Doctoral School in Cognitive and Brain Sciences
XXVI Cycle

Doctor of Philosophy

Multimodal distributional semantics

Elia Bruni

Advisors: Prof. Marco Baroni

December 2013

Thesis Committee:

Marco Baroni, Thesis Advisor

Daniel Gatica-Perez

Sebastian Padó

Fabio Massimo Zanzotto

To Elisa, my love

Acknowledgements

The completion of my thesis and subsequent Ph.D. has been a wonderful adventure.

First of all, I am grateful to my advisor Marco Baroni for life. Working with you has been an illuminating experience. You have been a steady support throughout my Ph.D. career; you have always been patient but also motivating in times of difficulties. Above all, I have found a friend in you.

My sincere thanks to Google Inc. for having supported my Ph.D. with a Google Research Award assigned to Marco Baroni. Thanks to Massimiliano Ciaramita, for his unconditional support to my research.

A very special thank goes to Jasper Uijlings, whose computer vision understanding is tremendous and whose scientific work has inspired me. My work has greatly benefited from your suggestions.

Another special thank goes to Claudio Martella, who patiently explained me everything about programming.

Thank you to Roberto Zamparelli and Raffaella Bernardi for your support. Thanks to Leah Mercanti, the secretary that anyone would want to have!

I am also indebted to the students I had the pleasure to work with. You have been an invaluable help day in, day out, during all these years. Thanks to Giang Binh Tran, Nam Khan Tran, Ulisse Bordignon, Adam Liška and Irina Sergienya.

I'd also like to give a heartfelt, special thanks to Eva Vecchi, you have been of irreplaceable support during these three years (and I am still sorry I didn't chase away that big monster under your bed in Jeju)!

Thanks to all CLICers for the precious feedback, and for providing such a pleasant and productive working atmosphere. Thank you Moto for your delicious dinners! Thanks to Victoria Yanulevskaya, Gemma Boleda, Nicu Sebe, David Melcher, Francesca Bacci, Elisa Zamboni, Massimo Poesio and Andrew Anderson, it has been a pleasure to work with you.

A very big thank goes to Sebastian Petersson and Jacopo Amidei, who hosted me in Siena during the thesis write-up: It had been a really magic time!

Finally, I would like to thank my mom, dad, brother and sister for their support throughout everything; and of course Elisa, for her infinite patience and love.

Abstract

Although being one very simple statement, the distributional hypothesis - namely, words that occur in similar contexts are semantically similar - has been granted the role of main assumption in many computational linguistic techniques. This is mostly due to the fact that it allows to easily and automatically construct a representation of word meaning from a large textual input.

Among the computational linguistic techniques that are corpus-based and adopt the distributional hypothesis, Distributional semantic models (DSMs) have been shown to be a very effective method in many semantic-related tasks. DSMs approximate word meaning by vectors that keep track of the patterns of co-occurrence of words in the processed corpora. In addition, DSMs have been shown to be a very plausible computational model for human concept cognition, since they are able to simulate several psychological phenomena.

Despite their success, one of their strongest limitations is that they entirely represent word meaning in terms of connections with other words. Cognitive scientists have argued that, in this way, DSMs neglect that humans rely also on non-verbal experiences and have access to rich sources of perceptual knowledge when they learn the meaning of words.

In this work, the lack of perceptual grounding of distributional models is addressed by exploiting computer vision techniques that automatically identify discrete “visual words” in images, so that the distributional representation of a word can be extended to also encompass its co-occurrence with the visual words of images it is associated with.

A flexible architecture to integrate text- and image-based distributional information is introduced and tested on a set of empirical evaluations, showing that an integrated model is superior to a purely text-based approach, and it provides somewhat complementary semantic information with respect to the latter.

Contents

List of Figures

List of Tables

Publications

Parts of this thesis (ideas, figures, results, and discussions) have been presented previously in the following publications:

Elia Bruni, Nam Khan Tran and Marco Baroni. Multimodal distributional semantics. Submitted.

Andrew Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio and Marco Baroni. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In *Proceedings of EMNLP 2013 (Conference on Empirical Methods in Natural Language Processing)*, East Stroudsburg PA: ACL.

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

Elia Bruni, Jasper Uijlings, Marco Baroni and Nicu Sebe Distributional Semantics with Eyes: Using Image Analysis to Improve Computational Representations of Word Meaning In *Proceedings of the ACM Multimedia 2012*, New York NY: ACM, 1219-1228.

Elia Bruni, Gemma Boleda, Marco Baroni and Nam Khan Tran Distributional Semantics in Technicolor In *Proceedings of the ACL 2012 (50th Annual Meeting of the Association for Computational Linguistics)*, East Stroudsburg PA: ACL.

Chapter 1

Introduction

1.1 Semantic space models

Distributional semantics is the branch of computational linguistics that develops methods to approximate the meaning of words based on their distributional properties in large textual corpora. The basis of such methods relies on the **distributional hypothesis**: Words that occur in similar context are semantically similar. Although the distributional hypothesis has multiple theoretical underpinnings in psychology, linguistics, lexicography and philosophy of language [Firth, 1957; Harris, 1954; Miller and Charles, 1991; Wittgenstein, 1953], nowadays its strong influence is mainly due to its practical consequence: Harvesting meaning becomes the very straightforward operation of recording the contexts in which words occur and using their co-occurrence statistics to represent their meanings. Distributional semantic models (**DSMs**) are among the approaches which take full advantage of the distributional hypothesis by storing distributional information into vectors that can be utilized to compute the degree of semantic relatedness of two or more words in terms of geometric distance (see e.g., Clark [2013]; Turney and Pantel [2010]). For example, both *sea* and *ocean* might often appear with words such as *water*, *boat*, *fish* and *wave* and, as a result, their distributional vectors will be very close, indicating that the two words are very similar. The way in which DSMs operationalize the distributional hypothesis has led to very effective approaches in many semantic-related tasks (see Section 2.1

for some references), also helping confirming the validity of the hypothesis.

1.2 The symbol grounding problem

Despite its great success, distributional semantics has the clear limitation of reducing the acquisition of word meaning solely to the linguistic input, ignoring other important channels of information such as the perceptual one. A long tradition of studies which goes from philosophy to cognitive science has developed a strong objection against models which represent the meaning of symbols (e.g., words) in terms of other symbols (e.g., other words) and without any connection to the outside world, called the **symbol grounding problem** [Harnad, 1990]. DSMs have also to be considered defective with respect of the symbol grounding problem and have come under attack for their lack of grounding [Glenberg and Robertson, 2000].

Although the specific criticisms vented at them might not be entirely well-founded [Burgess, 2000], there can be little doubt that the limitation to textual contexts makes DSMs very dissimilar from humans, who, thanks to their senses, have access to rich sources of perceptual knowledge when learning the meaning of words – so much so that some cognitive scientists have argued that meaning is directly *embodied* in sensory-motor processing (for different views on embodiment in cognitive science de Vega et al. [2008]). Indeed, in the last decades a large amount of behavioural and neuroscientific evidence has been amassed indicating that our knowledge of words and concepts is inextricably linked with our perceptual and motor systems. For example, perceiving action-denoting verbs such as *kick* or *lick* involves the activation of areas of the brain controlling foot and tongue movements, respectively [Pulvermueller, 2005]. Hansen et al. [2006] asked subjects to adjust the color of fruit images objects until they appeared achromatic. The objects were generally adjusted until their color was shifted away from the subjects' gray point in a direction opposite to the typical color of the fruit, e.g., bananas were shifted towards blue because subjects' overcorrected for their typical yellow color. Typical color also influences lexical access: For example, subjects are faster at naming a pumpkin in a picture in which it is presented in orange than in a grayscale representation, slowest if it is in another color [Therriault et al., 2009]. As a final example, Kaschak et al. [2005] found that subjects are slower at processing a sentence describing an action if the sentence is presented concurrently to a visual stimulus depicting motion in the opposite direction of that described (e.g., *The car approached you* is harder to process concurrently to the perception of motion away from you). See Barsalou [2008] for a review of more evidence that conceptual and linguistic competence is strongly embodied.

