical entailment measure, and SLQS, the entropy-based method of Santus et al. (2014). Note, however, that these are on a subsampled corpus of 1277 word pairs from BLESS, so the results are indicative but not directly comparable. On WBLESS we compare to the reported results of Weeds et al. (2014): we include results for the highest-performing supervised method (WeedsSVM) and the highest-performing unsupervised method (WeedsUnSup).

For BLESS, both dispersion and centroid distance reach or outperform the best other measure (SLQS). They beat the frequency baseline by a large margin (+30% and +29%). Taking the entropy of the mean image representations does not appear to do as well as the other two methods but still outperforms the baseline and WeedsPrec (+25% and +20% respectively).

In the case of WBLESS and BIBLESS, we see a similar pattern in that dispersion and centroid distance perform best. For WBLESS, these methods outperform the other unsupervised approach, WeedsUnsup, by +17% and match the best-performing support vector machine (SVM) approach in Weeds et al. (2014). In fact, Weeds et al. (2014) report results for a total of 6 supervised methods (based on SVM and k-nearest neighbor (k-NN) classifiers): the unsupervised image dispersion method outperforms all of these except for the highest-performing one, reported in the Table.

The task becomes increasingly difficult as we go from directionality to detection to the combination: the dispersion-based method goes from 0.88 to 0.75 to 0.57, for example. BIBLESS is the most difficult, as shown by the frequency baseline obtaining only 0.39. The proposed methods do much better than this baseline (+18%). Image dispersion appears to be the most robust measure.

To examine the results further, we divided the test data into buckets by the shortest WordNet path connecting word pairs (Miller, 1995). We expect generality-based methods like the ones proposed here to be less accurate on word pairs with short paths, since the difference in generality may be difficult to discern. It has also been suggested that very abstract hypernyms such as *object* and *entity* are difficult to detect because their linguistic distributions are not supersets of their hyponyms' distributions (Rimell, 2014), a factor that should not affect the visual modality. We find that concept comparisons

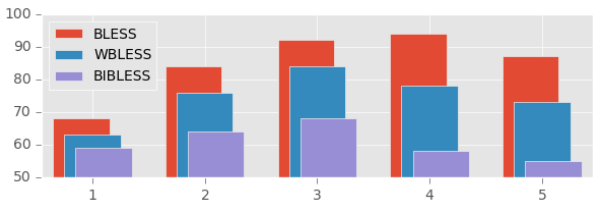

Figure 4.2: Accuracy by WordNet shortest path bucket (1 is shortest, 5 is longest).

with a very short path (bucket 1) are indeed the least accurate. We also find some drop in accuracy on the longest paths (bucket 5), especially for WBLESS and BIBLESS, perhaps because semantic similarity is difficult to detect in these cases. For a histogram of the accuracy scores according to WordNet similarity, see Figure 2.

4.1.3 Conclusion

This section introduced and evaluated three unsupervised methods for determining the generality of a concept based on its visual properties. The best-performing method, image dispersion, reaches the state-of-the-art on two standard lexical entailment datasets. A novel, more difficult task called BIBLESS was introduced, which combines hypernym detection and directionality. The proposed vision-based measures outperform a frequency baseline by a large margin.

Image generality may be particularly suited to entailment detection because it does not suffer from the same issues as linguistic distributional generality. Herbelot and Ganesalingam (2013) found that general terms like *liquid* do not always have higher entropy distributions than their hyponyms, since speakers use them in very specific contexts, e.g. *liquid* is often coordinated with *gas*.

One arguable weakness of the proposed approach is that it depends to some degree on Google's search algorithms, which may or may not include explicit diversification. We acknowledge that Google is something of a black box, but feel that this does not detract from the utility of the method: the fact that general concepts achieve greater maximum image dispersion than specific concepts is not dependent on any particular diversification algorithm. In other words, if we had used ImageNet or the ESP Game dataset, the same intuitions would hold. The reason Google Images was chosen is because it allowed us to get full coverage over the evaluation datasets.

4.2 Bilingual lexicon induction

Bilingual lexicon induction is the task of finding words that share a common meaning across different languages. It plays an important role in a variety of tasks in information retrieval and natural language processing, including cross-lingual information retrieval (Lavrenko et al., 2002; Levow et al., 2005) and statistical machine translation (Och and Ney, 2003). It is an attractive alternative to the time-consuming and expensive process of manually building high-quality resources for a wide variety of language pairs and domains. Early approaches relied on limited and domain-restricted parallel data, and the induced lexicons were typically a by-product of word alignment models (Och and Ney, 2003). That is, although parallel corpora have been used successfully for inducing bilingual lexicons for some languages (Och and Ney, 2003), these corpora are either too small or unavailable for many language pairs. To alleviate the issue of low coverage, a large body of work has been dedicated to lexicon learning from more abundant and less restricted comparable data (Fung and Yee, 1998; Rapp, 1999; Gaussier et al., 2004; Shezaf and Rappoport, 2010; Tamura et al., 2012).

However, these models typically rely on the availability of bilingual seed lexicons to produce shared bilingual spaces, as well as large repositories of comparable data. Therefore, several approaches attempt to learn lexicons from large monolingual data sets in two languages (Koehn and Knight, 2002; Haghighi et al., 2008), but their performance again relies on language pair-dependent clues such as orthographic similarity. An alternative approach removes the requirement of seed lexicons, and induces lexicons using bilingual spaces spanned by multilingual probabilistic topic models (Vulić et al., 2011; Liu et al., 2013; Vulić and Moens, 2013b). However, these models require document alignments as initial bilingual signals.

These approaches work by mapping language pairs to a shared bilingual space and extracting lexical items from that space. Bergsma and Van Durme (2011) showed that this bilingual space need not be linguistic in nature: they used labeled images from the Web to obtain bilingual lexical translation pairs based on the visual features of corresponding images. Local features are computed using SIFT (Lowe, 2004) and color histograms (Deselaers et al., 2008) and aggregated as bags of visual words (BOVW) (Sivic and Zisserman, 2003) to get bilingual representations in a shared visual space. Their highest performance is obtained by combining these visual features with normalized edit distance, an orthographic similarity metric (Navarro, 2001).

There are several advantages to having a visual rather than a linguistic intermediate bilingual space: while images are readily available for many languages through resources such as Google Images, language pairs that have sizeable comparable, let alone parallel, linguistic corpora are relatively scarce. Having an intermediate visual space means that words in different languages can be grounded in the same space. In fact, using vision as an intermediate space is a natural thing to do: when we communicate with someone who does not speak our language, we often communicate by directly referring to our surroundings. Languages that are linguistically far apart will, by cognitive necessity, still refer to objects in the same visual space. While some approaches to bilingual lexicon induction rely on orthographic properties (Haghighi et al., 2008; Koehn and Knight, 2002) or properties of frequency distributions (Schafer and Yarowsky, 2002) that will work only for closely related languages, a visual space can work for any language, whether it's English or Chinese, Arabic or Icelandic, or all Greek to you.

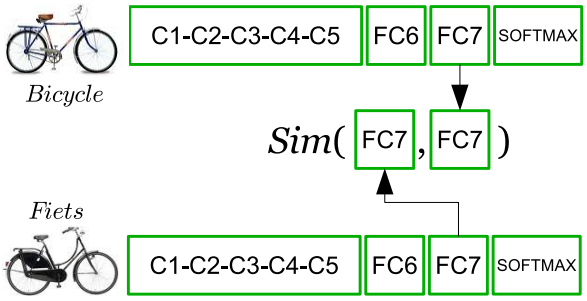

Figure 4.3: Illustration of calculating similarity between images from different languages.

Here, we apply CNN-derived visual features to the task of bilingual lexicon induction. To obtain a translation of a word in a source language, we find the nearest neighbors from words in the target language, where words in both languages reside in a shared visual space made up of CNN-based features. In other words, we test the ability of purely visual data to induce shared bilingual spaces and to consequently learn bilingual word correspondences in these spaces. By compiling images related to linguistic concepts given in different languages, the potentially prohibitive data requirements and language pair-dependence from prior work is removed.

4.2.1 Approach

The underlying assumption is that the best translation, or matching lexical item, of a word w_s (in the source language) is the word w_t (in the target language) that is the nearest cross-lingual neighbor to w_s in the shared bilingual visual space. Hence, a similarity (or distance) score between lexical items from different languages is required. In what follows, we describe: one, how to build image representations from sets of images associated with each lexical item, i.e. how to induce a shared bilingual visual space in which all lexical items are represented; and two, how to compute the similarity between lexical items using their visual representations in the shared bilingual space.

Google Images is used to extract the top n ranked images for each lexical item in the evaluation datasets. Using Google Images has the advantage that it has full coverage and is multi-lingual, as opposed to other potential image sources such as ImageNet or the ESP Game dataset. For each search query we specify the target language corresponding to the lexical item's language. We extract the pre-softmax layer of an AlexNet (Krizhevsky et al., 2012) for each image. See Figure 4.3 for a simple diagram illustrating the approach on the comparison *bicycle-fiets* (the Dutch word for bicycle). Figure 4.4 gives some example images retrieved using the same query terms in different languages.

4.2.1.1 Visual similarity

Consider the *bicycle* and *fiets* example. Each of the two words has n images associated with it — the top n as returned by Google image search, using *bicycle* and *fiets* as separate query terms. Hence to calculate the similarity, a measure is required which takes two sets of images as input.

A standard approach would be to derive visual representations through an aggregation function. As before, we try two aggregation functions: CNN-Mean and CNN-Max. To calculate the similarity between the visual representations of words and translation

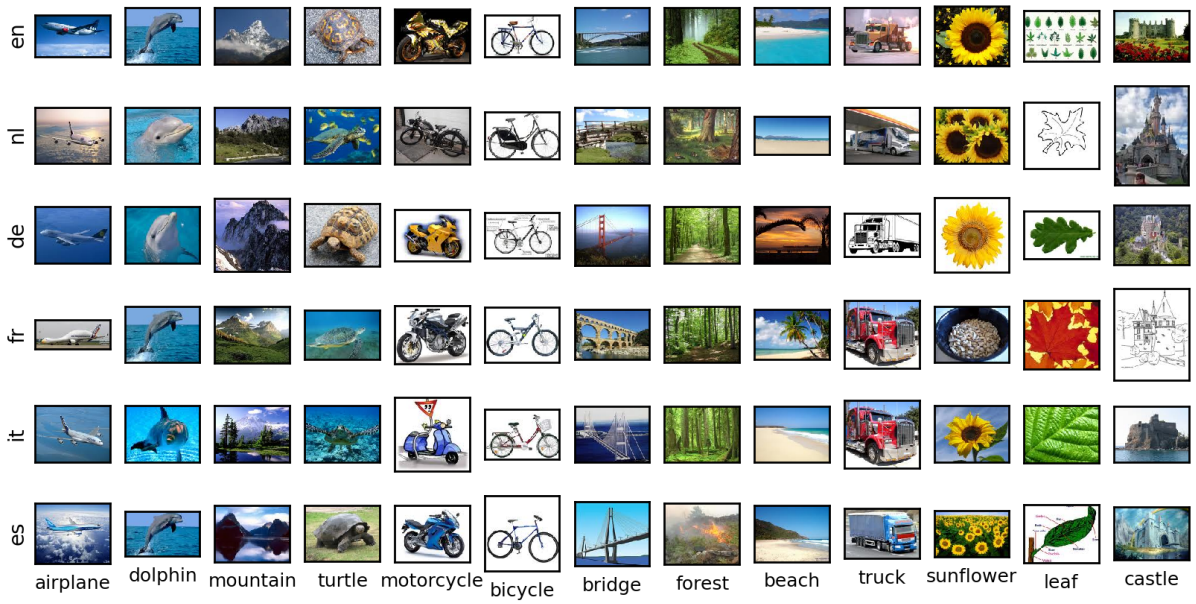

Figure 4.4: Example images for the languages in the Bergsma and Van Durme dataset.

candidates, we can then simply use cosine similarity.

An alternative strategy is to consider the similarities between individual images instead of their aggregated representations. Where CNN-Mean and CNN-Max would be *middle fusion* methods, aggregating similarity scores would constitute *late fusion* (see Section 2.2.5). Bergsma and Van Durme (2011) propose two similarity metrics based on this principle: taking the average of the maximum similarity scores (AvgMax), or the maximum of the maximum similarity scores (MaxMax) between associated images. Continuing with our example, for each of the n images for *bicycle*, the maximum similarity is found by searching over the n images for *fiets*. AvgMax then takes the average of those n maximum similarities; MaxMax takes their maximum. To avoid confusion, we will refer to the CNN-based models that use these metrics as CNN-AvgMax and CNN-MaxMax. Formally, these metrics are defined as in Table 4.3.

4.2.1.2 Evaluation

Bergsma and Van Durme's primary evaluation dataset consists of a set of five hundred matching lexical items for fifteen language pairs, based on six languages (i.e., $\binom{6}{2} = 15$). The data is publicly available online.[1] In order to get the five hundred lexical items, they first rank nouns by the conditional probability of them occurring in the pattern "{*image,photo,photograph,picture*} *of* {*a,an*} _____" in the web-scale Google N-gram corpus (Lin et al., 2010), and take the top five hundred words as their English lexicon. For each item in the English lexicon, they obtain corresponding items in the other languages—Spanish, Italian, French, German and Dutch—through Google Translate. This dataset will be referred to as Bergsma500.

In addition, we evaluate on a dataset constructed to measure the general performance of bilingual lexicon learning models from comparable Wikipedia data (Vulić and Moens, 2013a). The dataset comprises 1,000 nouns in three languages: Spanish (ES), Italian (IT),

[1] http://www.clsp.jhu.edu/~sbergsma/LexImg/

CNN-AvgMax	$\frac{1}{n} \sum_{i_s \in \mathcal{I}(w_s)} \max_{i_t \in \mathcal{I}(w_t)} sim(i_s, i_t)$
CNN-MaxMax	$\max_{i_s \in \mathcal{I}(w_s)} \max_{i_t \in \mathcal{I}(w_t)} sim(i_s, i_t)$
CNN-Mean	$sim(\frac{1}{n} \sum_{i_s \in \mathcal{I}(w_s)} i_s, \frac{1}{n} \sum_{i_t \in \mathcal{I}(w_t)} i_t)$
CNN-Max	$sim(\max' \mathcal{I}(w_s), \max' \mathcal{I}(w_t))$

Table 4.3: Visual similarity metrics between two sets of n images. $\mathcal{I}(w_s)$ represents the set of images for a given source word w_s, $\mathcal{I}(w_t)$ the set of images for a given target word w_t; $\max'$ takes a set of vectors and returns the single element-wise maximum vector.

and Dutch (NL), along with their one-to-one gold-standard word translations in English (EN) compiled semi-automatically using Google Translate and manual annotators for each language. This dataset will be referred to as Vulic1000[2]. The test set is accompanied with comparable data for training, for the three language pairs ES/IT/NL-EN on which text-based models for bilingual lexicon induction were trained (Vulić and Moens, 2013a).