One might argue that the concerns about DSMs not being grounded or embodied are exaggerated, because they overlook the fact that the patterns of linguistic co-occurrence exploited by DSMs reflect semantic knowledge we acquired through perception, so that linguistic and perceptual information are strongly correlated [Louwerse, 2011]. Because dogs are more often brown than pink, we are more likely to talk about *brown dogs* than *pink dogs*. Consequently, a child can learn useful facts about the meaning of the concept denoted by *dog* both by direct perception and through linguistic input (this explains, among other things, why congenitally blind subjects can have an excellent knowledge of color terms; see, e.g., Connolly et al. [2007]). One could then hypothesize that the meaning representations extracted from text corpora are indistinguishable from those derived from perception, making grounding redundant. However, there is by now a fairly extensive literature showing that this is not the case. Many studies [Andrews et al., 2009; Baroni and Lenci, 2008; Baroni et al., 2010; Riordan and Jones, 2011] have underlined how text-derived DSMs capture encyclopedic, functional and discourse-related properties of word meanings, but tend to miss their concrete aspects. Intuitively, we might harvest from text the information that *bananas* are *tropical* and *eatable*, but not that they are *yellow* (because few authors will write down obvious statements such as "*bananas are yellow*"). On the other hand, the same studies show how, when humans are asked to describe concepts, the features they produce (equivalent in a sense to the contextual features exploited by DSMs) are preponderantly of a perceptual nature: *Bananas* are *yellow*, *tigers* have *stripes*, and so on.[1]

[1] To be perfectly fair, this tendency might in part be triggered by the fact that, when subjects are asked to describe concepts, they might be encouraged to focus on their perceptual aspects by the experimenters' instructions. For example McRae et al. [2005] asked subjects to list first "physical properties, such as internal and external parts, and how [the object] looks."

This discrepancy between DSMs and humans is not, *per se*, a proof that DSMs will face empirical difficulties as computational semantic models. However, if we are interested in the potential implications of DSMs as models of how humans acquire and use language –as is the case for many DSM developers [Griffiths et al., 2007; Landauer and Dumais, 1997; Lund and Burgess, 1996]– then their complete lack of grounding in perception is a serious blow to their psychological plausibility, and exposes them to all the criticism that classic ungrounded symbolic models have received. Even at the empirical level, it is reasonable to expect that DSMs enriched with perceptual information would outperform their purely textual counterparts: Useful computational semantic models must capture human semantic knowledge, and human semantic knowledge is strongly informed by perception.

If we accept that grounding DSMs into perception is a desirable avenue of research, we must ask where we can find a practical source of perceptual information to embed into DSMs. Several interesting recent experiments use features produced by human subjects in concept description tasks (so-called "semantic norms") as a surrogate of true perceptual features [Andrews et al., 2009; Johns and Jones, 2012; Silberer and Lapata, 2012; Steyvers, 2010]. While this is a reasonable first step, and the integration methods proposed in these studies are quite sophisticated, using subject-produced features is unsatisfactory both practically and theoretically (see however for a crowdsourcing project that is addressing both kinds of concerns Kievit-Kylar and Jones [2011]). Practically, using subject-generated properties limits experiments to those words that denote concepts described in semantic norms, and even large norms contain features for just a few hundred concepts. Theoretically, the features produced by subjects in concept description tasks are far removed from the sort of implicit perceptual features they are supposed to stand for. For example, since they are expressed in words, they are limited to what can be conveyed verbally. Moreover, subjects tend to produce only salient and distinctive properties. They do not state that dogs have a head, since that's hardly a distinctive feature for an animal!

1.3 The proposed approach

The work presented in this thesis aims at filling the gap between the automatically constructed distributional semantic models and the human semantic memory, by building new DSMs that are perceptually grounded. In particular, we exploit recent advances in image analysis to extract compact representation of meaning from pictures, by extracting co-occurrence counts of target words and visual collocates from large datasets of tagged images. Thanks to these techniques, it is indeed possible to summarize an image by discretizing its content in vectors that keep track of visual unit counts. Moreover, we compose the obtained image-based DSMs with text-based DSMs and obtain multimodal representation of meaning.

1.4 Outline of the thesis

The main topic of this manuscript consists in the integration into DSMs of a more natural source of visual perceptual information (the relevant background from traditional and multimodal semantics is presented in Chapter 2). We exploit recently introduced image analysis techniques which allow us to encode the visual information in a way that is compatible with standard text-based distributional models of semantics. More in the detail, as visual perceptual source, we use collections of images naturally co-occurring with words (i.e., words appear as tags describing the image content). As feature extraction pipeline, we exploit recent advances in computer vision that can be broadly divided into two main steps. First, we use algorithms which encode the image contents in terms of low-level features. Low-level features are indeed ubiquitous in computer vision since are capable of automatically detecting and describing the most salient parts of an image. In the second step, we use the low level features extracted at step one to induce a more abstract model based on the well-established bags-of-visual-words method to represent images. The bag-of-visual-words method has the great advantage of discretizing the image content into a fixed-dimensionality feature vector and is a key transformation for our multimodal semantic representation.

In Chapter 3, Sections 3.1 and 3.2, the entire visual feature extraction pipeline is described.

Processing an image collection via the visual pipeline sketched above is just the first step to obtain a multimodal distributional model. Once the visual features are extracted and they act as a purely image-based representation of meaning, they have to be integrated in a multimodal space where textual and visual semantic features can cohabit. Chapter 4 is devoted exactly to this problem. The task of merging together two different channels of information can be pursued with increasingly sophisticated strategies. We will explore a first naive combination method that directly concatenates the visual and the textual vectors after a first normalization step (see Section 4.1). As an alternative, more advanced fusion strategy we propose a framework in which the textual and the visual features are projected into a common multidimensional space where they can interact, by promoting new connections between them (see Section 4.2). Moreover, in both Sections 4.1 and 4.2, each word in our framework is treated as requiring an equal amount of perceptual information, while it is natural to distinguish between very concrete, imageable words that require a fully perceptually informed feature representation, and abstract, non-imageable words, that are not “groundable" and therefore do not require perceptual features. Therefore, in Section 4.3, we explore different measures and ways to incorporate them in a new concatenation system, which is able to model textual and visual feature fusion locally, at a word-by-word level.

Chapter 5 and Chapter 6 of the thesis address the evaluation of the proposed multimodal framework. Since there is not a unique test which is capable of measuring if and how visual features convey meaningful information into a distributional semantic model, we approach the evaluation problem from different angles. The core part of the evaluation is presented in Chapter 5. Here we conduct three different tests, one of which is qualitative in nature and tries to asses the overall pattern of semantic relations that the model is able to capture (Section 5.1), while the other two are quantitative analyses, testing the model on word relatedness tasks (Section 5.2 and 5.3). In the qualitative test we can spot some significant differences between a traditional text-based semantic model and an image-based semantic model, while in both the quantitative tests adding visual features to state-of-the-art textual features systematically augments the performance. The framework evaluation continues in Chapter 6, where, after double-checking the validity of an enlarged set of multimodal models on word relatedness, we tackle two tasks where visual information is highly relevant, as they focus on color. In the first task we try to discover the color of of 52 concrete objects. In the second task we try to discriminate between literal and nonliteral uses of color terms. We show that especially here visual information has a determinant role which leads to the absolute best performance both with visual features standalone and combined with textual features.

Chapter 7 explores how information about object location can be used to advance multimodal distributional semantics. In particular, in Section 7.1 we exploit location information to improve visual feature extraction to tackle a word relatedness test. Interestingly, we show that a visual semantic model extracted only from within the precise location where the object appears in the image performs worse in the word relatedness task compared to a visual distributional model constructed with the information coming from the surrounding of the object only. In Section 7.2, we test whether image-based models capture the semantic patterns that emerge from fMRI recordings of the neural signal. Our result show that there is indeed a significant correlation between image-based and brain-based semantic similarities, and that image-based models complement text-based models so that the best correlation are obtained when the two modalities are combined in a multimodal distributional model. Chapter 8 contains our conclusive remarks and future work about multimodal distributional semantics.