Given the way that the Bergsma500 dataset was created, in particular the use of the pattern described above, it contains largely concrete linguistic concepts (since, e.g., *image of a democracy* is unlikely to have a high corpus frequency). In contrast, Vulic1000 was designed to capture general bilingual word correspondences, and contains several highly abstract test examples, such as *entendimiento (understanding)* and *desigualdad (inequality)* in Spanish, or *scoperta (discovery)* and *cambiamento (change)* in Italian. Using the two evaluation datasets can potentially provide some insight into how purely visual models for bilingual lexicon induction behave with respect to both abstract and concrete concepts.

In each case, performance is measured in the standard way using the mean-reciprocal rank:

$$MRR = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{rank(w_s, w_t)} \tag{4.6}$$

where $rank(w_s, w_t)$ denotes the rank of the correct translation w_t (as provided in the gold standard) in the ranked list of translation candidates for w_s, and M is the number of test cases. We also use *precision at N* (P@N) (Gaussier et al., 2004; Tamura et al., 2012; Vulić and Moens, 2013a), which measures the proportion of test instances where the correct translation is within the top N highest ranked translations.

4.2.2 Results on Bergsma500

The four similarity metrics are evaluated on the Bergsma500 dataset, comparing the results to the systems of Bergsma and Van Durme, who report results for the AvgMax

[2]http://people.cs.kuleuven.be/~ivan.vulic/software/

Method	P@1	P@5	P@20	MRR
B&VD Visual-Only	31.1	41.4	53.7	0.367
B&VD Visual + NED	48.0	59.5	68.7	0.536
CNN-AvgMax	**56.7**	**69.2**	**77.4**	**0.658**
CNN-MaxMax	42.8	60.0	64.5	0.529
CNN-Mean	50.5	62.7	71.1	0.586
CNN-Max	51.4	64.9	74.8	0.608

Table 4.4: Performance on Bergsma500 compared to Bergsma and Van Durme (B&VD).

function, having concluded that it performs better than MaxMax on English-Spanish translations. For comparison, we consider their best-performing visual-only system, which combines SIFT-based descriptors with color histograms, as well as their best-performing overall system, which combines the visual approach with normalized edit distance (NED). Results are averaged over fifteen language pairs.

The results can be seen in Table 4.4. Each of the CNN-based methods outperforms the B&VD systems. The best performing method overall, CNN-AvgMax, provides a 79% relative improvement over the B&VD visual-only system on the MRR measure, and a 23% relative improvement over their best-performing approach, which includes non-visual information in the form of orthographic similarity. Moreover, their methods include a tuning parameter λ that governs the contributions of SIFT-based, color histogram and normalized edit distance similarity scores, whilst our approach does not require any parameter tuning.

4.2.3 Results on Vulic1000

In short, the CNN-based approach does much better than B&VD's approaches. However, we should also compare the visual-only approach to linguistic approaches for bilingual lexicon induction. Since Bergsma500 has not been evaluated with such methods, we evaluate on the Vulic1000 dataset of Vulić and Moens (2013a). This dataset has been used to test the ability of bilingual lexicon induction models to learn translations from comparable data. One should not necessarily expect visual methods to outperform linguistic ones, given that linguistic methods have held the state of the art since the beginning, but even the comparison is instructive to see.

We compare our visual models against the current state-of-the-art lexicon induction model using comparable data (Vulić and Moens, 2013b). This model induces translations from comparable Wikipedia data in two steps: (1) It learns a set of highly reliable one-to-one translation pairs using a shared bilingual space obtained by applying the multilingual probabilistic topic modeling (MuPTM) framework (Mimno et al., 2009). (2) These highly reliable one-to-one translation pairs serve as dimensions of a word-based bilingual semantic space (Gaussier et al., 2004; Tamura et al., 2012). The model then bootstraps from the high-precision seed lexicon of translations and learns new dimensions of the bilingual space until convergence. This model, which we call BootStrap, obtains the current best results on the evaluation dataset of Vulić and Moens (2013b).

Table 4.5 shows the results for the language pairs in the Vulic1000 dataset. Of the four similarity metrics, CNN-AvgMax again performs best, as it did for Bergsma500.

Language Pair	Method	P@1	P@5	P@10	P@20	MRR
ES ⇒ EN	BootStrap	57.7	74.7	80.9	84.8	0.652
	CNN-AvgMax	41.9	54.6	59.1	65.6	0.485
	CNN-MaxMax	34.9	47.4	53.7	58.5	0.414
	CNN-Mean	35.4	48.5	51.7	55.8	0.416
	CNN-Max	33.3	46.3	50.3	54.5	0.395
IT ⇒ EN	BootStrap	64.7	80.6	85.6	89.7	0.716
	CNN-AvgMax	28.3	40.6	44.8	50.9	0.343
	CNN-MaxMax	22.6	33.5	38.6	44.4	0.282
	CNN-Mean	22.7	33.2	37.9	42.6	0.281
	CNN-Max	21.3	32.7	36.8	41.5	0.269
NL ⇒ EN	BootStrap	20.6	35.7	43.4	51.3	0.277
	CNN-AvgMax	38.4	48.5	53.7	58.6	0.435
	CNN-MaxMax	30.8	42.6	47.8	52.9	0.367
	CNN-Mean	32.3	42.3	46.5	50.1	0.373
	CNN-Max	30.4	41.0	44.3	49.3	0.356

Table 4.5: Performance on Vulic1000 compared to the linguistic bootstrapping method of Vulić and Moens (2013b).

The linguistic BootStrap method outperforms the visual approach for two of the three language pairs, but, for the NL-EN language pair, the visual methods in fact perform better. This can be explained by the observation that Vulić and Moens's NL-EN training data for the BootStrap model is less abundant (2-3 times fewer Wikipedia articles) and of lower quality than the data for their ES-EN and IT-EN models. Thus, these results are highly encouraging: while purely visual methods cannot yet reach the peak performance of linguistic approaches that are trained on sufficient amounts of high-quality text data, they outperform linguistic state-of-the-art methods when there is less or lower quality text data available—which one might reasonably expect to be the default scenario.

4.2.4 Conclusion

This section presented a novel approach to bilingual lexicon induction that uses convolutional neural network-derived visual features. Using only such visual features, we are able to outperform existing visual and orthographic systems, and even a state-of-the-art linguistic approach for one language, on standard bilingual lexicon induction tasks. This was the first work to provide a comparison of visual and state-of-the-art linguistic approaches to bilingual lexicon induction. The beauty of the current approach is that it is completely language agnostic and closely mirrors how humans would perform bilingual lexicon induction: by referring to the external world.

4.3 Conclusions & discussion

This chapter showed how CNN-derived features lead to improvements over uni-modal linguistic approaches, as well as over SIFT-based approaches. In the first section, we exploited the notion of *generality* with visual representations to model hypernymy. The problem becomes a very natural one for humans to solve, once we rephrase it visually: how different are all the *elephants* that I've seen from each other, compared to all the *animals* that I've seen? It is a simple test of category membership, and the fact that it outperforms all linguistic methods, including supervised ones, means that the intuition is not only sound, but practically useful.

In a similar fashion, we recast bilingual lexicon induction as a much more "human" problem: when two humans meet and don't speak the same language, they communicate though a common and shared visual world. This notion is easy to exploit using an intermediate visual space. Although Bergsma & Van Durme came to the same idea earlier, the fact that this work made use of high-quality CNN-derived features allowed it to beat linguistic approaches on at least some of the language pairs. In work that followed on from this thesis, the state-of-the-art in bilingual lexicon induction was obtained for all languages by a multi-modal approach that builds on the visual approach proposed here (Vulić et al., 2016).

Part III

Grounding in non-visual modalities

CHAPTER 5

AUDITORY GROUNDING

The usage of raw image data has become the *de facto* method for grounding representations in perception. If the objective is to ground semantic representations in perceptual information, though, why should we stop at image data? The meaning of *violin* is surely not only grounded in its visual properties, such as its shape, color and texture, but also in its sound, pitch and timbre. To understand how perceptual input leads to conceptual representation, which is one of the central questions that this thesis aims to shed light on, we should cover as many perceptual modalities as possible. In this chapter, we introduce the first ever attempt to perform grounding with neural networks in raw auditory, rather than visual, perception.

The same types of evaluations and a very similar approach to visual grounding are used. We evaluate on human similarity and relatedness ratings, just like before. Since this work constitutes the first time auditory grounding is performed, we introduce a method for constructing auditory concept representations, called bag of audio words (BoAW). We then show that we can improve over BoAW by applying deep learning, specifically through training convolutional neural networks on sound files to obtain neural auditory embeddings (NAEs).

5.1 Evaluation

The MEN test collection (Bruni et al., 2014) is used for evaluation. Evidence suggests, however, that the inclusion of visual representations only improves performance for certain concepts, and that in some cases the introduction of visual information is detrimental to performance on similarity and relatedness tasks (Kiela et al., 2014, not included in this thesis). The same is likely to be true for other perceptual modalities: in the case of comparisons such as *guitar-piano*, the auditory modality is certainly meaningful, whereas in the case of *democracy-anarchism* it is probably less so. This is even more likely to be the case for less dominant modalities such as auditory perception.

Therefore, we had two graduate students annotate the MEN dataset according to whether auditory perception is relevant to the pairwise comparison. The annotation criterion was as follows: if both concepts in a pairwise comparison have a distinctive associated sound, the modality is deemed relevant. Inter-annotator agreement was high, with $\kappa = 0.93$. Some examples of relevant pairs can be found in Table 6.1. Hence, we now have two evaluation datasets for conceptual similarity and relatedness: the MEN

MEN	human rating	relevant
automobile-car	1.00	✓
rain-storm	0.98	✓
cat-feline	0.96	✓
pregnancy-pregnant	0.96	
jazz-musician	0.88	✓
bird-eagle	0.88	✓
highway-traffic	0.88	✓
guitar-piano	0.86	✓
foliage-tulip	0.64	

Table 5.1: Illustrative examples of pairs in the datasets where auditory information is or is not relevant, together with their corresponding similarity rating as provided by human annotators.

Dataset	MEN	AMEN
Textual	3000	258
Auditory	2590	233

Table 5.2: Number of concept pairs for which representations are available in each modality.

test collection **MEN**, and its auditory-relevant subset **AMEN**. Due to the nature of the auditory data sources, it is not possible to build auditory representations for all concepts in the test sets. Hence, we only evaluate on the covered subsets to ensure a fair comparison, that is, we only use the comparisons that have coverage in both the textual and auditory modalities, as shown in Table 5.2. While 258 concept pairs were annotated as auditorily relevant, auditory representations were only available for 233 of those.

5.2 Approach

One reason for using raw image data in multi-modal models is that there are many high quality resources available that contain tagged images, such as ImageNet and the ESP Game dataset, and image search engines such as Google and Bing. Such human annotated high-quality resources do not exist for audio files, but there does exist an online search engine that provides tagged sound files: Freesound[1] (Font et al., 2013). Hence, we use FreeSound to obtain audio files. Freesound is a collaborative database released under Creative Commons licenses, in the form of snippets, samples and recordings, that is aimed at sound artists. The Freesound API allows users to easily search for audio files that have been tagged using certain keywords. For each of the concepts in the evaluation datasets, we used the Freesound API to obtain samples encoded in the standard open source OGG

[1] http://www.freesound.org.

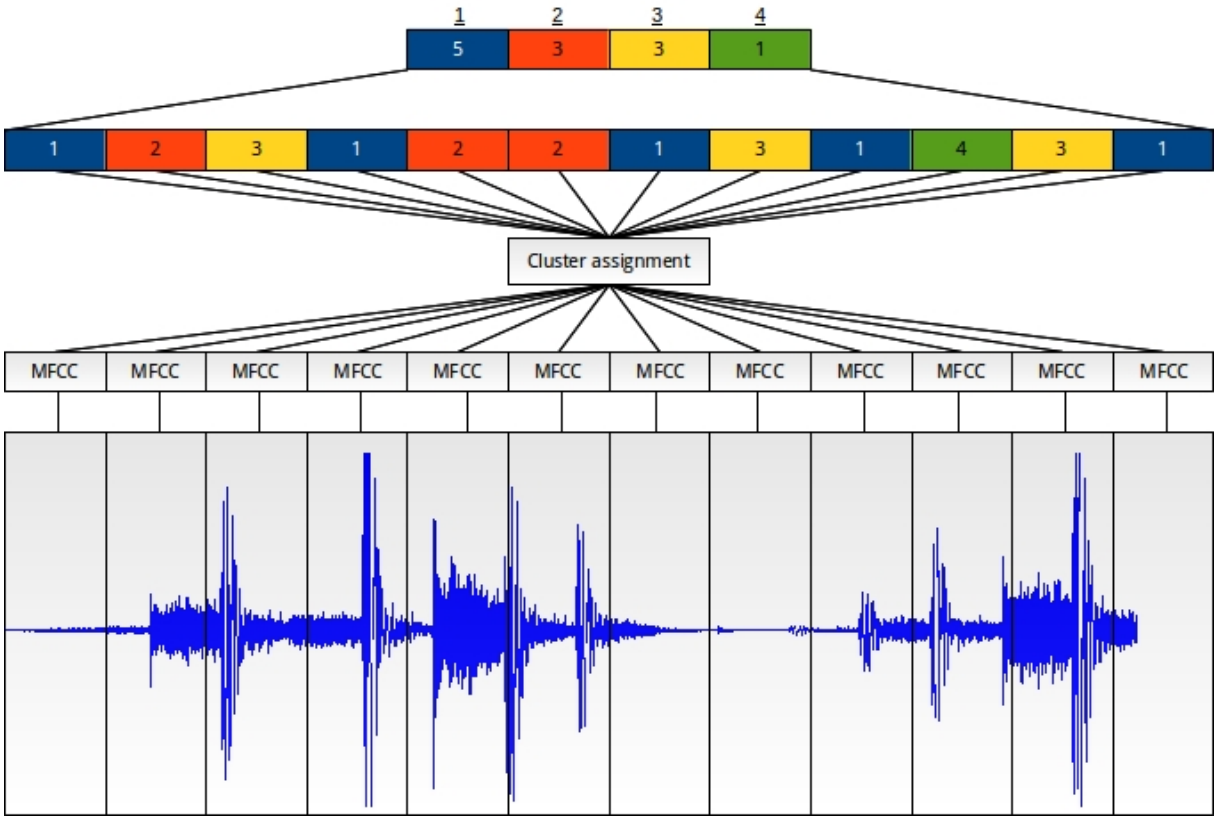

Figure 5.1: Illustration of the BoAW method. Each of the MFCC descriptors is assigned to a cluster. Assignments are subsequently quantized into a bag of audio words representation. In this illustration, $k = 4$ in k-means, which means there are four clusters and the value for each of the k clusters is the number of datapoints belonging it.

format[2]. The Freesound API allows for various degrees of keyword matching: we opted for the strictest keyword matching, in that the audio file needs to have been purposely tagged with the given word (the alternative includes searching the text description for matching keywords).