Finally, Appendix 9 introduces an off-the-shelf freely distributed library to build an image-based semantic model.

Chapter 2

Background

2.1 Distributional semantics

In the last few decades, a number of different distributional semantic models (DSMs) of word meaning have been proposed in computational linguistics, all relying on the assumption that word meaning can be learned directly from the linguistic environment.

Semantic space models are one of the most common types of DSM. They approximate the meaning of words with vectors that record their distributional history in a **corpus** [Turney and Pantel, 2010]. A distributional semantic model is encoded in a matrix whose m rows are **semantic vectors** representing the meanings of a set of m **target words**. Each component of a semantic vector is a function of the occurrence counts of the corresponding target word in a certain context (see Lowe [2001], for a formal treatment). Definitions of **context** range from simple ones (such as documents or the occurrence of another word inside a fixed window from the target word) to more linguistically sophisticated ones (such as the occurrence of certain words connected to the target by special syntactic relations) [Padó and Lapata, 2007; Sahlgren, 2005; Turney and Pantel, 2010]. After the raw target-context counts are collected, they are transformed into **association scores** that typically discount the weights of components whose corresponding word-context pairs have a high probability of chance co-occurrence [Evert, 2005]. The rank of the matrix containing the semantic vectors as rows can optionally be decreased by **dimensionality reduction**, that might provide beneficial smoothing by getting rid of noise components and/or allow more efficient storage and computation [Landauer and Dumais, 1997; Sahlgren, 2005; Schütze, 1997]. Finally, the distributional semantic similarity of a pair of target words is estimated by a **similarity function** that takes their semantic vectors as input and returns a scalar similarity score as output.

There are many different semantic space models in the literature. Probably the best known is Latent Semantic Analysis (LSA, Landauer and Dumais [1997]), where a high dimensional semantic space for words is derived by the use of co-occurrence information between words and the passages where they occur. Another well-known example is the Hyperspace Analog to Language model (HAL, Lund and Burgess [1996]), where each word is represented by a vector containing weighted co-occurrence values of that word with the other words in a fixed window. Other semantic space models rely on syntactic relations instead of windows [Curran and Moens, 2002; Grefenstette, 1994; Padó and Lapata, 2007]. For general overviews of semantic space models see Clark [2013]; Erk [2012]; Manning and Schütze [1999]; Sahlgren [2006]; Turney and Pantel [2010].

More recently, probabilistic topic models have been receiving increasing attention as an alternative implementation of DSMs [Blei et al., 2003; Griffiths et al., 2007]. Probabilistic topic models also rely on co-occurrence information from large corpora to derive meaning but, differently from semantic space models, they are based on the assumption that words in a corpus exhibit some probabilistic structure connected to topics. Words are not represented as points in a high-dimensional space but as a probability distribution over a set of topics. Conversely, each topic can be defined as a probability distribution over different words. Probabilistic topic models solve the problem of meaning representation with a statistical inference: use the word corpus to infer the hidden topic structure.

Distributional semantic models, whether of the geometric or the probabilistic kind, ultimately are mainly used to provide a similarity score for arbitrary pairs of words, and that is how we will also employ them. Indeed, such models have shown to be very effective in modeling a wide range of semantic tasks including judgments of semantic relatedness and word categorization [Almuhareb, 2006; Baroni and Lenci, 2010; Budanitsky and Hirst, 2006; Radinsky et al., 2011; Reisinger and Mooney, 2010; Rothenhäusler and Schütze, 2009].

There are several data sets to assess how well a DSM captures human intuitions about semantic relatedness, such as the Rubenstein and Goodenough set [Rubenstein and Goodenough, 1965] and WordSim353 [Finkelstein et al., 2002]. Usually they are constructed by asking subjects to rate a set of word pairs according to a similarity scale. Then, the average rating for each pair is taken as an estimate of the perceived relatedness between the words (e.g., *dollar-buck*: 9.22, *cord-smile*: 0.31). To measure how well a distributional model approximates human semantic intuitions, usually a correlation measure between the similarity scores generated by the model and the human ratings is computed. The highest correlation we are aware of on the WordSim353 set we will also employ below is of 0.80 and it was obtained by a purely textual model called Temporal Semantic Analysis, which captures patterns of word usage over time and where concepts are represented as time series over a corpus of temporally-ordered documents [Radinsky et al., 2011]. This temporal knowledge could be integrated with the perceptual knowledge we encode in our model.

Humans are very good at grouping together words (or the concepts they denote) into classes based on their semantic relatedness [Murphy, 2002], therefore a cognitive-aware representation of meaning must show its proficiency also in categorization (e.g., Baroni et al. [2010]; Poesio and Almuhareb [2005]). Concept categorization is moreover useful for applications such as automated ontology construction and recognizing textual entailment. Unlike similarity ratings, categorization requires a discrete decision to group coordinates/cohyponyms into the same class and it is performed by applying standard clustering techniques to the model-generated vectors representing the words to be categorized. An example of a categorization data set is the Almuhareb-Poesio [Almuhareb and Poesio, 2005] data set, that we we also employ below, and which includes 402 concepts from WordNet (see Section 5.3.1 below), balanced in terms of frequency and degree of ambiguity. Rothenhäusler and Schütze [2009] present a text-based approach that constitutes the state of the art on the Almuhareb-Poesio data set (maximum clustering purity: 0.79).

See Baroni and Lenci [2010] for a survey of other semantic tasks that DSMs have been applied to, and Turney and Pantel [2010] for some of the applications in which DSMs are employed, including document classification, clustering and retrieval, question answering, automatic thesaurus generation, word sense disambiguation, query expansion, textual advertising.

2.2 Multimodal distributional semantics

The availability of large amounts of mixed media on the Web, on the one hand, and the discrete representation of images as visual words on the other has not escaped the attention of computational linguists interested in enriching distributional representations of word meaning with visual features.

Feng and Lapata [2010] propose the first multimodal distributional semantic model. Their generative probabilistic setting requires the extraction of textual and visual features from the same mixed-media corpus, because latent dimensions are here estimated through a probabilistic process which assumes that a document is generated by sampling both textual and visual words. Words are then represented by their distribution over a set of latent multimodal dimensions or "topics" [Griffiths et al., 2007] derived from the surface textual and visual features. Feng and Lapata experiment with a collection of documents downloaded from the BBC News website as corpus. They test their semantic representations on a subset of 254 pairs from the WordSim353 *Word Similarity* and *Word Association* test collections, obtaining gains in performance for both test sets when visual information is taken into account (correlations with human judgments of 0.12 and 0.32 respectively), compared to the textual modality standalone (0.08 and 0.25 respectively), even if performance is still well below state-of-the-art for WordSim353 (see Section 2.1 above).

The main drawbacks of this approach are that the textual and visual data must be extracted from the same corpus, thus limiting the choice of the corpora to be used, and that the generative probabilistic approach, while elegant, does not allow much flexibility in how the two information channels are combined. Below, we re-implement the Feng and Lapata method (MixLDA) training it on the ESP-Game data set, the same source of labeled images we adopt for our model. This is possible because the data set contains both images and the textual labels describing them. More in general, we recapture Feng and Lapata's idea of a common latent semantic space in the latent multimodal mixing step of our pipeline (see Section 5 below).

Leong and Mihalcea [2011] also exploit textual and visual information to obtain a multimodal distributional semantic model. While Feng and Lapata merge the two sources of information by learning a joint semantic model, Leong and Mihalcea propose a strategy akin to what we will call Scoring Level fusion below: Come up with separate text- and image-based similarity estimates, and combine them to obtain the multimodal score. In particular, they use two combination methods: summing the scores and computing their harmonic mean. Differently from Feng and Lapata [2010], here visual information for meaning representation is extracted not from a corpus but from a manually coded resource, namely the ImageNet[1] database [Deng et al., 2009], a large-scale ontology of images. Using a handcoded annotated visual resource such as ImageNet faces the same sort of problems that using a manually developed lexical database such as WordNet faces with respect to textual information, that is, applications will be severely limited by ImageNet coverage (for example, ImageNet is currently restricted to nominal concepts), and the interest of the model as a computational simulation of word meaning acquisition from naturally occurring language and visual data is somewhat reduced (humans do not learn the meaning of "mountain" from a set of carefully annotated images of mountains with little else crowding or occluding the scene). In the evaluation, Leong and Mihalcea experiment with small subsets of WordSim, obtaining some improvements, although not at the same level we report (the highest reported correlation is 0.59 on just 56 word pairs). Furthermore they use the same data set to tune and test their models.