5.2.1 Auditory representations

We experiment with two methods for obtaining auditory representations: bag of audio words and transfering a layer from a trained convolutional neural network. The former is a relatively simple approach that does not take into account any interdependencies between local feature descriptors, whereas the latter is more sophisticated and able to extract more elaborate patterns and interactions. While these methods vary significantly, in that they use different input features (local feature descriptors of frames versus spectrograms), their representations are constructed from the same sound files, allowing us to compare the two methods.

[2]http://www.vorbis.com.

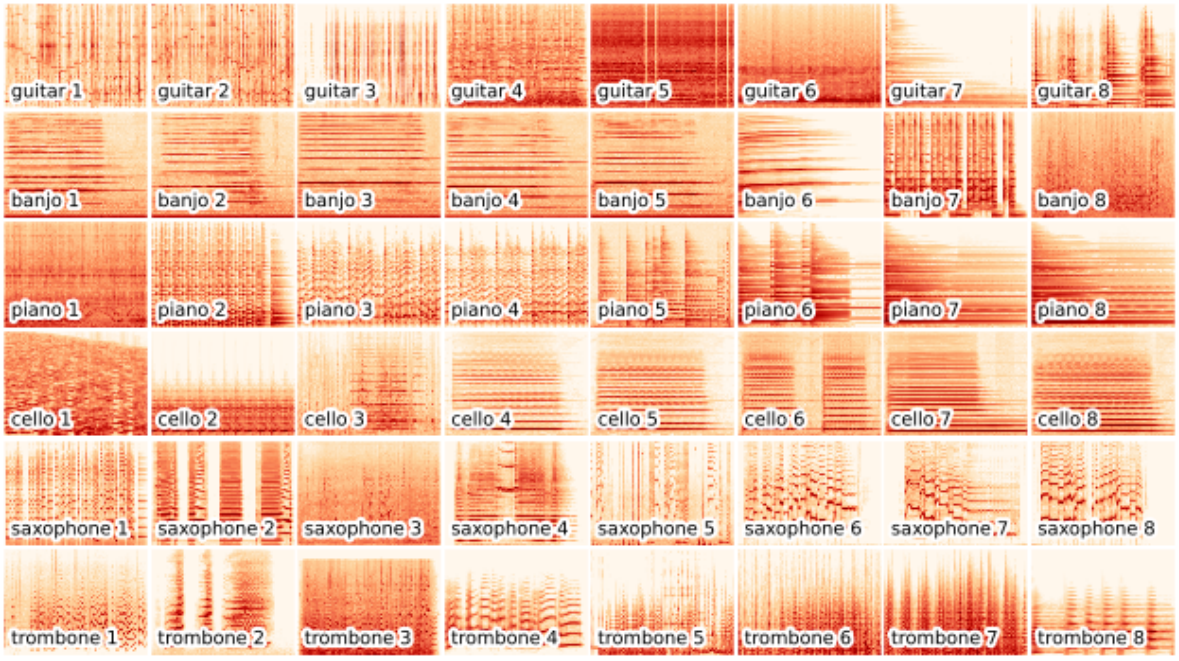

Figure 5.2: Examples of spectrograms, plotted for various musical instruments.

5.2.1.1 Bag of audio words (BoAW)

A common approach to obtaining acoustic features of audio files is the Mel-scale Frequency Cepstral Coefficient (MFCC) (O'Shaughnessy, 1987). MFCC features are abundant in a variety of applications in audio signal processing, ranging from audio information retrieval, to speech and speaker recognition, and music analysis (Eronen, 2003). Such features are derived from the mel-frequency cepstrum representation of an audio fragment (Stevens et al., 1937). In MFCC, frequency bands are spaced along the mel scale, which has the advantage that it approximates human auditory perception more closely than e.g. linearly-spaced frequency bands. Hence, MFCC takes human perceptual sensitivity to audio frequencies into consideration, which makes it suitable for e.g. compression and recognition tasks, but also for our current objective of modeling auditory perception.

After having obtained MFCC descriptors, we cluster them using mini-batch k-means (Sculley, 2010) and quantize the descriptors into a "bag of audio words" (BoAW) (Foote, 1997) representation by comparing the MFCC descriptors to the cluster centroids. We set $k = 300$ and do not apply any additional weighting[3]. See Figure 5.1 for an illustration of the process for a single audio file. Auditory representations for a concept are obtained by taking the mean of the BoAW representations of the relevant audio files.

5.2.1.2 Neural auditory embeddings (NAE)

In Chapter 3, it was shown that it is possible to transfer and aggregate convolutional neural network layers in order to obtain a visual semantic representation. Here, we examine whether a similar methodology can be applied to auditory representations. We obtain

[3]Weighting might improve performance, but runs the risk of "fitting" to the dataset if we only obtain audio files for the words in the dataset, biasing performance.

Figure 5.3: Illustration of the Neural Auditory Embedding method, using a convolutional neural network. The auditory signal is converted to a spectrogram which is fed to the neural network for classification. The pre-softmax layer, FC7, is transferred and taken as the neural audio embedding (NAE) for the given sound file.

sound file representations by transfering the pre-softmax layer from a convolutional neural network trained on audio classification. The advantage of using a convolutional neural network instead of a recurrent one (RNN) is that it requires less memory and suffers less from the vanishing or exploding gradients problem (Pascanu et al., 2012). RNNs are especially susceptible to this problem for the current study, given that sound files can vary considerably in the number of frames (i.e., their duration in milliseconds), which means padding mini-batches is cumbersome and leads to unnecessary memory allocations.

We use a spectrogram (Flanagan, 2013) of the sound file as the input to the network, i.e., the input is a three-dimensional representation of the spectrum of frequencies as they vary with time. A spectrogram can be interpreted as a visual rendering of an auditory signal, which means that we can apply a similar network architecture to deep neural networks used in computer vision, for classifying auditory patterns. Figure 5.2 shows how simple visual inspection already reveals some clear patterns for certain musical instruments, which convolutional networks are well-equipped to exploit.

Our architecture is identical to the Krizhevsky et al. (2012) network architecture: the network consists of 5 convolutional layers, followed by two fully connected rectified linear unit (ReLU) layers that feed into a softmax for classification. The network learns through a multinomial logistic regression objective. We obtain audio embeddings by performing a forward pass with a spectrogram and subsequently taking the 4096-dimensional fully connected layer that precedes the softmax (FC7) as the representation of that sound file (see Figure 5.3).

One of the reasons behind the success of convolutional neural networks in computer vision is that they can be trained on millions of images. This allows for the lower layers of the network to become very good "edge detectors", and to become more specific to the final classification decision in higher layers, as shown by Zeiler and Fergus (2014). Since there are fewer sound files available, we exploit the power of vision-based CNNs by applying a transfer learning technique, where we "finetune" a network that has already been trained on ILSVRC 2012. This means we can rely on the network to already perform well at recognizing visual patterns. In particular, we set the learning rate to a small number for the first five layers, and learn the fully connected weights that lead to the new softmax with different labels from scratch with a higher learning rate. This allows the use of edge-detectors that were trained on a massive dataset of images, but enables the fine-tuning of parameters for the particular task at hand, in this case the classification of auditory signals as represented by spectrograms. We use standard stochastic gradient descent (SGD) optimization, with an initial learning rate of 0.01 for the fully connected layers and 0.001 for the earlier convolutional layers. The learning rate was set to degrade

accordion	bagpipe	balalaika	banjo	baritone
bass	bassoon	bell	bongo	bugle
carillon	castanets	celeste	cello	chimes
clarinet	claves	clavichord	clavier	conga
cornet	cowbell	cymbals	didgeridoo	drum
fiddle	flute	glockenspiel	gong	guitar
harmonica	harp	harpsichord	horn	keyboard
lute	lyre	mandolin	maracas	marimba
oboe	organ	piano	piccolo	saxophone
sitar	tambourine	trombone	trumpet	tuba
ukulele	violin	xylophone	zither	

Table 5.3: Labels for the musical instruments classifier.

in a stepwise fashion by a factor of 0.1 every 1000 iterations, with 4000 training iterations in total.

We experiment with training the network on either a narrow dataset of musical instruments, or a broad dataset of naturally occurring environmental sounds. In other words, one model has to be good at fine-grained distinctions between similar sounds (e.g., distinguishing between a mandolin, a ukelele and a banjo), while the other needs to be able to recognize general sound categories that can vary substantially in their audio signatures (e.g. distinguishing scissors from cows and airplanes).

Instruments Classifier For a set of 54 musical instruments, we obtain up to 1000 sound files each, yielding a total of 25324 sound files. A training and validation set are constructed by sampling 75% for the former and taking the remainder for the latter. Using the training methodology described above, we obtain an accuracy of 92% on the validation set. See Table 5.3 for the labels. Embeddings transferred from this classifier are referred to as NAE-INST in what follows.

Environmental Sounds Classifier Gygi et al. (2007) performed an extensive psychological study of auditory perception and its relation to environmental sound categories. We obtain up to 2000 sound files for the 50 classes used in their acoustic similarity and categorization experiments, which results in 31432 sound files. The classifier achieves 54% accuracy on the validation set. This number is substantially lower than the instruments classifier, which indicates that it is a significantly harder problem. The environmental labels (see Table 5.4) are much more varied and it is likely that FreeSound returns noisier sound files for these categories. Ultimately, we are less interested in the performance of the trained classifier, but more in the quality of the representations that can be extracted from that classifier, in order to use them for downstream tasks or applications. This set of labels has been specifically designed with the similarity and categorization of human auditory perception in mind, and hence it spans a wide range of sound categories

airplane	axe	baby	basketball	bells
bird	bowling	bubbling	car accelerating	car start
cat	claps	clock	cough	cow
cymbals	dog	door	drums	footsteps
gallop	glass break	gun	harp	helicopter
honking	ice drop	keyboard	laugh	match
neigh	phone	ping pong	rain	rooster
saw	scissors	sheep	siren	sneeze
splash	thunder	toilet	train	typewriter
water	wave	whistle	wipers	zipper

Table 5.4: Labels for the environmental sound classifier, from Gygi et al. (2007).

and arguably reflects human auditory perception better than the instruments dataset. Embeddings transferred from this classifier are referred to as NAE-ENV.

5.2.1.3 Duration and number

The method for obtaining the auditory representations to be used in the conceptual similarity and relatedness evaluations is as follows: For each word, we retrieve the first 100 sound samples from FreeSound with a maximum duration of 1 minute. The same sound files are used as input in all models when extracting representations, to ensure direct comparability.

5.2.2 Textual representations

We compare against textual representations, and combine auditory representations with textual representations to obtain multi-modal representations. For the textual representations we use the continuous vector representations from the log-linear skip-gram model of Mikolov et al. (2013a). Specifically, 300-dimensional vector representations were obtained by training on a dump of the English Wikipedia plus newswire (8 billion words in total)[4]. These types of representations have been found to yield the highest performance on a variety of semantic similarity tasks.

5.2.3 Multi-modal fusion strategies

Since multi-modal semantics relies on two or more modalities, there are several ways of combining or *fusing* linguistic and perceptual cues. In Chapter 2 we referred to these as early, middle and late fusion. Here, we experiment with the latter two fusion strategies.

[4]The demo-train-big-model-v1.sh script from http://word2vec.googlecode.com was used to obtain this corpus.

(a): MEN and AMEN

Modality	MEN	AMEN
TEXT	0.69	0.59
BOAW	0.23	0.43
NAE-INST	0.27	0.49
NAE-ENV	0.32	0.56
MM-BOAW$_{\alpha=0.5}$	0.62	0.64
MM-NAE-INST$_{\alpha=0.5}$	0.62	0.65
MM-NAE-ENV$_{\alpha=0.5}$	0.63	0.67

(b): MEN with tuned α on devset

Modality	MEN
TEXT	0.687
MM-MIDDLE-BOAW$_{\alpha=0.7}$	0.693
MM-MIDDLE-NAE-INST$_{\alpha=0.8}$	0.689
MM-MIDDLE-NAE-ENV$_{\alpha=0.7}$	0.697
MM-LATE-BOAW$_{\alpha=0.9}$	0.693
MM-LATE-NAE-INST$_{\alpha=0.9}$	0.691
MM-LATE-NAE-ENV$_{\alpha=0.9}$	0.695

Table 5.5: Spearman ρ_s correlation comparison of uni-modal and multi-modal representations. All correlations are significant.

5.2.3.1 Middle fusion

Middle fusion allows for individual training objectives and independent training data. Similarity between two multi-modal representations is calculated as follows:

$$sim(w_1, w_2) = g(f(r^l_{w_1}, r^a_{w_1}), f(r^l_{w_2}, v^a_{w_2}))$$

where $g = \frac{x \cdot y}{|x||y|}$ (cosine similarity), $r^l_{w_i}$ are textual representations, and $r^a_{w_i}$ are the auditory ones. We use $f(x, y) = \alpha\, x \| (1-\alpha) y$, where $\|$ is concatenation. We call this model MM-MIDDLE.

5.2.3.2 Late fusion

Late fusion can be seen as the converse of middle fusion, in that the similarity function is computed first before the similarity scores are combined:

$$sim(w_1, w_2) = h(g(r^l_{w_1}, r^l_{w_2}), g(r^a_{w_1}, r^a_{w_2}))$$

where we use cosine similarity and h is a way of combining similarities, in our case a weighted arithmetic average: $h(x, y) = \alpha\, x + (1-\alpha) y$. We call this model MM-LATE.

5.3 Results

We evaluate the quality of auditory representations by calculating the Spearman ρ_s correlation between the ranking of the concept pairs produced by the automatic similarity metric and that produced by the gold-standard similarity scores.

The results are reported in Table 5.5(a). Since cosine similarity is the normalized dot-product, and the uni-modal representations are themselves normalized, middle and late fusion are equivalent if we take the unweighted average (i.e., $\alpha = 0.5$). Since they obtain the same performance, we call these models MM-* and omit whether they use middle or late fusion in that table.

TEXT							
engine	monster	children	dinner	splash	weather	birds	dawn
gasoline	zombie	kids	lunch	bucket	rain	mammals	dusk
vehicle	dragon	girls	wedding	skateboard	storm	animals	sunrise
airplane	creatures	women	breakfast	ink	fog	rodents	moon
aircraft	clown	people	cocktail	cocktail	cold	reptiles	night
motor	dog	boys	holiday	dripping	tropical	amphibians	misty
BOAW							
engine	monster	children	dinner	splash	weather	birds	dawn
motor	dead	female	eat	wet	storm	fabric	garden
car	zombie	cow	food	run	cold	summer	summer
storm	guitar	kids	tiles	lake	winter	forest	pond
drive	ship	animals	breakfast	wave	ford	village	parrot
automobile	dark	lady	floor	sea	building	food	birds
NAE-INST							
engine	monster	children	dinner	splash	weather	birds	dawn
automobile	zombie	kids	school	wet	storm	morning	morning
motor	dead	farm	bar	lake	highway	garden	birds
vehicle	guy	party	bottle	run	wind	zoo	tropical
auto	fun	women	lunch	dripping	motorcycle	tropical	zoo
car	action	girls	mac	stone	ocean	mountain	village
NAE-ENV							
engine	monster	children	dinner	splash	weather	birds	dawn
motor	zombie	protest	lunch	wet	storm	summer	tropical
automobile	dead	kids	coffee	lake	wind	morning	zoo
drive	guy	party	bar	dripping	alley	forest	birds
vehicle	lion	happy	mug	run	rain	tropical	morning
car	man	women	rusty	river	ocean	zoo	dusk

Table 5.6: Example nearest neighbors in MEN for textual representations and auditory BoAW and NAE representations.

While the performance of textual representations is lower on AMEN than on MEN, the performance of uni-modal auditory representations is higher on AMEN than on MEN. This indicates that our auditory representations are better at judging auditory-relevant comparisons than they are at non-auditory ones, as we might expect. Both types of neural audio embeddings (NAE-*) outperform bag of audio words (BOAW) without tuning, in the uni-modal as well as in the multi-modal case, indicating that these neural auditory representations are better at capturing human similarity and relatedness judgments than the simpler model. Embeddings extracted from the broad environmental sounds classifier (NAE-ENV) outperform embeddings extracted from the narrow musical instruments classifier (NAE-INST).

Focusing on AMEN, we see a large increase in performance when using multi-modal representations, including for MM-BOAW. Performance for the uni-modal NAEs is close to the performance of the textual model (TEXT). Although that datset has been tagged with auditory relevance in mind, many of the comparisons (e.g. *cat-kittens* or *car-automobile*) are dominated by visual or linguistic information, which means that these auditory representations must be of a high quality if they still mirror the human judgments.

Textual models (TEXT) outperform multi-modal ones (MM-*) on the full MEN dataset. This is understandable given how few pairwise comparisons are auditory-relevant. In many cases we are still able to obtain sound files, but these tend to be of poor quality and lead to noisy representations (e.g., what does "sunlight" sound like?). As discussed in Section 5.2.3, the mixing parameter α plays an important role in the middle and late fusion models. It was fixed at 0.5 for the MM model in Table 5.5(a), but it is possible to use a development set to obtain a more optimal weighting. Hence, 100 comparisons were

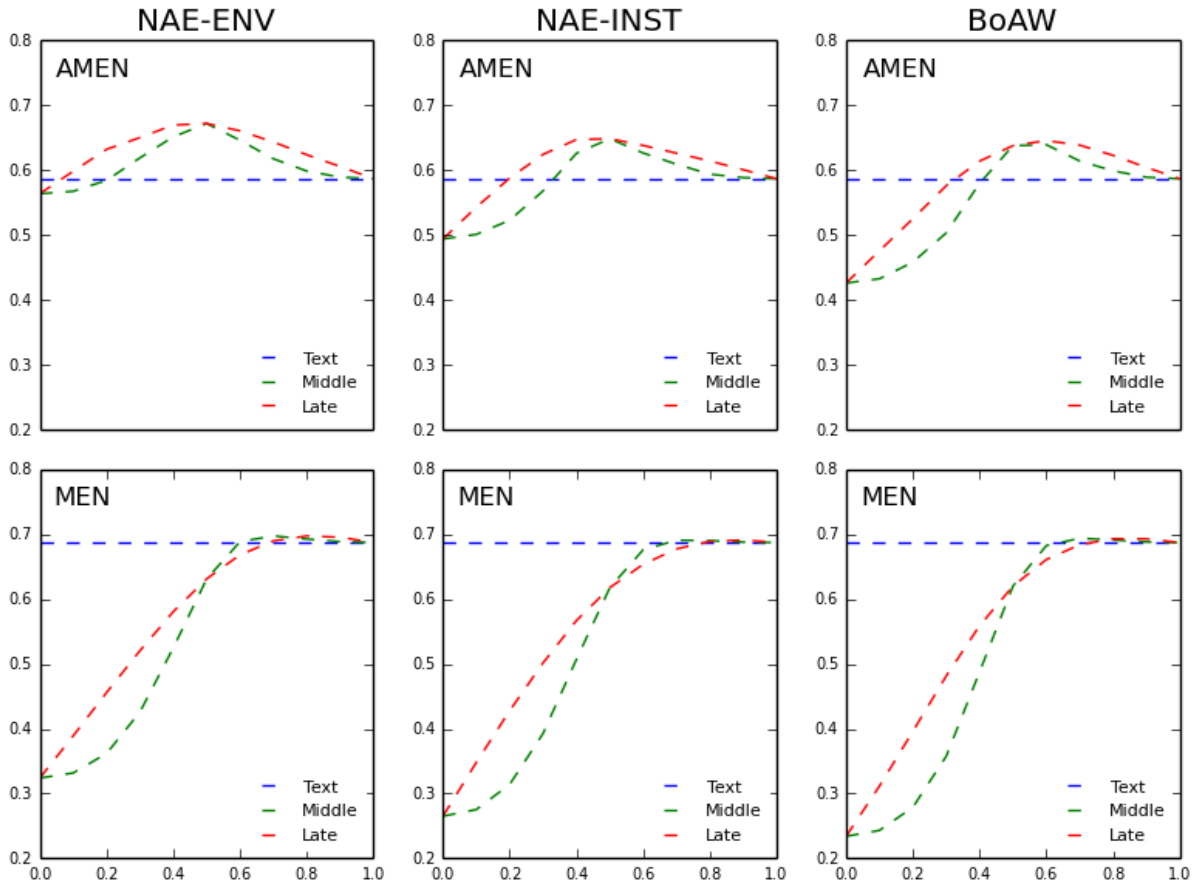

Figure 5.4: Performance of middle and late multi-modal fusion models compared to textual representations on both datasets when varying the α mixing parameter on the x-axis.

sampled from MEN and used as a development set for tuning the α parameter. The results for the full dataset using this tuned parameter can be found in Table 5.5(b), reporting up to three decimal places because of the smaller differences. Note that although the differences in performance are probably not significant, they indicate that the inclusion of auditory information is not detrimental to performance with a more intelligent choice of α.

A small qualitative analysis of the auditory representations for the words in the MEN dataset is shown in Table 5.6. The nearest neighbors are remarkably semantically coherent. For example, the auditory models group together sounds produced by cars and engines. Nearest neighbors for the textual model tend to be of a more abstract nature: where we find *wet* and *lake* as auditory neighbors for *splash*, the textual model gives us concepts like *bucket*, which can make splashes but does not sound like them. While auditory neighbors of *dawn* are related to sounds one might hear at that time of day (*birds* chirping on the *morning* in *summer*), the textual model knows that *dawns* come after the *night* when the *moon* makes way for the *sunrise*. We observe that neighbors of *birds* in the textual model are all other types of animals—i.e., categorically related—while the auditory neighbors are related in a much more associative manner.

Model	Mean	Max
TEXT	0.30 ± 0.06	0.42
BOAW	0.20 ± 0.05	0.37
NAE-INST	0.25 ± 0.07	0.42
MM-BOAW	0.32 ± 0.07	0.50
MM-NAE-INST	0.37 ± 0.07	0.54

Table 5.7: V-measure performance for clustering musical instruments. Mean is over 100 runs of k-means.

woodwind	accordion, bassoon, clarinet, didgeridoo, flute, harmonica, oboe
string	balalaika, banjo, bass, cello, fiddle, guitar, harp, lute, lyre, mandolin, sitar, tambourine, ukelele, violin, zither
brass	baritone, bugle, cornet, horn, saxophone, trombone, trumpet, tuba
percussion	bell, bongo, castanets, chimes, claves, conga, cowbell, cymbals, drum, glockenspiel, gong, maracas, marimba, xylophone
piano	carillon, celeste, clavichord, clavier, harpsichord, keyboard, piano, piccolo

Table 5.8: Musical instruments and their classes.

5.3.1 Fusion strategies

Another question to examine is how much input of which modality is most useful for predicting human similarity and relatedness ratings? This can be examined by experimenting with varying the α parameter for the full MEN dataset and plotting correlation. The results are shown in Figure 5.4, where moving to the right on the x-axis uses more textual input and moving to the left uses more auditory input.

There are some interesting observations to be made. The environmental sound embeddings perform better than the other auditory embeddings, for all alpha values. The late model consistently outperforms the middle fusion model on AMEN, which is probably because it is less susceptible to noise in the auditory representation. Optimal performance seems to be between $0.6 \leq \alpha \leq 0.9$ for both middle and late fusion strategies on MEN, indicating that it is better to include more textual than auditory input. It appears that any α in that range (i.e., where we have more textual input but still some auditory signal), outperforms the purely textual representation. In the case of the auditory-relevant subsets, we see a consistent improvement for a wide range of α parameter settings.

5.3.2 Musical instrument clustering

To further examine the finding that multi-modal representations perform well on the auditory-relevant datasets, we evaluate on an altogether different task, namely that of unsupervised musical instrument classification. The set of instruments in Table 5.3 was manually divided into 5 classes: brass, percussion, piano-based, string and woodwind in-

	TEXT
1	piccolo
2	flute, lute, harpsichord, marimba, zither, harp, clavichord, sitar, didgeridoo, carillon, lyre, keyboard
3	harmonica, mandolin, banjo, guitar, accordion, ukulele, fiddle, bass
4	xylophone, tambourine, glockenspiel, claves, maracas, castanets, cymbals, celeste, horn, balalaika, clavier, cowbell, bongo, bugle, drum, conga, chimes, bell, gong
5	clarinet, trombone, bassoon, cello, saxophone, piano, violin, oboe, tuba, trumpet, cornet, baritone
	BOAW
1	xylophone, glockenspiel, cowbell, tambourine, chimes, celeste, maracas, bell, conga
2	flute, piano, violin, clarinet, saxophone, mandolin, harmonica, harp, oboe, banjo, lute, trumpet, zither, harpsichord, sitar, marimba, accordion, cornet, ukulele, clavichord, fiddle, horn, cymbals, balalaika, claves, lyre, keyboard, castanets, bugle, drum
3	trombone, tuba, cello, guitar, bassoon, baritone, didgeridoo, bass, piccolo, carillon, bongo
4	gong
5	clavier
	NAE-INST
1	accordion, balalaika
2	trombone, piano, cello, violin, saxophone, flute, banjo, oboe, tuba, mandolin, clarinet, harmonica, guitar, harpsichord, bassoon, cornet, trumpet, marimba, sitar, harp, lute, ukulele, zither, didgeridoo, clavichord, fiddle, horn, bugle, baritone, bass
3	xylophone, glockenspiel, celeste, claves, carillon, clavier, chimes, cowbell, piccolo, keyboard, bongo, lyre, bell, conga
4	gong
5	tambourine, cymbals, drum, castanets, maracas

Table 5.9: Instruments closest to cluster centroid by cosine distance for textual and multi-modal representations.

struments (see Table 5.8). For each instrument, as many audio files as available were obtained from FreeSound. We then performed k-means clustering with five cluster centroids and compared results between textual, bag of audio words and NAE representations. We experiment with the NAE-INST embeddings, which are specialized for musical instrument identification.

This is an interesting problem because instrument classes are determined somewhat by convention (is a *saxophone* a brass or a woodwind instrument?). What is more, how instruments sound is rarely described in detail in text, so corpus-based linguistic representations cannot take this information into account. Table 5.7 shows the mean and standard deviation of V-measure scores (Rosenberg and Hirschberg, 2007), a well-known clustering evaluation metric that represents the harmonic mean between the homogeneity (how many datapoints in the same cluster are in the same class) and completeness (how many datapoints in the same class are in the same cluster)[5], obtained by applying the clustering algorithm a total of 100 times in order to mitigate differences due to the random seeding phase in k-means. The results clearly show that the multi-modal representation, which utilizes both linguistic information and auditory input, performs better on this task than the uni-modal representations.

It is interesting to observe that the textual representations perform better than the auditory ones: a possible explanation for this result is that audio files in FreeSound are rarely samples of a single individual instrument, so if a bass is often accompanied by a drum this will affect the overall representation. The clusters that were obtained by the maximally performing model are reported in Table 5.9: for the 5 clusters under the three uni-modal models, it shows the nearest instruments to the cluster centroids, qualitatively demonstrating the greater cluster coherence for the multi-modal models, in particular

[5]We find the same patterns in the results with other clustering metrics such as purity and B-cubed.

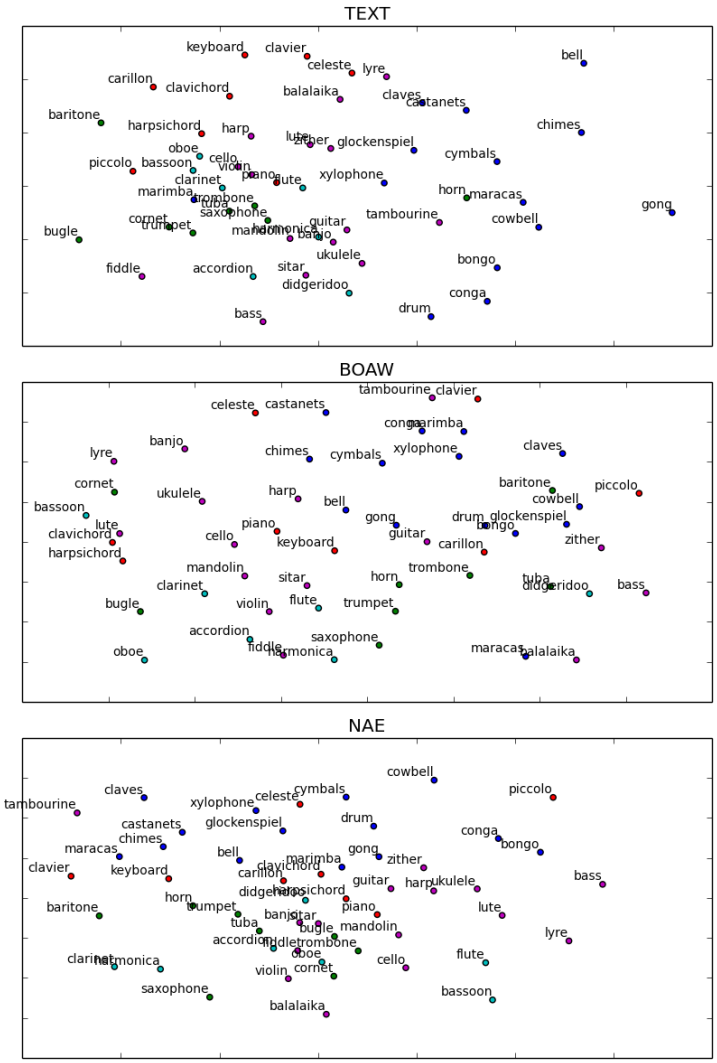

Figure 5.5: Multi-dimensional scaling of instrument representations.

the one based on NAEs. Percussive instruments appear relatively easy to pick out using the auditory signal (e.g. cluster 5 for NAE-INST), except for some of the obvious ones (drums, bongos, gongs). Piano-based instruments (e.g. cluster 1 for BOAW and cluster 3 for NAE-INST) are also grouped together, but that cluster interestingly never includes the piano instrument.