In Bruni et al. [2011] we propose instead to directly concatenate the text- and image-based vectors to yield a single multimodal vector to represent words, as in what we call Feature Level fusion below. The text-based distributional vector representing a word, taken there from a state-of-the-art distributional semantic model [Baroni and Lenci, 2010], is concatenated with a vector representing the same word with visual features, extracted from all the images in the ESP Game collection we also use here. We obtain promising performance on WordSim and

[1] http://image-net.org/

other test sets, although appreciably lower than the results we report here (we obtained a maximum correlation of 0.52 when text- and image-based features are used together; compare to Table 5.2 below). Moreover, in Bruni et al. [2012a] we evaluate our multimodal models in the task of discovering the color of concrete objects, showing that the relation between words denoting concrete things and their typical color is better captured when visual information is also taken into account. Moreover, we show that multimodality helps in distinguishing literal and nonliteral uses of color terms.

Attempts to use multimodal models derived from text and images to perform more specific semantic tasks have also been reported. Bergsma and Goebel [2011] use textual and image-based cues to model selectional preferences of verbs (which nouns are likely arguments of verbs). Their experiment shows that in several cases visual information is more useful than text in this task. For example, by looking in textual corpora for words such as *carillon*, *migas* or *mamey*, not much useful information is obtained to guess which of the three is a plausible argument for the verb *to eat*. On the other hand, by exploiting Google image search functionality,[1] enough images for these words are found that a vision-based model of edible things can classify them correctly.

Silberer and Lapata [2012] present a first survey about grounded models of semantic representation. The authors conduct a comparative study of semantic models that incorporate linguistic and perceptual information. They experiment with a model that combines the two different channels in a concatenated multimodal space [Johns and Jones, 2012] and with two joint models, which construct the multimodal representation from a joint distribution of the two channels [Andrews et al., 2009] or from a joint “consensus" based on the correlation between the two channels [Hardoon et al., 2004]. They conclude that all models benefit form the integration of perceptual information since they obtain closer correspondence to human data, with a slightly better performance for the joint models. The novel comparative approach offers a nice and systematic overview of some fusion strategies in multimodal semantics. On the other hand, its main drawback is that the models under study cannot be really considered state-of-the-art multimodal representations. First, these models introduce perceptual information

[1] http://images.google.com/

with subject-produced features and not with more direct ways such as visual information via image analysis, as in this paper. Second, their performance in the semantic tasks is well below state-of-the-art in the literature.

More recently, Silberer and Lapata [2013] show that visual attribute classifiers can act as an effective substitute for feature norms to physically ground word meaning. Describing images by their attributes is a very recent idea initially explored in works such as Ferrari and Zisserman [2007] and then successfully applied to object recognition starting from Farhadi et al. [2009]. Learning to describe images through their attributes allows to generalize to objects never seen before and even to transcend the category level, while providing a more general description of the visual input. Silberer and Lapata successfully applied visual attributes in multimodal distributional semantics, by first creating their own dataset of images and visual attributes for the nouns contained in the McRae et al. [2005] norms. Images were downloaded from ImageNet, while the a attributes were manually annotated by the authors. They proceeded by training a classifier for each attribute in their list. Each concept in their visual semantic space is then represented by a vector summarizing the attribute prediction scores of each image tagged with the target concept. They tested the resulting visual model standalone and combined with a purely textual model. As combination methods, they used a simple concatenation method and a more sophisticated approach based on Canonical Correlation Analysis [Hardoon et al., 2004]. Their results demonstrate that, compared to a purely textual (topic) model, visual attributes improve the performance of distributional models across different settings of a semantic association task based on the Nelson norms [Nelson et al., 1998]. In particular, the performance was improved by visual attributes either standalone or combined with the textual model.

2.3 Other work on combining text and image data

Nowadays huge image collections are freely available on the Web, often incorporating additional textual data such as tags, which provide complementary information related to the image content. The multimedia and computer vision communities have fruitfully used textual tags of image data to supplement image analysis and to help bridging the semantic gap that visual features alone cannot easily fill. Taking inspiration from methods originally used in text processing, algorithms for image labeling, search and retrieval have been built upon the connection between text and visual features. Barnard et al. [2003] present one of the first attempts to model multimodal sets of images with associated text, learning the joint distribution of image regions and concrete concepts. Their model has been recently extended to attributes such as *yellow* or *striped* [Berg et al., 2010; Farhadi et al., 2009; Lampert et al., 2009; Wang and Forsyth, 2009], enabling transfer learning to recognize attributes without hand-labeled training data [Farhadi et al., 2009; Lampert et al., 2009] and even unseen object recognition by using a fast object localization system which integrates additional properties such as shape [Lampert et al., 2009]. Rohrbach et al. [2010] enhance transfer learning for attribute-based classification by using semantic relatedness values that they extract from textual knowledge bases.

The works reviewed above focus mostly on modeling the visual domain as opposed to the textual one, where, except for Berg et al. [2010], most use keywords rather than natural language captions. Both Farhadi et al. [2010] and Kulkarni et al. [2011] aim to associate more natural descriptions to images than just tags. They first use visual features to predict the content of an image in terms of objects and attributes. Then they use a natural language generation system to create image captions. Zha et al. [2009] present a system for visual query suggestion in image search. When users type a query, the system recommends them additional textual and visual queries that are semantically related with the original one in order to assist their searching process. Jamieson et al. [2010] propose an algorithm to simultaneously learn the names and the appearances of the objects represented in an unstructured collection of images containing a variety of objects within cluttered scenes.

Another interesting line of research exploits the connection between text and images with the goal to enhance human-robot interaction. For example, Chen and Mooney [2011] present an automatic system that understands natural-language navigation instructions by transforming them into an executable navigation plan.

For the task, a semantic parser for interpreting the navigation instructions is learnt by observing how human followers act. Matuszek et al. [2012] present instead a model for grounded attribute learning. Using joint textual and visual information, they build a system capable of producing a set of (visual) attribute models that help in an object selection task. Given a set of objects G and a sentence such as “Here are the yellow ones,", the system (i.e., the robot) has to select only the yellow objects from G.

Chapter 3

Extraction of visual and textual representations

In this Chapter we introduce the visual and the textual pipelines to construct image- and text-based distributional models respectively. Section 3.1 is entirely devoted to image processing. More in the detail, Section 3.1.1 introduces the means to obtain a first, low-level representation of the image content. Section 3.1.2 explains how to subsume such low-level information in a more abstract environment, where the visual units are the extensively used *visual words*. Section 3.2 describes the actual image sources and parameters we utilized for our experiments. Section 3.3 provides all the specifications for the state-of-the-art textual models we used.

3.1 Extraction of visual features from images

This section overviews the recent advances in the field of image analysis. In particular, it introduces state-of-the-art algorithms to automatically produce a discrete representation of the image content. This is obtained in two main steps, namely by first encoding the image content in terms of *local features* and then by inducing a more abstract representation, based on the *bag-of-visual-words* technique.

3.1.1 Extraction of local features from images

In the last years, there has been a tremendous progress on the development of local features for analyzing images in order to tackle common computer vision tasks such as object recognition and image retrieval [Grauman and Leibe, 2011]. The key aspect which makes local features so effective is their invariance to a series of image transformations such as translation, rotation, scaling and affine deformation. Invariance is indeed at the basis of recognition and representation approaches which render them robust to a variety of viewing conditions, occlusions and image clutter.