The differences between the representations, together with the cluster assignments, can be visualized through multi-dimensional scaling (Hout et al., 2013): Figure 5.5 shows the instruments over the first two components, which shows how neatly some of the instruments are clustered.

5.3.3 Acoustic similarity

The idea that learned representations can also shed light on cognitive questions goes back at least to Landauer and Dumais (1997), and was reiterated by Lenci (2008) specifically for distributional semantics models. This is probably even more the case for grounded distributional models such as discussed here. In particular, multi-modal representations open up interesting possibilities for interdisciplinary studies between psychological, neuro-

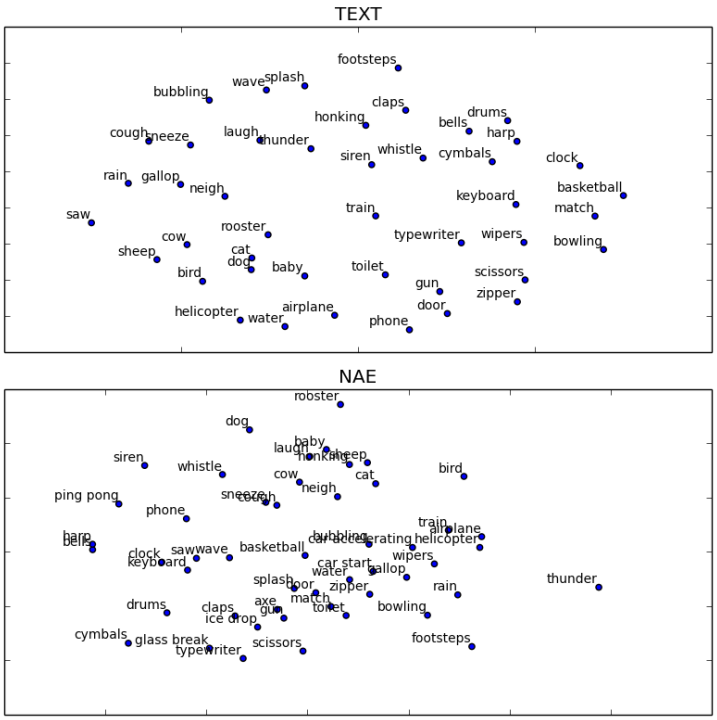

Figure 5.6: Multi-dimensional scaling of environmental sounds

logical and computational representation learning approaches (Kriegeskorte et al., 2008).

Cognitive psychologists have been interested in similarity and categorization judgments of auditory inputs. The auditory representations that we learn here may be used in a variety of cognitive science experiments. To illustrate this, we compare learned representations with an experiment from Gygi et al. (2007) where multi-dimensional scaling is used to examine categorizations of environmental sounds. Gygi et al. (2007) find some clearly defined groups in their cognitive studies: impacts, continuous sounds, and vocalizations and signals.

Figure 5.6 plots the first two components in multi-dimensional scaling for both the textual and the NAE representations. The NAE-ENV classifier was trained on the same set of sound labels (not the same sound files), so we can study how sounds are categorized with such representations and whether this matches the psychological sound groups. It can be seen that the textual model clusters neatly in terms of relatedness—e.g., for instruments and animal sounds. The clusters, however, do not apply to auditory perception: a *gallop* and a *neigh* do not at all sound similar. The NAE representations are much more intuitive in that sense: a *sneeze* and a *cough* sound similarly, a *baby*'s sound is a *laugh* which (arguably) sounds quite similar to a *sheep*'s "baa".

When examining whether Gygi et al. (2007)'s clearly defined cognitive groups can also be found here, we notice that while impact sounds are not clustered for textual representations, they are clearly grouped in the bottom left corner for NAEs. In a similar fashion, signals and vocalizations (which include animal sounds) are grouped at the top. The set of continuous sounds is a harder to identify, but there is a clear group of machine-sounds (trains, airplanes, helicopters, cars) at the bottom of the center.

In short, this preliminary multi-dimensional scaling analysis is remarkably similar to psychological findings, which indicates that auditory representations are not only useful for improving representations to be used in semantics tasks, but also for cognitive science

experiments that involve auditory data (and if that observation applies to audio data, it probably also applies to vision and other modalities).

5.4 Discussion

In these experiments we have relied on FreeSound and its community efforts in uploading and tagging sound files. Although we did somewhat restrict the queries, the reliance on FreeSound may explain some of our findings. The fact that *gong* is in its own cluster for BOAW and NAE-INST, for instance, seems to indicate that the audio samples already make it an outlier, as opposed to gongs having some special properties. In that respect, it is all the more interesting that such relatively noisy sound signals already lead to considerable semantic improvements, especially on auditory-relevant tasks. A better, or more cleaned up, source of auditory data (e.g., with more stringent labelling or with outliers removed) might increase representational quality further.

The auditory representations learned here could be used in a variety of audio-related tasks that are not necessarily related to semantics, from musical preference prediction to identifying environmental background noise in video. We chose to evaluate on semantic relatedness in this case, because it shows how well the learned representations reflect human similarity and relatedness judgments. This type of intrinsic evaluation has long been used as an indicator of representation quality. However, such similarity and relatedness judgment datasets are not modality-specific, which means that they are susceptible to priming, i.e., if a previous comparison was very clearly visual, e.g. *bright-light*, subjects might rely more on the visual modality for judging the next comparison. Furthermore, the dominance of vision in perceptually grounded cognition (Gazzaniga, 1995) probably biases similarity and relatedness judgments of concrete word pairs towards that modality. This might explain why visual grounding yields higher relative improvements than auditory grounding. In cases where auditory information is relevant, auditory grounding leads to large improvements, which merits further exploration of that particular area.

5.5 Conclusions

We have studied grounding semantic representations in raw auditory perceptual information, using a bag of audio words model and neural audio embeddings (NAEs) transferred from a convolutional neural network. NAEs were obtained by extracting the final layer from networks trained on audio recognition tasks. The auditory representations were compared to textual representations and combined with them using a variety of fusion strategies. Following previous work in multi-modal semantics, we evaluated on conceptual similarity and relatedness datasets and performed a detailed analysis of our findings. To show the applicability of auditory representations to auditory-relevant tasks, we examined musical instrument clustering. To show how such auditory representations might be used in cognitive science studies, we performed a preliminary analysis comparing learned representations with psychological acoustic similarity experiments. We found that multi-modal representations perform much better than auditory or textual representations on musical instrument clustering, and that NAEs are very useful for cognitive modeling of auditory perception, closely mirroring human categorizations of audio signals.

CHAPTER 6

Olfactory grounding

We can repeat the question from the previous chapter: if our objective is to ground semantic representations in perceptual information, why stop at image data and sound files? Visual and auditory data are easiest to obtain, to be sure, but the meaning of *lavender* is probably more grounded in its smell than in the visual properties of the flower that produces it. Olfactory (smell) perception is of particular interest for grounded semantics because it is much more primitive compared to the other perceptual modalities (Carmichael et al., 1994; Krusemark et al., 2013). As a result, natural language speakers might take aspects of olfactory perception "for granted", which would imply that text is a relatively poor source of such perceptual information. A multi-modal approach would overcome this problem, and might prove useful in, for example, metaphor interpretation (*the sweet smell of success; rotten politics*) and cognitive modelling, as well as in real-world applications such as automatically retrieving smells or even producing smell descriptions. Here, we explore grounding semantic representations in olfactory perception.

We obtain olfactory representations by constructing a novel bag of chemical compounds (BoCC) model. We evaluate on well known conceptual similarity and relatedness tasks and on zero-shot learning through induced cross-modal mappings. To our knowledge this is the first work to explore using olfactory perceptual data for grounding linguistic semantic models.

6.1 Tasks

The performance of olfactory representations is evaluated on two standard multi-modal evaluation tasks.

6.1.1 Conceptual similarity and relatedness

We evaluate performance on SimLex-999 (Hill et al., 2015) and MEN (Bruni et al., 2014). In the previous chapter we constructed an auditory-relevant dataset, motivated by evidence that in some cases the introduction of perceptual information may be detrimental to performance. The same is likely to be true for other perceptual modalities: in the case of a comparison such as *lily-rose*, the olfactory modality certainly is meaningful, while this is probably not the case for *skateboard-swimsuit*. Some examples of relevant pairs are in Table 6.1.

Olfactory-Relevant Examples					
MEN		sim	SimLex-999		sim
bakery	bread	0.96	steak	meat	0.75
grass	lawn	0.96	flower	violet	0.70
dog	terrier	0.90	tree	maple	0.55
bacon	meat	0.88	grass	moss	0.50
oak	wood	0.84	beach	sea	0.47
daisy	violet	0.76	cereal	wheat	0.38
daffodil	rose	0.74	bread	flour	0.33

Table 6.1: Examples of pairs in the evaluation datasets where olfactory information is relevant, together with the gold-standard similarity score.

MEN	3000
OMEN	311
SimLex	999
OSLex	65

Table 6.2: Number of pairwise comparisons for the datasets and their olfactory-relevant subsets.

Hence, just like in the auditory case in the previous chapter, we had two annotators rate the two datasets according to whether smell is relevant to the pairwise comparison. The annotation criterion was as follows: if both concepts in a pairwise comparison have a distinctive associated smell, then the comparison is relevant to the olfactory modality. Only if both annotators agree is the comparison deemed olfactory-relevant. This annotation leads to a total of four evaluation sets: the MEN test collection and its olfactory-relevant subset **OMEN**; and the SimLex-999 dataset and its olfactory-relevant subset **OSLex**. See Table 6.2. The inter-annotator agreement on the olfactory relevance judgments was high ($\kappa = 0.94$ for the MEN test collection and $\kappa = 0.96$ for SimLex-999).

6.1.2 Cross-modal zero-shot learning

Cross-modal semantics, instead of being concerned with improving semantic representations through grounding, focuses on the problem of reference (see Chapter 2). Using, for instance, mappings between visual and textual space, the objective is to learn which words refer to which objects (Lazaridou et al., 2014). This problem is very much related to the object recognition task in computer vision, but instead of using just visual data and labels, these cross-modal models also utilize textual information (Socher et al., 2014; Frome et al., 2013). This approach allows for *zero-shot learning*, where the model can predict how an object relates to other concepts just from seeing an image of the object, but without ever having seen the object previously (Lazaridou et al., 2014).

We evaluate cross-modal zero-shot learning performance through the average percentage correct at N (P@N), which measures how many of the test instances were ranked

		Chemical Compound				
		Phenethyl acetate	Isoamyl butyrate	Anisyl butyrate	Myrcene	Syringaldehyde
Smell label	Melon	✓	✓			
	Pineapple	✓				✓
	Licorice			✓		
	Anise			✓	✓	
	Beer				✓	✓

Table 6.3: A BoCC model.

within the top N highest ranked nearest neighbors. A chance baseline is obtained by randomly ranking a concept's nearest neighbors. We use partial least squares regression (PLSR) to induce cross-modal mappings from the linguistic to the olfactory space and vice versa.[1]

Due to the nature of the olfactory data source (see Section 6.2), it is not possible to build olfactory representations for all concepts in the test sets. However, cross-modal mappings yield an additional benefit: since linguistic representations have full coverage over the datasets, we can project from linguistic space to perceptual space to also obtain full coverage for the perceptual modalities. This technique has been used to increase coverage for feature norms (Fagarasan et al., 2015). Consequently, we are in a position to compare perceptual spaces directly to each other, and to linguistic space, over the entire dataset, as well as on the relevant olfactory subsets. When projecting into such a space and reporting results, the model is prefixed with an arrow ($\rightarrow$) in the corresponding table.

6.2 Olfactory perception

The Sigma-Aldrich Fine Chemicals flavors and fragrances catalog[2] (henceforth SAFC) is one of the largest publicly accessible databases of semantic odor profiles that is used extensively in fragrance research (Zarzo and Stanton, 2006). It contains organoleptic labels and the chemical compounds—or more accurately the perfume raw materials (PRMs)—that produce them. By automatically scraping the catalog we obtained a total of 137 organoleptic smell labels from SAFC, with a total of 11,152 associated PRMs. We also experimented with Flavornet[3] and the LRI and odour database[4], but found that the data from these were more noisy and generally of lower quality.

[1]To avoid introducing another parameter, we set the number of latent variables in the cross-modal PLSR map to a third of the number of dimensions of the perceptual representation.

[2]http://www.sigmaaldrich.com/industries/flavors-and-fragrances.html

[3]http://www.flavornet.org

[4]http://www.odour.org.uk

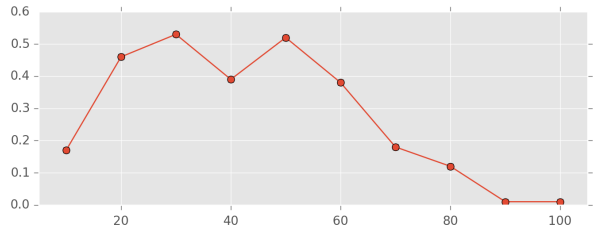

Figure 6.1: Performance of olfactory representations when using SVD to reduce the number of dimensions.

Dataset	Linguistic	BoCC-Raw	BoCC-SVD
OMEN (35)	0.40	0.42	0.53

Table 6.4: Comparison of olfactory representations on the covered OMEN dataset.

For each of the smell labels in SAFC we count the co-occurrences of associated chemical compounds, yielding a bag of chemical compounds (BoCC) model. Table 6.3 shows an example sub-space of this model. Although the SAFC catalog is considered sufficiently comprehensive for fragrance research (Zarzo and Stanton, 2006), the fact that PRMs usually occur only once per smell label means that the representations are rather sparse. Hence, we apply dimensionality reduction to the original representation to get denser vectors. We call the model without any dimensionality reduction BoCC-Raw and use singular value decomposition (SVD) to create an additional BoCC-SVD model with reduced dimensionality. Positive pointwise mutual information (PPMI) weighting is applied to the raw space before performing dimensionality reduction.