A typical local feature extraction pipeline is composed of two steps: (i) Feature detection and (ii) Feature description. More details about each of the two phases follow.

3.1.1.1 Feature detection

Feature detection is the first important stage for local feature extraction, in which a set of distinctive keypoints are localized in the image. This is a delicate process because the detection of the selected keypoints must be repeatable under varying image conditions, viewpoint changes and the presence of noise. More technically, the extraction of the keypoints should produce the same feature coordinates no matter if the image is rotated or translated. Of course, not all points do satisfy these restrictive requirements. The motion of *points lying on a uniform region* is indeed not detectable since they are indistinguishable from its neighbors, while the motion of *points lying on a straight line* can be traced only if it is perpendicular to the line. Therefore, we are naturally led to consider points that exhibit signal changes in (at least) two directions as the proper candidates. In most of the cases these points happen to be *corners*: Imagine to examine the change of intensity in an image due to shifts in a local window; around a corner, the image intensity will change greatly as a local window is shifted in arbitrary directions.

Two common algorithms which implement these criteria are the *Hessian detector* [Beaudet, 1978] and the *Harris detector* [Harris and Stephens, 1988][1]. The

[1] The list of feature detectors is not exhaustive. A much detailed discussion can be found in Mikolajczyk and Schmid [2005]

Hessian detector looks at strong changes in two orthogonal directions and it is based on the matrix of the second derivatives of the image points, so-called *Hessian*. The points detected by this algorithm are mainly located on corners and in intensively textured areas. The Harris detector looks even more strictly at corners only. It proceeds by "searching for points $\mathbf{x}$ where the second-moment matrix $\mathbf{C}$ around $\mathbf{x}$ has two large eigenvalues" [Grauman and Leibe, 2011]. In Figure 3.1 the Harris responses are compared to those of the Hessian detector, showing that the former is slightly more accurate in individuating corners, while Harris returns also regions with strong textures.

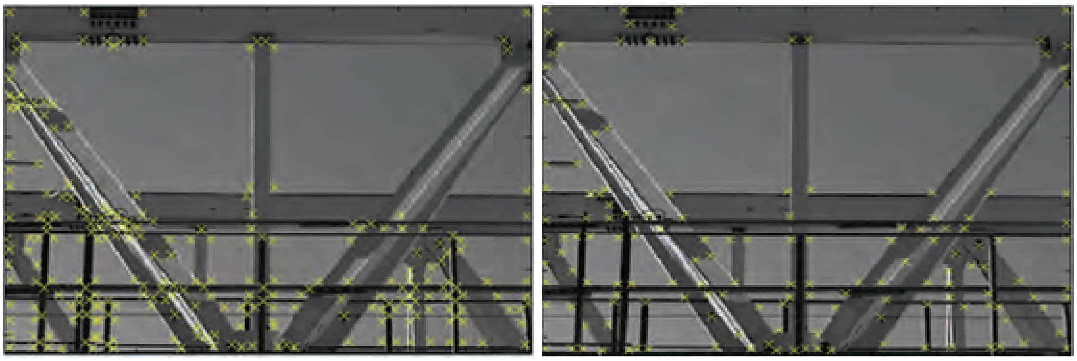


Figure 3.1: Example results of the (left) Hessian detector; (right) Harris detector. [Figure from Krystian Mikolajczyk]

While both the Hessian and the Harris detectors are exceptionally robust to image variations such as plane rotations, illumination changes and noise [Schmid et al., 2000], they still cannot offer locations which are robust enough to scale changes. If the location individuated by one of these detectors appear on a significantly larger scale within another image, the extracted structure will be different. This problem is due to the fact that both detectors utilize (Gaussian) derivatives computed on a single, fixed scale σ.

The solution consists in designing detection algorithms which become invariant to scale change by sampling the image at a range of scales (i.e., values of σ) automatically (see Figure 3.2).

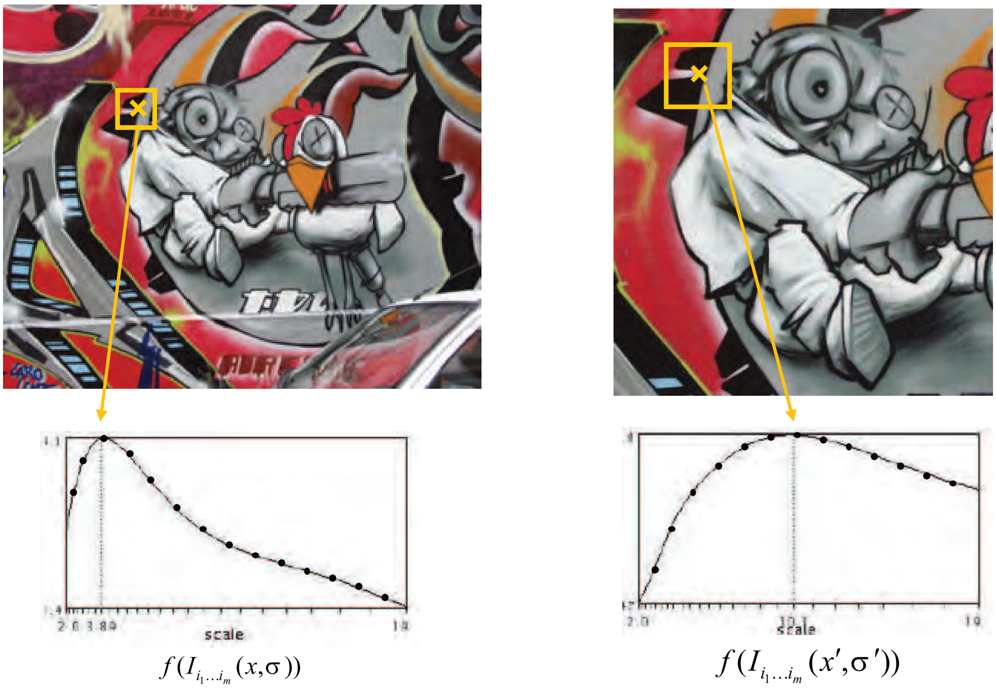


Figure 3.2: Automatic scale selection: Given a keypoint location, a scale-dependent signature function of the region around the keypoint is computed and the resulting value are plotted as a function of the scale. [Figure from Krystian Mikolajczyk]

One of the most utilized is the Laplacian-of-Gaussian **LoG**, which evaluates a scale-dependent signature function on the keypoint neighborhood and returns a value which is function of the scale [Lindeberg, 1998] (see Figure 3.3). One very popular version which approximates LoG is the Difference of Gaussians (DoG). DoG was introduced as the detection algorithm of the very popular feature descriptor Scale Invariant Feature Transform (**SIFT**), which we present below.

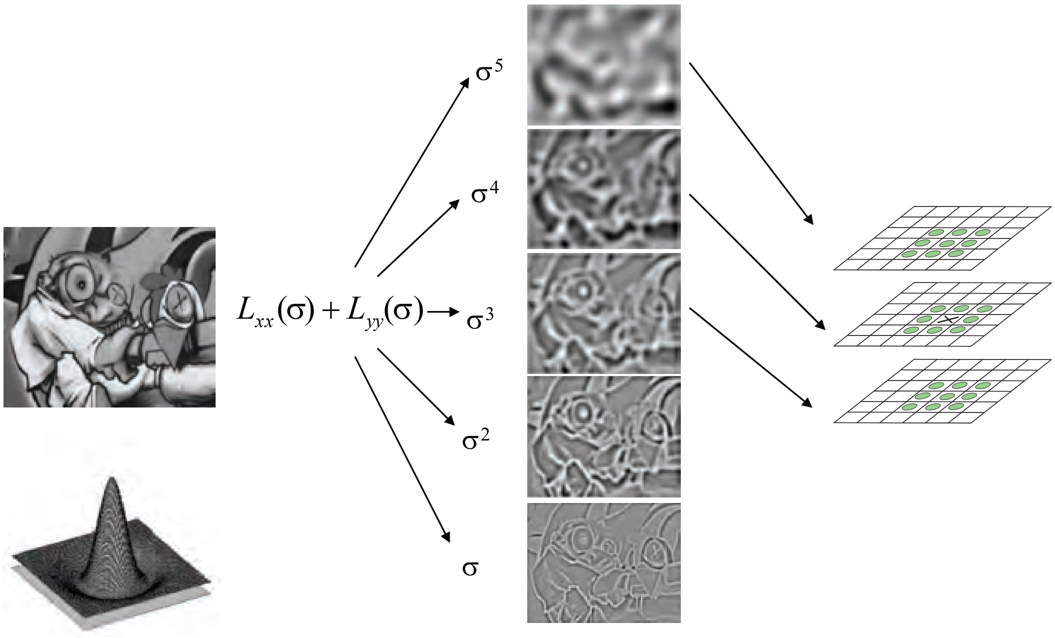


Figure 3.3: The Laplacian-of-Gaussian (LoG) detector searches for 3D scale space extrema of the LoG function. [Figure from Krystian Mikolajczyk]

An additional step after having detected a scale-invariant region is that of normalizing the content for *rotation invariance*. The typical way to do it is by finding the region's dominant direction and then by rotating the region content in accordance with this angle.