The number of dimensions in human olfactory space is a hotly debated topic in the olfactory chemical sciences (Buck and Axel, 1991; Zarzo and Stanton, 2006). Recent studies involving multi-dimensional scaling on the SAFC catalog revealed approximately 32 dimensions in olfactory perception space (Mamlouk et al., 2003; Mamlouk and Martinetz, 2004). We examine this finding by evaluating the Spearman ρ_s correlation on the pairs of OMEN that occur in the SAFC database (35 pairs). The coverage on SimLex was not sufficient to also try that dataset (only 5 pairs). Figure 6.1 shows the results. It turns out that the best olfactory representations are obtained with 30 dimensions. In other words, our findings appear to corroborate recent evidence suggesting that olfactory space (at least when using SAFC as a data source) is best modeled using around 30 dimensions.

6.2.1 Linguistic representations

As before, for the linguistic representations we use the continuous vector representations from the log-linear skip-gram model of Mikolov et al. (2013a), specifically the 300-dimensional vector representations trained on part of the Google News dataset (about 100 billion words) that have been released on the Word2vec website.[5]

[5]https://code.google.com/archive/p/word2vec/

	MEN	OMEN	SLex	OSLex
Linguistic	0.78	0.38	0.44	0.30
→BoCC-Raw	0.38	0.36	0.19	0.23
→BoCC-SVD	0.46	0.51	0.23	0.48
Multi-modal	0.69	0.53	0.40	0.49

Table 6.5: Comparison of linguistic, olfactory and multi-modal representations.

Mapping	P@1	P@5	P@20	P@50
Chance	0.0	3.76	13.53	36.09
Olfactory ⇒ Ling.	1.51	8.33	24.24	47.73
Ling. ⇒ Olfactory	4.55	15.15	43.18	67.42

Table 6.6: Zero-shot learning performance for BoCC-SVD.

6.2.2 Conceptual similarity

Results on the 35 covered pairs of OMEN for the two BoCC models are reported in Table 6.4. Olfactory representations outperform linguistic representations on this subset. In fact, linguistic representations perform poorly compared to their performance on the whole of MEN. The SVD model performs best, improving on the linguistic and raw models with a 33% and 26% relative increase in performance, respectively.

We use a cross-modal PLSR map, trained on all available organoleptic labels in SAFC, to extend coverage and allow for a direct comparison between linguistic representations and cross-modally projected olfactory representations on the entire datasets and relevant subsets. The results are shown in Table 6.5. As might be expected, linguistic performs better than olfactory on the full datasets. On the olfactory-relevant subsets, however, the projected BoCC-SVD model outperforms linguistic for both datasets. Performance increases even further when the two representations are combined into a multi-modal representation by concatenating the L2-normalized linguistic and olfactory (→BoCC-SVD) vectors.

6.2.3 Zero-shot learning

We learn a cross-modal mapping between the two spaces and evaluate zero-shot learning. We use all 137 labels in the SAFC database that have corresponding linguistic vectors for the training data. For each term, we train the map on all other labels and measure whether the correct instance is ranked within the top N neighbors. We use the BoCC-SVD model for the olfactory space, since it performed best on the conceptual similarity task. Table 6.6 shows the results. It appears that mapping linguistic to olfactory is easier than mapping olfactory to linguistic, which may be explained by the different number of dimensions in the two spaces. One could say that it is easier to find the chemical composition of a "smelly" word from its linguistic representation, than it is to linguistically represent or describe a chemical composition.

apple	bacon	brandy	cashew
pear	smoky	rum	hazelnut
banana	roasted	whiskey	peanut
melon	coffee	wine-like	almond
apricot	mesquite	grape	hawthorne
pineapple	mossy	fleshy	jam
chocolate	lemon	cheese	caramel
cocoa	citrus	grassy	nutty
sweet	geranium	butter	roasted
coffee	grapefruit	oily	maple
licorice	tart	creamy	butterscotch
roasted	floral	coconut	coffee

Table 6.7: Example nearest neighbors for BoCC-SVD representations.

6.2.4 Qualitative analysis

We also examined the BoCC representations qualitatively. As Table 6.7 shows, the nearest neighbors are remarkably semantically coherent. The nearest neighbors for *bacon* and *cheese*, for example, accurately sum up how one might describe those smells. The model also groups together nuts and fruits, and expresses well what *chocolate* and *caramel* smell (or taste) like.

6.3 Conclusions

We have studied grounding semantic representations in raw olfactory perceptual information. We used a bag of chemical compounds model to obtain olfactory representations and evaluated on conceptual similarity and cross-modal zero-shot learning, with good results.

This work opens up interesting possibilities in analyzing smell and even taste. It could be applied in a variety of settings beyond semantic similarity, from chemical information retrieval to metaphor interpretation to cognitive modelling. A speculative blue-sky application based on this, and other multi-modal models, would be an NLG application describing a wine based on its chemical composition, and perhaps other information such as its color and country of origin.

A theme in this thesis has been to show that deep learning methods outperform more traditional bag of words approaches. In this case the same idea is likely to apply, but we are faced with the difficulty that there is not enough data available. Deep learning approaches work particularly well when there is a lot of data available. This is not the case here, with only 137 organoleptic labels available in SAFC. It would be very interesting to use neural networks to learn smell repesentations, i.e., a *smell2vec*, but we would first need more data.

It may well be the case that the auditory and olfactory modalities are better suited for other evaluations or particularly useful in specific downstream tasks, but we have chosen

to follow standard evaluations in multi-modal semantics to allow for a direct comparison. A central question for the second part of this thesis is, why stop at the visual modality? We hope to have shown that similar advances to those achieved by visually grounded models may be possible with non-visually grounded models as well. Our findings point towards fruitful applications of grounded representations in real downstream tasks that relate to audio and olfaction, as well as to a wholly unexplored area of linking grounded representations with cognitive studies. This will, hopefully, ultimately lead to perceptually grounded models in artificial intelligence that rely on data from all modalities, as a unified model that captures human semantic knowledge and experience.

Part IV

Discussion & conclusions

CHAPTER 7

Discussion

This thesis is an exponent of the recent trend in artificial intelligence to move towards more interdisciplinary research involving multiple modalities. Many of AI's subdisciplines have matured considerably. Interdisciplinary research provides an interesting new frontier for the field.

While many of the core tasks in natural language processing remain essentially linguistic by nature, in many cases improved language understanding can lead to improved performance. Grounding is a natural way to improve concept-level understanding: it has sound theoretical motivations, as discussed in Chapter 2, and clearly leads to practical improvements as well, as evidenced by the higher-quality representations obtained in Chapter 3, and their applications as described in Chapter 4.

The uni-modal focus of AI research, where a subdiscipline focuses on a single modality, has improved narrow AI quite dramatically. Ultimately, however, if we want to achieve artificial general intelligence (AGI), systems will have to be multi-modal almost by necessity, performing grounding in complex perceptual environments at every step.

7.1 Full virtual embodiment

One of the aims of this work has been to show that grounding is possible and leads to better representations. The other has been to show that this grounding need not be limited to the visual modality: any (perceptual) modality can be used. In fact, while vision is clearly the dominant perceptual modality, modalities like olfaction are much harder to capture in words, implying that olfactory aspects of meaning are much harder to learn from linguistic corpora or from easily accessible data on the Web. This has repercussions all across natural language understanding, from metaphor detection to modality-specific meaning representation.

While gustatory perception is closely related to olfactory perception and can be modeled using the same methods (both being essentially chemical receptors), the obvious perceptual modality that is still missing from the current treatment of grounding is haptic or tactile perception. Even more so than in the olfactory or gustatory case, such information is hard to obtain from a passive data source: that is, more than any of the other modalities, it requires active embodiment. If we want true grounding, in the most human of ways, we would probably need an actively embodied agent—e.g., a robot—who learns the meaning of "bumping into a wall" by actually bumping into a wall. There have

been studies of grounding in robotics (Fitzpatrick and Metta, 2003; Coradeschi et al., 2013). However, one might argue that having a robot learn meaning by "bumping into things" is possibly not the best way to develop semantics from the ground up, given current technological limitations and how long it would take to learn things that way.

In short, what we would want is a fully embodied agent, that gets input from various perceptual modalities simultaneously, that is, jointly, and can learn from that. An interesting possibility, that has these joint input characteristics, is virtual or augmented reality, where we either know the constraints of the virtual world, or know what an agent is looking at. This has the limitation, however, that it is fully reliant on human involvement: ideally, we would have a way to learn from humans, but not need to have them around all the time to teach us things. Recent developments in deep reinforcement learning (Mnih et al., 2015; Silver et al., 2016) point the way towards agents learning from each other: video games are the ideal platform for full virtual embodiment.

Video games make a lot of sense as AI's next frontier: the real world is enormously complex, and performing common sense reasoning in such a complicated environment has long been one of the classic AI problems (McCarthy and Hayes, 1969, often called "the frame problem"). Video games have the benefit that we can start with relatively simple games, and make them progressively more complicated as our capabilities advance. They are also ideally suited for artificial agents to play against and learn *from humans*, but to also play against and learn from *each other*. Ethically speaking, focusing attention on fully capable, generally intelligent (artificial general intelligence, AGI), agents in a virtual world has a benefit that is arguably of existential importance to human kind: we can learn to define the ethical guidelines by which artificial agents must function in a virtual world, so as to ensure that artificial agents behave as desired in the real world. This would, for instance, weed out "paperclip maximizers" that destroy the world by following an erroneous objective function (Bostrom, 2003). It would also frame the debate about the threats of AI and allow for riskless experimentation with regard to defining sensible ethical constraints. It would constitute the ultimate "game with a purpose". It is worthwhile outlining what the desired properties of such a video game (or such video games) would be. Such a game should have the following necessary and sufficient properties:

- Multi-player: the game needs to support multiple agents (human or artificial), interacting in the game world.
- Mixed agency: it should be playable by (and enjoyable for) humans, but humans should not be essential to game play, so that machines can learn from each other.
- Varying degree of state-access: the state of the game world should be accessible in varying degrees. The more complex the game world, the more access can initially be given to the game state, or a subset thereof. A hard constraint on game state access, as some have advocated (Narasimhan et al., 2015), is probably not necessary.
- Multi-modal synchrony: input from different modalities (vision, audio, language) should occur simultaneously to allow for joint learning.
- Non-deterministic: the game should not be learnable by learning fixed patterns, its environment should be non-deterministic.
- Concrete goals with manageable rules: the goals and rewards should be clearly defined within the game (though not necessarily be accessible to agents), and not

be too far removed from achieving a given state or performing a certain action. The rules to obtain goals should not be overly complex.

- No single objective: there should not be a single objective function which can be optimized, but rather a set of objective functions (or one weighted super-objective function). One of the characteristics of narrow AI is that it tends to have a single objective, which should not apply in the general case.
- Planning and complex strategies: goals should involve a good deal of planning and obtaining goals should not be easy but rather require complex reasoning.
- Level playing field with human bias: the game should aim to provide for a level playing field between humans and artificial agents, which means that e.g. superior memory of machines should not affect in-game performance. A human bias ensures machines must learn from humans. Introducing the notion of "common sense" would be an obvious human bias.
- Benign: a first-person shooter game is not what we want artificial agents to excel at, the objectives of the game should be benign in nature and inspire confidence, if only to alleviate potential hostility from e.g. the media.
- Language-heavy: language should be viewed as a *sine qua non* for intelligent agents. This differentiates narrow AI, which does not necessarily require language, from the type of general AI that could ultimately be developed within the game's constraints. The true test of artificial intelligence is language capability: it should not be seen as one of many expressions of intelligence, but rather as an essential aspect thereof. It is too easy to take an intentional stance (Dennett, 1989) towards agents without language capabilities and overestimate the level of intelligence achieved.
- Bot-friendly: contrary to most games, the game should welcome bots (i.e., artificial agents). If the parameters are defined in such a way that humans can still win, this would only improve game play: consider being the captain of a bot-squad, where the bots understand the captain's commands.

No game with the above characteristics currently exists. However, this set of properties might serve as a guide for future development. If artificial intelligence is inevitable, as some seem to think (Kurzweil, 2006), it is worthwhile fleshing out how we should grow these capabilities, and within what set of constraints.

7.2 Open problems

Farfetched as it may sound, if we're ever going to decipher the language of dolphins and whales[1], or are ever going to communicate with alien species, it is going to be through some form of grounding and embodiment. The shared visual space that was used for bilingual lexicon induction is a good example of how this would work: communication becomes meaningful with respect to a shared perceptual environment with grounded symbols. That is, grounding is an absolutely central aspect of meaning, that is still too often overlooked.

[1]This is actually not that farfetched: there have been studies where wearables were attached to dolphins' heads to allow for grounded communication (Kohlsdorf et al., 2013).

If we are solely interested in specific applications and in beating the current state-of-the-art on a somewhat contrived NLP task, then we may be excused, but if we are interested in more general AI problems, then multi-modal research is nothing but essential. That said, there are a few important open questions that will need to be answered, specifically in the context of multi-modal semantics:

- The problem of fusion: fusion is often still done in very crude ways, making assumptions about the availability and applicability of multi-modal data. As we have seen, perceptual information is not relevant for every meaning: it depends on the concept, and on the task that the representation is applied to. A better approach would be to learn a single multi-modal space, as is done e.g. by Lazaridou et al. (2015b), but with more perceptual modalities, and with a more sophisticated scheme for presenting data to the learning algorithm, for instance by exploiting curriculum learning (Bengio et al., 2009). An alternative is to first independently learn uni-modal representations but then separately learn a method for combining these into a single multi-modal one. A first attempt at this was the work of Silberer and Lapata (2014), but it may be better to learn to combine representations for the task at hand, instead of simply learning a single-space representation of lower dimensionality. Ultimately, these are questions about the correct level of fusion, which is something that cognitive science may shed light on, or, conversely, that may lead to new insights in cognitive science.

- The problem of extra information: do multi-modal representations lead to improvements because they introduce extra information, or because perceptual information is complementary to what we can learn from linguistic corpora? The answer is probably that it is a combination of both, but it is important to establish how this trade-off works: can we get away with just having more and more linguistic data, or is perceptual input, as has been assumed in this thesis, a sine qua non for human-level meaning representation?

- The problem of transfer learning: transfer learning has some great benefits, since it allows us to have very high-quality representations, that do not require end-to-end learning for each task. It is an open question whether, and if so to what extent, this may be detrimental to performance. Should we always learn end-to-end, if we have the option, or is the adverse effect of transfer learning negligible?

- The problem of representational quality: how is it that transferred visual representations work so well? What is it exactly that FC7 captures that allows for representing meaning in a way that is close to, or sometimes even outperforms, linguistic methods? Are there ways to improve this, e.g. by discarding certain images, or by selecting appropriate segments of images? Or is every part of an image equally relevant to meaning, that is, does the fact that images of mountains often depict blue skies at the top contribute to the meaning of *mountain*?