3.1.1.2 Feature description

Once a set of interesting regions is detected in an image by using one of the feature detectors introduced above, their content has to be described and encoded in a suitable feature vector. This is done by a feature descriptor. The most popular and effective descriptor is the Scale Invariant Feature Transform (**SIFT**), introduced by Lowe [1999, 2004]. As mentioned above, SIFT was originally introduced together with the DoG feature detector. Later, SIFT has been applied to other detectors such as Hessian, Harris and many more, achieving generally good performance as shown by Mikolajczyk and Schmid [2005]. More recently, it has also been applied to dense grids (dense SIFT), which has been shown to yield better performance in tasks such as object recognition. Note that the dense SIFT is the solution that we also adopt.

The success obtained by SIFT is due to its robustness to variation of the image conditions such as lighting and small position shifts of the detected keypoints. In order to achieve such robustness, SIFT encodes the image information in a localized set of *gradient orientation histograms*. The computation begins from one of the regions localized by one of the feature detectors (or by dense sampling). First of all, the image gradient magnitude and orientation is sampled around the keypoint location at a particular region scale. The sampling is computed on a 16×16 regular grid covering the region of interest. For each location, the gradient orientation is stored into a smaller 4×4 subgrid of gradient orientation histograms with 8 orientations bins each. Furthermore, each bin is weighted by the corresponding pixel's gradient magnitude. Once all orientation histograms have been computed, the resulting entries are concatenated to form a single 4×4×8=128 dimensional feature vector. Figure 3.4 depicts this procedure for a smaller 2×2 grid.

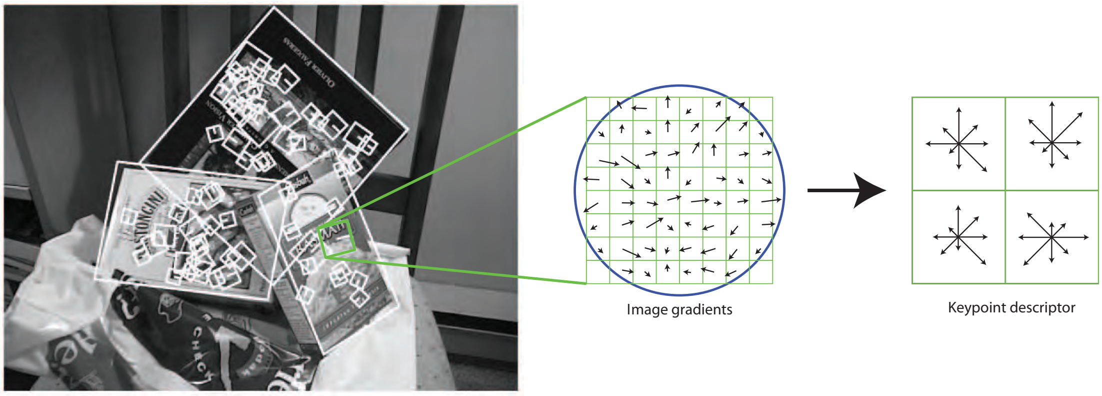


Figure 3.4: The SIFT descriptor. For each localized region, image gradients are computed on a regular grid and then encoded into a 4×4 grid of local gradient orientations (the figure shows only a 2×2 grid).

A last step of normalization to unit length is performed, in order to adjust for image contrast.

3.1.2 Bag of visual words

Ideally, to build a multimodal DSM, we would like to extract visual information from images in a way that is similar to how we do it for text. Thanks to a well-known image analysis technique, namely bag-of-visual-words (**BoVW**), it is indeed possible to discretize the image content and produce visual units somehow comparable to words in text, known as **visual words** [Bosch et al., 2007; Csurka et al., 2004; Nister and Stewenius, 2006; Sivic and Zisserman, 2003; Yang et al., 2007]. Therefore, semantic vectors can be extracted from a corpus of images associated with the target (textual) words using a similar pipeline to what is commonly used to construct text-based vectors: Collect co-occurrence counts of target words and discrete image-based contexts (visual words), and approximate the semantic relatedness of two words by a similarity function over the visual words representing them.

The BoVW technique to extract visual word representations of documents was inspired by the traditional bag-of-words (BoW) method in Information Retrieval. BoW in turn is a dictionary-based method to represent a (textual) document as a "bag" (i.e., order is not considered), which contains words from the dictionary. BoVW extends this idea to visual documents (namely images), describing them as a collection of discrete regions, capturing their appearance and ignoring their spatial structure (the visual equivalent of ignoring word order in text). A bag-of-visual-word representation of an image is convenient from an image-analysis point of view because it translates a usually large set of high-dimensional local features into a single sparse vector representation across images. Importantly, the size of the original set varies from image to image, while the bag-of-visual-word representation is of fixed dimensionality. Therefore, machine learning algorithms which by default expect fixed-dimensionality vectors as input (e.g., for supervised classification or unsupervised clustering) can be used to tackle typical image analysis tasks such as object recognition, image segmentation, video tracking, motion detection, etc.

More specifically, similarly to terms in a text document, an image has local interest points or **keypoints** defined as salient image patches that contain rich local information about the image. However “keypoint types” in images do not come off-the-shelf like word types in text documents. Local interest points have to be grouped into types (i.e. visual words) within and across images, so that an image can be represented by the number of occurrences of each type in it, analogously to BoW. The following pipeline is typically followed. From every image of a data set, local features are extracted and represented as vectors as described in 3.1.1. Feature vectors are then grouped across images into a number of clusters based on their similarity in descriptor space. Each cluster is treated as a discrete visual word. With its keypoints mapped onto visual words, each image can then be represented as a BoVW feature vector recording how many times each visual word occurs in it. In this way, we move from representing the image by a varying number of high-dimensional keypoint descriptor vectors to a representation in terms of a single visual word count vector of fixed dimensionality across all images, with the advantages we discussed above.

What kind of image content a visual word captures exactly depends on a number of factors, including the descriptors used to identify and represent keypoints, the clustering algorithm and the number of target visual words selected. In general, local interest points assigned to the same visual word tend to be patches with similar low-level appearance; but these local patterns need not be correlated with object-level parts present in the images [Grauman and Leibe, 2011]. Visual word assignment and its use to represent the image content is exemplified in Figure 1, where two images with a similar content are described in terms of bag-of-visual-word vectors.