- The problem of compositionality: this thesis has largely been concerned with word-level representations. How do we extend these findings to phrases and sentences, or even documents? Is there a difference between visual or perceptual and linguistic composition? Does composition even mean the same thing in these different modalities?

- The problem of context: we should not limit ourselves to a single concept-level multi-modal representation that is to be used in every task. Can we improve things by introducing attention, or memory, into the process, so that we can contextualize representations based on the task that they are being used in?

There are many more open questions, but these are some of the most pertinent ones.

CHAPTER 8

Conclusions

This thesis has been concerned with advancing the field of multi-modal semantics.

First, it was shown that better visually grounded representations can be obtained through deep convolutional neural networks than by using the standard bag of visual words approach. Such features were demonstrated to work much better, a finding that was extended to different neural network architectures that share similar characteristics. Furthermore, search engines such as Google and Bing were found to yield images of a similarly high quality as the carefully human-annotated ImageNet dataset. These novel representations were subsequently successfully applied to two natural language processing problems: lexical entailment and bilingual lexicon induction.

Second, grounding was taken beyond the visual modality and into the previously unexplored territory of the auditory and olfactory perceptual modalities. In the case of auditory perception, bag of audio words was introduced as a baseline method for performing auditory grounding. Along the same lines as in the vision case, deep learning methods, through what we called neural auditory embeddings, yielded even better representations. A first attempt at grounding in olfactory perception was achieved as well, through a bag of chemical compounds model.

Finally, full virtual embodiment through video games was proposed as a new frontier for AI research and the properties of such video games were discussed. In addition, several open problems were raised that multi-modal semantics should address.

8.1 Main findings

The main findings of this thesis are as follows:

- Transferred convolutional neural network layers work better than bag of visual words representations for modelling similarity and relatedness.
- The type of network does not matter very much, but the data source for images matters a lot. Search engines are similar in performance to human-annotated datasets such as ImageNet, but have the advantage that they have better coverage.
- Visual representations may be applied to the tasks of lexical entailment and bilingual lexicon induction, with excellent performance in both cases. This shows multi-modal research's potential for improving the state-of-the-art.

- Grounding can successfully be achieved in auditory perception, through bag of audio words as well as through neural auditory embeddings.
- For the olfactory modality, grounding can be successfully performed through bag of chemical compounds models.

8.2 Future work

There are many ways to improve multi-modal semantics, at each step of the process. Polymodal fusion, for instance, where we have more than one perceptual modality, is an important issue: now that we have non-visual modalities as well, how do we fuse all of these uni-modal representations into a multi-modal whole? Should we include just as much information from each modality, or should we learn to decide which modality to include? Many of these questions were already discussed in the previous chapter.

One particular area where the field needs to improve is on evaluations: intrinsic evaluations are not always good indicators of performance in extrinsic downstream tasks (Faruqui et al., 2016). It will become more and more necessary to construct datasets with an explicit focus on multi-modality, or with a more general focus on the type of "common sense" reasoning that humans excel at, in part because of their grounding in (the same) complex perceptual environments.

Bibliography

Agirre, E., Alfonseca, E., Hall, K. B., Kravalova, J., Pasca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *NAACL*, pages 19–27.

Andrews, M., Vigliocco, G., and Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2015). Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *arXiv preprint arXiv:1502.03520*.

Atkin, A. (2013). Peirce's theory of signs. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Baayen, R. H. and Lieber, R. (1996). Word frequency distributions and lexical semantics. *Computers and the Humanities*, 30(4):281–291.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. Association for Computational Linguistics.

Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Baroni, M., Bernardi, R., Do, N.-Q., and Shan, C.-C. (2012). Entailment above the word level in distributional semantics. In *Proceedings of EACL*, pages 23–32.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Baroni, M. and Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55—88.

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

Baroni, M. and Lenci, A. (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop*, pages 1–10.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences*, 22(04):637–660.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59(1):617–645.

Bay, H., Ess, A., Tuytelaars, T., and Gool, L. J. V. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359.

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

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., et al. (2007). Greedy layer-wise training of deep networks. *Advances in neural information processing systems (NIPS)*, 19:153—160.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM.

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L. (2006). Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer.

Berg, T. L., Berg, A. C., and Shih, J. (2010). Automatic attribute discovery and characterization from noisy web data. In *Proceedings of ECCV*, pages 663–676. Springer.

Bergsma, S. and Goebel, R. (2011). Using visual information to predict lexical preference. In *Proceedings of RANLP*, pages 399–405.

Bergsma, S. and Van Durme, B. (2011). Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI*, pages 1764–1769.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *arXiv preprint arXiv:1601.03896*.

Biran, O. and McKeown, K. (2013). Classifying taxonomic relations between pairs of wikipedia articles. In *Proceedings of IJCNLP*, pages 788–794.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M.-G., Ruel, J., Venuti, P., and Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development*, 75(4):1115–1139.

Bosch, A., Zisserman, A., and Muñoz, X. (2007). Image classification using random forests and ferns. In *Proceedings of ICCV*, pages 1–8.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pages 277–284.

Bruni, E., Boleda, G., Baroni, M., and Tran, N. (2012). Distributional semantics in technicolor. In *ACL*, pages 136–145.

Bruni, E., Tran, G. B., and Baroni, M. (2011). Distributional semantics from text and images. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 22–32. Association for Computational Linguistics.

Bruni, E., Tran, N., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artifical Intelligence Research*, 49:1–47.

Buck, L. and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65(1):175–187.

Bulat, L., Kiela, D., and Clark, S. (2016). Vision and Feature Norms: Improving automatic feature norm learning through cross-modal maps. In *Proceedings of NAACL-HLT 2016*, San Diego, CA.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods*, 44(3):890–907.

Cangelosi, A. and Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive science*, 30(4):673–689.

Carmichael, S. T., Clugnet, M.-C., and Price, J. L. (1994). Central olfactory connections in the macaque monkey. *Journal of Comparative Neurology*, 346(3):403–434.

Christiansen, M. H. and Chater, N. (1992). Connectionism, learning and meaning. *Connection Science*, 4(3-4):227–252.

Christiansen, M. H. and Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences*, 5(2):82–88.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Clark, A. (1999). An embodied cognitive science? *Trends in cognitive sciences*, 3(9):345–351.

Clark, S. (2015). Vector Space Models of Lexical Meaning. In Lappin, S. and Fox, C., editors, *Handbook of Contemporary Semantic Theory*, chapter 16. Wiley-Blackwell, Oxford.

Clarke, D. (2009). Context-theoretic semantics for natural language: An overview. In *Proceedings of the GEMS 2009 Workshop*, pages 112–119.

Cole, D. (2015). The Chinese Room argument. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2015 edition.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Coradeschi, S., Loutfi, A., and Wrede, B. (2013). A short review of symbol grounding in robotic and intelligent systems. *KI-Künstliche Intelligenz*, 27(2):129–136.

Curran, J. R. (2004). *From distributional to semantic similarity*. PhD thesis, College of Science and Engineering: School of Informatics, University of Edinburgh.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Deselaers, T. and Ferrari, V. (2011). Visual and semantic similarity in ImageNet. In *Proceedings of CVPR*, pages 1777–1784.

Deselaers, T., Keysers, D., and Ney, H. (2008). Features for image retrieval: An experimental comparison. *Information Retrieval*, 11(2):77–107.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655.

Driancourt, X. and Bottou, L. (1990). TDNN-extracted features. In *Neuro Nimes 90*.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of ACL*, page 216.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6(10):635–653.

Eronen, A. (2003). Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications*, volume 2, pages 133–136.

Evans, V. (2015). *The Crucible of Language: How Language and Mind Create Meaning*. Cambridge University Press, Cambridge, UK.

Fagarasan, L., Vecchi, E. M., and Clark, S. (2015). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, pages 52–57, London, UK.

Fairchild, M. D. (2005). Status of CIE color appearance models. In *Proceedings of the International Society for Optical Engineering (SPIE)*.

Farah, M. J. and McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120(4):339—357.

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.*, pages 1778–1785. IEEE.

Faruqui, M. and Dyer, C. (2014). Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, USA. Association for Computational Linguistics.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*.

Feng, Y. and Lapata, M. (2010). Visual information in semantic representation. In *Proceedings of NAACL*, pages 91–99.

Fergus, R., Li, F., Perona, P., and Zisserman, A. (2005). Learning object categories from Google's image search. In *Proceedings of ICCV*, pages 1816–1823.

Ferrari, V. and Zisserman, A. (2007). Learning visual attributes. In *Advances in Neural Information Processing Systems*, pages 433–440.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116—131.

Firth, J. R. (1957). *A synopsis of linguistic theory*. Blackwell.

Fitzpatrick, P. and Metta, G. (2003). Grounding vision through experimental manipulation. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 361(1811):2165–2185.

Flanagan, J. L. (2013). *Speech analysis synthesis and perception*, volume 3. Springer Science & Business Media.

Font, F., Roma, G., and Serra, X. (2013). Freesound technical demo. In *Proceedings of the 21st acm international conference on multimedia*, pages 411–412. ACM.

Foote, J. T. (1997). Content-based retrieval of music and audio. In *Voice, Video, and Data Communications*, pages 138–147.

Frege, G. (1892). Über sinn und bedeutung.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013). DeViSE: A Deep Visual-Semantic Embedding Model. In *Proceedings of NIPS*, pages 2121–2129.

Fu, R., Guo, J., Qin, B., Che, W., Wang, H., and Liu, T. (2014). Learning semantic hierarchies via word embeddings. In *Proceedings of ACL*, pages 1199–1209.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202.

Fung, P. and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *ACL*, pages 414–420.

Gallese, V. (2007). Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1480):659—669.

Gallese, V. and Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, 22(3-4):455—479.

Garrette, D., Erk, K., and Mooney, R. (2011). Integrating logical representations with probabilistic information using Markov logic. In *Proceedings of IWCS*, pages 105–114.

Gaussier, É., Renders, J., Matveeva, I., Goutte, C., and Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of ACL*, pages 526–533.

Gazzaniga, M. S., editor (1995). *The Cognitive Neurosciences*. MIT Press, Cambridge, MA.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

Glenberg, A. M. and Kaschak, M. P. (2002). Grounding language in action. *Psychonomic bulletin & review*, 9(3):558–565.

Grefenstette, G. (1994). *Explorations in automatic thesaurus construction*. Kluwer Dordrecht.

Gupta, A., Srinivasan, P., Shi, J., and Davis, L. S. (2009). Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009*, pages 2012–2019. IEEE.

Gygi, B., Kidd, G. R., and Watson, C. S. (2007). Similarity and categorization of environmental sounds. *Perception & psychophysics*, 69(6):839–855.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL*, pages 771–779.

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

Harnad, S. (1993). Grounding symbols in the analog world with neural nets. *Think*, 2(1):12–78.

Harris, Z. (1954). Distributional Structure. *Word*, 10(23):146—162.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Massachusetts Institute of Technology, Cambridge, MA, USA.

Herbelot, A. and Ganesalingam, M. (2013). Measuring semantic content in distributional vectors. In *Proceedings of ACL*, pages 440–445.

Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. *Proceedings of ICLR*.

Hill, F. and Korhonen, A. (2014). Learning abstract concept embeddings from multimodal data: Since you probably can't see what I mean. In *Proceedings of EMNLP*, pages 255–265.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Hinton, G., McClelland, J., and D.E., R. (1986). Distributed representations. In Rurnelhart, D. and McClelland, J., editors, *Parallel distributed processing*, volume 1, pages 77—109. MIT Press, Cambridge, MA.

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

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hotton, S. and Yoshimi, J. (2011). Extending dynamical systems theory to model embodied cognition. *Cognitive Science*, 35(3):444–479.

Hout, M. C., Papesh, M. H., and Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(1):93–103.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 873–882. Association for Computational Linguistics.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press, Cambridge, MA.

Jackendoff, R. (2002). *Foundations of Language*. Oxford University Press, Oxford.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R. B., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678.

Jones, S. S., Smith, L. B., and Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child development*, 62(3):499–516.

Kiela, D. and Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. In *Proceedings of EACL 2014, Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*.

Kiela, D., Hill, F., Korhonen, A., and Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*, pages 835–841.

Kievit-Kylar, B. (2014). *Reading between the domains*. PhD thesis, Indiana University, Bloomington.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *ULA'02 Workshop*, pages 9–16.

Kohlsdorf, D., Gilliland, S., Presti, P., Starner, T., and Herzing, D. (2013). An underwater wearable computer for two way human-dolphin communication experimentation. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ISWC '13, pages 147–148.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. (2010). Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1106–1114.

Krusemark, E. A., Novak, L. R., Gitelman, D. R., and Li, W. (2013). When the sense of smell meets emotion: anxiety-state-dependent olfactory processing and neural circuitry adaptation. *The Journal of Neuroscience*, 33(39):15324–15332.

Kurzweil, R. (2006). *The Singularity Is Near: When Humans Transcend Biology*. Penguin (Non-Classics).

Landau, B., Smith, L., and Jones, S. (1998). Object perception and object naming in early development. *Trends in cognitive sciences*, 2(1):19–24.

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

Lapesa, G. and Evert, S. (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

Lavrenko, V., Choquette, M., and Croft, W. B. (2002). Cross-lingual relevance models. In *SIGIR*, pages 175–182.

Lazaridou, A., Bruni, E., and Baroni, M. (2014). Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Proceedings of ACL*, pages 1403–1414.

Lazaridou, A., Dinu, G., Liska, A., and Baroni, M. (2015a). From visual attributes to adjectives through decompositional distributional semantics. *Transactions of the Association for Computational Linguistics (TACL)*, 3:183—196.

Lazaridou, A., Pham, N. T., and Baroni, M. (2015b). Combining language and vision with a multimodal skipgram model. In *Proceedings of NAACL*.

Lebret, R. and Collobert, R. (2014). Word embeddings through Hellinger PCA. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–490, Gothenburg, Sweden. Association for Computational Linguistics.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436—444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Leech, G., Garside, R., and Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 622–628. Association for Computational Linguistics.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Lenci, A. and Benotto, G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of *SEM*, pages 75–79.

Leong, C. W. and Mihalcea, R. (2011a). Going beyond text: A hybrid image-text approach for measuring word relatedness. In *Proceedings of IJCNLP*, pages 1403–1407.

Leong, C. W. and Mihalcea, R. (2011b). Measuring the semantic relatedness between words and images. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 185–194. Association for Computational Linguistics.

Leviant, I. and Reichart, R. (2015). Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

Levow, G.-A., Oard, D., and Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management*, 41:523 – 547.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Li, Y., Xu, L., Tian, F., Jiang, L., Zhong, X., and Chen, E. (2015). Word embedding revisited: A new representation learning and explicit matrix factorization perspective. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

Lin, D., Church, K. W., Ji, H., Sekine, S., Yarowsky, D., Bergsma, S., Patil, K., Pitler, E., Lathbury, R., Rao, V., Dalwani, K., and Narsale, S. (2010). New tools for Web-scale N-grams. In *LREC*, pages 2221–2227.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *CoRR*, abs/1312.4400.

Liu, X., Duh, K., and Matsumoto, Y. (2013). Topic models + word alignment = A flexible framework for extracting bilingual dictionary from comparable corpus. In *CoNLL*, pages 212–221.

Louwerse, M. M. (2008). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 59(1):617–645.

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

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

Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., and Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. *Proceedings of ICLR*.

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

Mamlouk, A. M., Chee-Ruiter, C., Hofmann, U. G., and Bower, J. M. (2003). Quantifying olfactory perception: Mapping olfactory perception space by using multidimensional scaling and self-organizing maps. *Neurocomputing*, 52:591–597.

Mamlouk, A. M. and Martinetz, T. (2004). On the dimensions of the olfactory perception space. *Neurocomputing*, 58:1019–1025.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

McCarthy, J. and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Readings in artificial intelligence*, pages 431–450.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Meteyard, L., Cuadrado, S. R., Bahrami, B., and Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, 48(7):788–804.

Meteyard, L. and Vigliocco, G. (2008). The role of sensory and motor information in semantic representation: A review. *Handbook of cognitive science: An embodied approach*, pages 293–312.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, Arizona, USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mimno, D. M., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of EMNLP*, pages 880–889.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Mnih, A. and Hinton, G. E. (2009). A scalable hierarchical distributed language model. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1081–1088.

Mnih, A. and Teh, Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of ICLR*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Mohler, M., Bracewell, D., Tomlinson, M., and Hinote, D. (2013). Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the 1st Workshop on Metaphor in NLP*.

Monner, D. D. and Reggia, J. A. (2011). Systematically grounding language through vision in a deep, recurrent neural network. In *Artificial General Intelligence*, pages 112–121. Springer.

Mooney, R. J. (2008). Learning to connect language and perception. In *Proceedings of AAAI*, pages 1598—1601.

Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *Proceedings of AISTATS*, volume 5, pages 246–252.

Murphy, B., Talukdar, P., and Mitchell, T. (2012). Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of 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*, pages 114–123. Association for Computational Linguistics.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of ICML*, pages 807–814.

Narasimhan, K., Kulkarni, T. D., and Barzilay, R. (2015). Language understanding for textbased games using deep reinforcement learning. In *Proceedings of EMNLP*.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88.

Nelson, D. L., McEvoy, C. L., , and Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, 36(3):402–407.

Newell, A. (1980). Physical symbol systems. *Cognitive science*, 4(2):135–183.

Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3):113–126.

Nowak, E., Jurie, F., and Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *Proceedings of ECCV*, pages 490–503. Springer.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of CVPR*, pages 1717–1724.

O'Shaughnessy, D. (1987). *Speech communication: human and machine*. Addison-Wesley series in electrical engineering: digital signal processing. Universities Press (India) Pvt. Limited.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1410–1418.

Paşca, M., Lin, D., Bigham, J., Lifchits, A., and Jain, A. (2006). Names and similarities on the web: fact extraction in the fast lane. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 809–816. Association for Computational Linguistics.

Pascanu, R., Mikolov, T., and Bengio, Y. (2012). On the difficulty of training recurrent neural networks. *Proceedings of ICLR*.

Peirce, C. (1931/1936). *The Collected Papers. Volumes 16*. Harvard University Press, Cambridge, MA.

Pennacchiotti, M., De Cao, D., Basili, R., Croce, D., and Roth, M. (2008). Automatic induction of framenet lexical units. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 457–465. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532—1543.

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes*, 25(2&3):363—377.

Quillan, M. R. (1966). Semantic memory. Technical report, DTIC Document.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL*.

Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., and Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Rei, M. and Briscoe, T. (2014). Looking for hyponyms in vector space. In *Proceedings of CoNLL*, pages 68–77.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

Rimell, L. (2014). Distributional lexical entailment by topic coherence. In *Proceedings of EACL*, pages 511–519.

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

Roller, S. and Schulte im Walde, S. (2013). A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of EMNLP*, pages 1146–1157.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Rumelhart, D. E., McClelland, J. L., Group, P. R., et al. (1986). Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1-2. *Cambridge, MA*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Santus, E., Lenci, A., Lu, Q., and im Walde, S. S. (2014). Chasing hypernyms in vector spaces with entropy. In *Proceedings of EACL*, pages 38–42.

Schafer, C. and Yarowsky, D. (2002). Inducing translation lexicons via diverse similarity measures and bridge languages. In *CoNLL*, pages 1–7.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of WWW*, pages 1177–1178. ACM.

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3(3):417–57.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Shekarpour, S., Höffner, K., Lehmann, J., and Auer, S. (2013). Keyword query expansion on linked data using linguistic and semantic features. In *Proceedings of the 7th IEEE International Conference on Semantic Computing*, pages 191–197.

Shezaf, D. and Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of ACL*, pages 98–107.

Shutova, E., Van de Cruys, T., and Korhonen, A. (2012). Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of COLING 2012*, pages 1121–1130.

Silberer, C., Ferrari, V., and Lapata, M. (2013). Models of semantic representation with visual attributes. In *Proceedings of ACL*, pages 572–582.

Silberer, C. and Lapata, M. (2012). Grounded models of semantic representation. In *Proceedings of EMNLP*, pages 1423–1433.

Silberer, C. and Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, pages 721–732.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Simmons, W. K. and Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive neuropsychology*, 20(3-6):451–486.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, pages 1470–1477.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences*, 11(01):1–23.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of ACL*, 2:207–218.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Srivastava, N. and Salakhutdinov, R. (2014). Multimodal learning with deep Boltzmann machines. *Journal of Machine Learning Research*, 15(1):2949–2980.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185—190.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Tamura, A., Watanabe, T., and Sumita, E. (2012). Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of EMNLP*, pages 24–36.

Tellex, S., Katz, B., Lin, J., Fernandes, A., and Marton, G. (2003). Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 41–47. ACM.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Turney, P. D. and Mohammad, S. M. (2015). Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering*.

Turney, P. D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, pages 680–690.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: vector space models of semantics. *Journal of Artifical Intelligence Research*, 37(1):141–188.

Vallet, G., Brunel, L., and Versace, R. (2010). The perceptual nature of the cross-modal priming effect: arguments in favor of a sensory-based conception of memory. *Experimental Psychology*, 57(5):376—82.

Van Gelder, T. (1991). Classical questions, radical answers: Connectionism and the structure of mental representations. In *Connectionism and the Philosophy of Mind*, pages 355–381. Springer.

Vedaldi, A. and Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`.

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *CHI*, pages 319–326.

Vulić, I., Kiela, D., Clark, S., and Moens, M. (2016). Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of ACL*.

Vulić, I. and Moens, M. (2013a). Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *NAACL*, pages 106–116.

Vulić, I. and Moens, M.-F. (2013b). A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of EMNLP*, pages 1613–1624.

Vulić, I., Smet, W. D., and Moens, M. (2011). Identifying word translations from comparable corpora using latent topic models. In *Proceedings of ACL*, pages 479–484.

Weeds, J., Clarke, D., Reffin, J., Weir, D., and Keller, B. (2014). Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING*, pages 2249–2259.

Weeds, J., Weir, D., and McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1015. Association for Computational Linguistics.

Weyand, T., Kostrikov, I., and Philbin, J. (2016). Planet - photo geolocation with convolutional neural networks. *CoRR*, abs/1602.05314.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4):625–636.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328.

Yu, H., Siddharth, N., Barbu, A., and Siskind, J. M. (2015). A compositional framework for grounding language inference, generation, and acquisition in video. *Journal of Artificial Intelligence Research*, 52(1):601–713.

Zarzo, M. and Stanton, D. T. (2006). Identification of latent variables in a semantic odor profile database using principal component analysis. *Chemical Senses*, 31(8):713–724.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer.

APPENDIX A

MMFeat: A toolkit for multi-modal feature representations

The MMFeat toolkit is written in Python and is available online[1]. There are two command-line tools (described below) for obtaining files and extracting representations that do not require any knowledge of Python. The Python interface maintains a modular structure and contains the following modules:

- mmfeat.miner
- mmfeat.bow
- mmfeat.cnn
- mmfeat.space

Source files (images or sounds) can be obtained with the *miner* module, although this is not a requirement: it is straightforward to build an index of a data directory that matches words or phrases with relevant files. The miner module automatically generates this index, a Python dictionary mapping labels to lists of filenames, which is stored as a Python pickle file *index.pkl* in the data directory. The index is used by the *bow* and *cnn* modules, which together form the core of the package for obtaining perceptual representations. The *space* package allows for the manipulation and combination of multi-modal spaces.

miner Three data sources are currently supported: Google Images[2] (GoogleMiner), Bing Images[3] (BingMiner) and FreeSound[4] (FreeSoundMiner). All three of them require API keys, which can be obtained online and are stored in the *miner.yaml* settings file in the root folder.

[1] https://github.com/douwekiela/mmfeat
[2] https://images.google.com
[3] https://www.bing.com/images
[4] https://www.freesound.org

bow The bag of words methods are contained in this module. BoVW and BoAW are accessible through the mmfeat.bow.vw and mmfeat.bow.aw modules respectively, through the BoVW and BoAW classes. These classes obtain feature descriptors and perform clustering and quantization through a standard set of methods. BoVW uses dense SIFT for its local feature descriptors; BoAW uses MFCC. The modules also contain an interface for loading local feature descriptors from Matlab, allowing for simple integraton with e.g. VLFeat[5]. The centroids obtained by the clustering (sometimes also called the "codebook") are stored in the data directory for re-use at a later stage.

cnn The CNN module uses Python bindings to the Caffe deep learning framework (Jia et al., 2014). It supports the pre-trained reference adaptation of AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015) and VGGNet (Simonyan and Zisserman, 2015). The interface is identical to the *bow* interface.

space An additional module is provided for making it easy to manipulate perceptual representations. The module contains methods for aggregating image or sound file representations into visual or auditory representations; combining perceptual representations with textual representations into multi-modal ones; computing nearest neighbors and similarity scores; and calculating Spearman ρ_s correlation scores relative to human similarity and relatedness judgments.

A.1 Dependencies

MMFeat has the following dependencies: *scipy*, *scikit-learn* and *numpy*. These are standard Python libraries that are easy to install using your favorite package manager. The BoAW module additionally requires *librosa*[6] to obtain MFCC descriptors. The CNN module requires Caffe[7]. It is recommended to make use of Caffe's GPU support, if available, for increased processing speeds. More detailed installation instructions are provided in the readme file online and in the documentation of the respective projects.

A.2 Tools

MMFeat comes with two easy-to-use command-line tools for those unfamiliar with the Python programming language.

A.2.1 Mining: *miner.py*

The *miner.py* tool takes three arguments: the data source (bing, google or freesound), a query file that contains a line-by-line list of queries, and a data directory to store the mined image or sound files in. Its usage is as follows:

```
miner.py {bing,google,freesound} query_file data_dir [-n int]
```

[5] http://www.vlfeat.org
[6] https://github.com/bmcfee/librosa
[7] http://caffe.berkeleyvision.org

The -n option can be used to specify the number of images to download per query. The following examples show how to use the tool to get 10 images from Bing and 100 sound files from FreeSound for the queries "dog" and "cat":

```
$ echo -e "dog\ncat" > queries.txt
$ python miner.py -n 10 bing queries.txt ./img_data_dir
$ python miner.py -n 100 freesound queries.txt ./sound_data_dir
```

A.2.2 Feature extraction: *extract.py*

The *extract.py* tool takes three arguments: the type of model to apply (boaw, bovw or cnn), the data directory where relevant files and the index are stored, and the output file where the representations are written to. Its usage is as follows:

```
extract.py [-k int] [-c string] [-o {pickle,json,csv}] [-s float] \
   [-m {vgg,alexnet,googlenet}] {boaw,bovw,cnn} data_dir out_file
```

The -k option sets the number of clusters to use in the bag of words methods (the k in k-means). The -c option allows for pointing to an existing codebook, if available. The -s option allows for subsampling the number of files to use for the clustering process (which can require significant amounts of memory) and is in the range 0-1. The tool can output representation in Python pickle, JSON and CSV formats. The following examples show how the three models can easily be applied:

```
python extract.py -k 100 -s 0.1 bovw ./img_data_dir ./output_vecs.pkl
python extract.py -gpu -o json cnn ./img_data_dir ./output_vecs.json
python extract.py -k 300 -s 0.5 -o csv boaw ./snd_data_dir ./out.csv
```

A.3 Getting started

The command-line tools mirror the Python interface, which allows for more fine-grained control over the process. In what follows, we walk through an example illustrating the process. The code should be self-explanatory.

Mining The first step is to mine some images from Google Images:

```
datadir = '/path/to/data'
words = ['dog', 'cat']
n_images = 10

from mmfeat.miner import *

miner = GoogleMiner(datadir, '/path/to/miner.yaml')
miner.getResults(words, n_images)
miner.save()
```

Applying models We then apply both the BoVW and CNN models, in a manner familiar to scikit-learn users, by calling the fit() method:

```
from mmfeat.bow import *
from mmfeat.cnn import *

b = BoVW(k=100, subsample=0.1)
c = CNN(modelType='alexnet', gpu=True)
b.load(data_dir)
b.fit()
c.load(data_dir)
c.fit()
```

Building the space We subsequently construct the aggregated space of visual representations and print these to the screen:

```
from mmfeat.space import *

for lkp in [b.toLookup(), c.toLookup()]:
    vs = AggSpace(lkp, 'mean')
    print vs.space
```

These short examples are meant to show how one can straightforwardly obtain perceptual representations that can be applied in a wide variety of experiments.

A.4 Demos

To illustrate the range of possible applications, the toolkit comes with a set of demonstrations of its usage. The following demos are available:

1-Similarity and relatedness The demo downloads images for the concepts in the well-known MEN and SimLex-999 datasets, obtains CNN-derived visual representations and calculates the Spearman ρ_s correlations for textual, visual and multi-modal representations.

2-ESP game To illustrate that it is not necessary to mine images or sound files and that an existing data directory can be used, this demo builds an index for the ESP Game dataset and obtains and stores CNN representations for future use in other applications.

3-Matlab interface To show that local feature descriptors from Matlab can be used, this demo contains Matlab code (*run_dsift.m*) that uses VLFeat to obtain descriptors, which are then used in the BoVW model to obtain visual representations.

4-Instrument clustering The demo downloads sound files from FreeSound for a set of instruments and applies BoAW. The mean auditory representations are clustered and the cluster assignments are reported to the screen, showing similar instruments in similar clusters.

5-Image dispersion This demo obtains images for the concepts of *elephant* and *happiness* and applies BoVW. It then shows that the former has a lower image dispersion score and is consequently more concrete than the latter.