3.2 Pipeline for visual representation

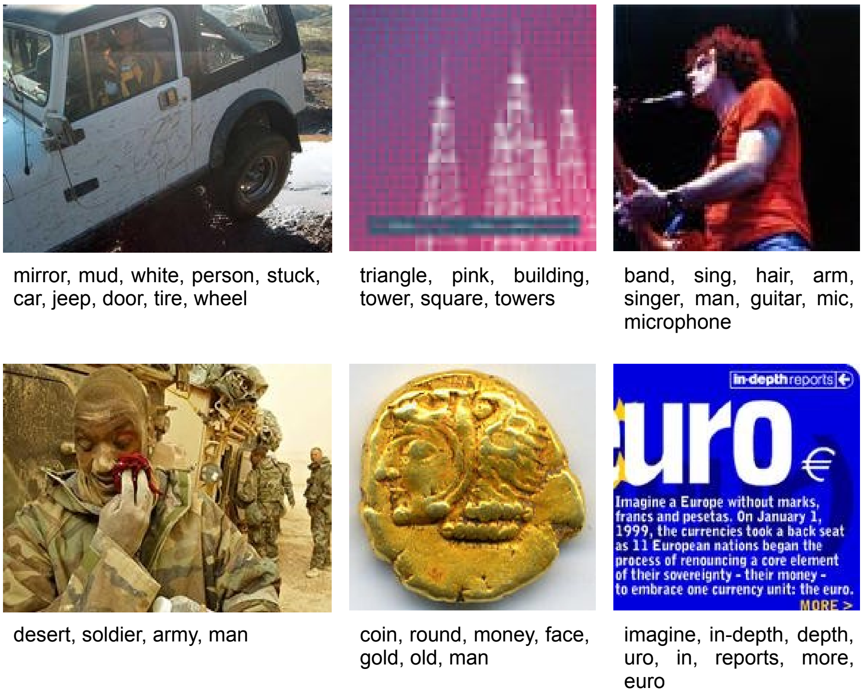
Given that image-based semantic vectors are a novelty with respect to text-based ones, in the next subsections we dedicate more space to how we constructed them, including full details about the source corpus we utilize as input of our pipeline (Section 3.2.1), the particular image analysis technique we choose to extract visual collocates and how we finally arrange them into semantic vectors that constitute the visual block of our distributional semantic matrix (Section 3.2.2).

3.2.1 Image source corpus

We adopt as our source corpus the ESP-Game data set[1] that contains 100K images, labeled through the famous "game with a purpose" developed by Louis von Ahn, in which two people partnered online must independently and rapidly agree on an appropriate word to label random selected images. Once a word is entered by both partners in a certain number of game rounds, that word is added as a tag for that image, and it becomes a taboo term for next rounds of the game involving the same image, to encourage players to produce more terms describing the image [Von Ahn, 2006]. The tags of images in the data set form a vocabulary of 20,515 distinct word types. Images have 14 tags on average (4.56 standard deviation), while a word is a tag for 70 images on average (737.71 standard deviation).

To have the words in the same format as in our text-based models, the tags are lemmatized and POS-tagged. To annotate the words with their parts of speech, we could not run a POS-tagger, since here words are out of context (i.e., each tag appears alphabetically within the small list of words labeling the same image and not within the ordinary sentence required by a POS-tagger). Thus we used a heuristic method, which assigned to the words in the ESP-Game vocabulary their most frequent tag in our textual corpora.

[1] http://www.cs.cmu.edu/~biglou/resources/


in-depth reports
uro €
Imagine a Europe without marks, francs and pesetas. On January 1, 1999, the currencies took a back seat as 11 European nations began the process of renouncing a core element of their sovereignty - their money - to embrace one currency unit: the euro.
MORE >

mirror, mud, white, person, stuck, car, jeep, door, tire, wheel

triangle, pink, building, tower, square, towers

band, sing, hair, arm, singer, man, guitar, mic, microphone

desert, soldier, army, man

coin, round, money, face, gold, old, man

imagine, in-depth, depth, uro, in, reports, more, euro

Figure 3.5: Samples of images and their tags from the ESP-Game data set

The ESP-Game corpus is an interesting data set from our point of view since, on the one hand, it is rather large and we know that the tags it contains are related to the images. On the other hand, it is not the product of experts labelling representative images, but of a noisy annotation process of often poor-quality or uninteresting images (e.g., logos) randomly downloaded from the Web. Thus, analogously to the characteristics of a textual corpus, our algorithms must be able to exploit large-scale statistical information, while being robust to noise. While cleaner and more illustrative examples of each concept are available in carefully constructed databases such as ImageNet (see Section 2.2), noisy tag annotations are available on a massive scale on sites such as Flickr[1] and Facebook,[2] so if we want to eventually exploit such data it is important that our methods can work on noisy input. A further advantage of ESP-Game with respect to ImageNet is that its images are associated not only with concrete noun categories but also

[1] http://www.flickr.com

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

with adjectives, verbs and nouns related to events (e.g., *vacation*, *party*, *travel*, etc). From a more practical point of view, "clean" data sets such as ImageNet are still relatively small, making experimentation with standard benchmarks difficult. In concrete, looking at the benchmarks we experiment with, as of mid 2013, ImageNet covers only just about half the pairs in the WordSim353 test set, and less than 40% of the Almuhareb-Poesio words. While in the future we want to explore to what extent higher-quality data sources can improve image-based models, this will require larger databases, or benchmarks relying on a very restricted vocabulary.

The image samples in Figure 3.5 exemplify different kinds of noise that characterize the ESP-Game data set. Both on top and bottom left and top right there are images where the scene is cluttered or partially occluded. The top center image is hardly a good representative of accompanying words such as *building*, *tower*(*s*) or *square*. Similarly, the center bottom image is only partially a good illustration of a coin, and certainly not a very good example of a man! Finally, the bottom right image is useless from a visual feature extraction perspective.

3.2.2 Image-based semantic vector construction

We collect co-occurrence counts of target words and image-based contexts by adopting the BoVW pipeline that, as we already explained in Section 3.1.2, is particularly convenient in order to discretize visual information into "visual collocates". We are adopting what is currently considered a standard implementation of BoVW. In the future, we could explore more cutting-edge ways to build image-based semantic vectors, such as local linear encoding [Wang et al., 2010] or Fisher encoding [Perronnin et al., 2010]. See Chatfield et al. [2011] for a systematic evaluation of several recent methods.

Our current implementation is composed of the following steps: (i) Extraction of the **local features**, which encode geometric or other information about the area around each keypoint, i.e., pixel of interest (here, SIFT features); (ii) **Encoding** the vector representation of an image by assigning the local descriptors to clusters corresponding to visual words, and recording their distribution across these clusters in the vector (this presupposes a preliminary step in which a clustering algorithm has been applied to the whole image collection or a sample, to determine the visual word vocabulary) (iii) Including some spatial information into the representation with **spatial binning**; (iv) Operate a series of image **transformations**, such as summing visual word occurrences across the list of images associated with a word label to obtain the co-occurrence counts associated with each word label and converting these counts into association scores, analogously to what is done in text analysis. The process (without spatial binning) is schematically illustrated in Figure 3.6, for a hypothetical example in which there are three images in the collection labeled with the word *monkey*. More details follow.

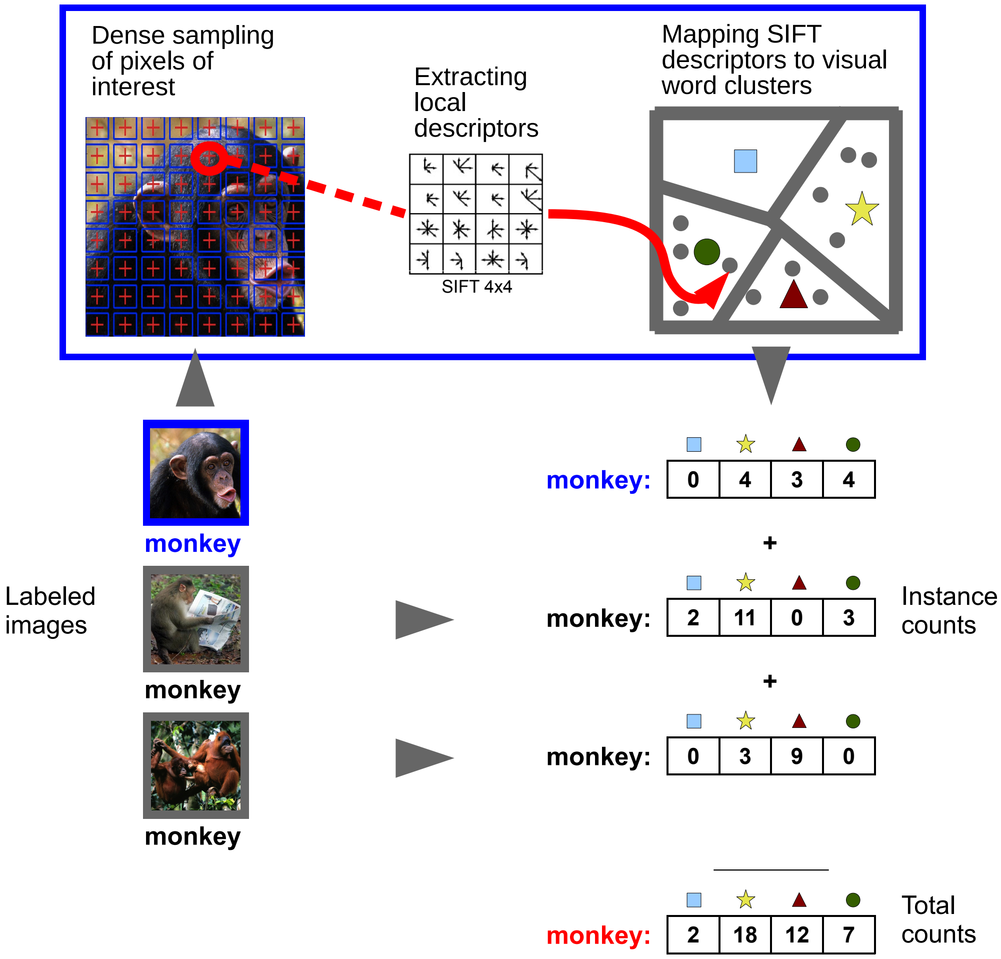


Figure 3.6: The procedure to build an image-based semantic vector for a target word. First, a bag-of-visual-word representation for each image labeled with the target word is computed (in this case, three images are labeled with the target word *monkey*). Then, the visual word occurrences across instance counts are summed to obtain the co-occurrence counts associated with the target word.

Local features To construct the local descriptors of pixels of interest we use Scale-Invariant Feature Transform (SIFT) [Lowe, 1999, 2004]. We chose SIFT for its invariance to image scale, orientation, noise, distortion and partial invariance to illumination changes. A SIFT vector is formed by measuring the local image gradients in the region around each location and orientation of the feature at multiple scales. In particular, the contents of 4 × 4 sampling subregions are explored around each keypoint. For each of the resulting 16 samples, the magnitude of the gradients at 8 orientations are calculated, which would already result in a SIFT feature vector of 128 components. However, we extract color SIFT descriptors in HSV (Hue, Saturation and Value) space [Bosch et al., 2008]. We use HSV because it encodes color information in a similar way to how humans do. We compute SIFT descriptors for each HSV component. This gives 3×128 dimensions per descriptor, 128 per channel. Color channels are then averaged to obtain the final 128-dimensional descriptors. We experimented also with different color scales, such as LUV, LAB and RGB, obtaining significantly worse performance compared to HSV on our development set introduced in 5.2.1, therefore we do not conduct further experiments with them. See van de Sande et al. [2010] for a systematic evaluation of color features.

Instead of searching for interesting keypoints with a salient patch detection algorithm, we use a more computationally intensive but also more thorough dense keypoint sampling approach, with patches of fixed size and localized on a regular grid covering the whole image and repeated over multiple scales. SIFT descriptors are computed on a regular grid every five pixels, at four scales (10, 15, 20, 25 pixel radii) and zeroing the low contrast descriptors. For their extraction we use the vl_phow command included in the VLFeat toolbox [Vedaldi and Fulkerson, 2010]. This implementation has been shown to be very close to Lowe's original but it is much faster for dense feature extraction. See Nowak et al. [2006] for a systematic evaluation of different patch sampling strategies.

Importantly, SIFT feature vectors are extracted from a large corpus of representative images to populate a feature space, which subsequently is quantized into a discrete number of visual words by clustering. Once this step is performed, every SIFT vector (local descriptor) from the original or new images can be translated into a visual word by determining which cluster it is nearest to in the quantized space.

Visual vocabulary To map SIFT descriptors to visual words, we first cluster all local descriptors extracted from all images in a training image corpus in their 3×128-dimensional space using the *k*-means clustering algorithm, and encode each descriptor by the index of the cluster (visual word) to which it belongs. *k*-

means is the most common way of constructing visual vocabularies [Grauman and Leibe, 2011]. Given a set $\mathbf{x_1}, ..., \mathbf{x}_n \in R^D$ of n training descriptors, k-means aims to partition the n descriptors into k sets ($k \leq n$) so as to minimize the cumulative approximation error $\sum_{\mathbf{i=1}}^{n} ||\mathbf{x}_i - \mu_{q_\mathbf{i}}||^\mathbf{2}$, with K centroids $\mu_\mathbf{1}, ..., \mu_K \in R^D$ and data-to-means assignments $\mathbf{q_1}, ..., \mathbf{q}_N \in \{1, ..., K\}$. We use an approximated version of k-means called Lloyd's algorithm [Lloyd, 1982] as implemented in the VLFeat toolbox.

To construct our visual vocabulary we extracted SIFT descriptors from all the 100K images of the ESP-Game data set. To tune the parameter k we used the MEN development set (see Section 5.2.1). By varying k between 500 and 5000 in steps of 500, we found the optimal k being 5000. It is most likely that the performance has not peaked even at 5000 visual words and enhancements could be attained by adopting larger visual vocabularies via more efficient implementations of the BoVW pipeline, as for example in Chatfield et al. [2011].

Encoding Given a set of descriptors $\mathbf{x_1}, ..., \mathbf{x}_n$ sampled from an image, let q_i be the assignment of each descriptor $\mathbf{x}_i$ to its corresponding visual word. The bag-of-visual-words representation of an image is a nonnegative vector $v \in R^k$ such that $v_k = |\{i : q_i = k\}|$, with q ranging from 1 to the number of visual words in the vocabulary (in our case, 5000). This representation is a vector of visual words obtained via hard quantization (i.e., assignment of each local descriptor vector to the single nearest codeword).

Spatial binning A consolidated way of introducing weak *geometry* in BoVW is the use of spatial histograms [Grauman and Darrell, 2005; Lazebnik et al., 2006]. The main idea is to divide the image in several (spatial) regions and to perform the entire visual word extraction and counting pipeline for each region and then concatenate the vectors. In our experiments the spatial regions are obtained by dividing the image in 4×4, for a total of 16 regions. Therefore, crossing the values for k with the spatial region, we increase the feature dimensions 16 times, for a total of 80,000 components in our vectors.

Transformations Once the BoVW representations are built, each target (textual) word is associated to the list of images which are labeled with it; the visual word occurrences across the list of images is summed to obtain the co-occurrence counts associated with the target (textual) word. In total, 20,515 target words (those that constitute ESP-Game tags) have an image-based semantic vector associated.

Also in the image-based semantic matrix, like in the text-based one, raw counts are transformed into nonnegative LMI. The difference is that here LMI is computed between a target element t that is a textual word and a context element c that is a visual word instead.

Note that, just like in the standard textual approach, we are accumulating visual words from all images that contain a word without taking into account the fact that words might denote concepts with multiple appearances, can be polysemous or even hide homonyms (our *bank* vector will include visual words extracted from river as well as building pictures). An interesting direction for further research would be to cluster the images associated to a word in order to distinguish the "visual senses" of the word, e.g., along the lines of what was done for textual models by Reisinger and Mooney [2010].

3.3 Pipeline for textual representation

As reviewed in Section 2.1 above, a text-based distributional model is encoded in a matrix whose rows are "semantic vectors" representing the meaning of a set of target words. Important parameters of the model are the choice of **target** and **contextual elements**, the **source corpora** used to extract co-occurrence information, the **context** delimiting the scope of co-occurrence, and the function to transform raw counts into statistical **association scores** downplaying the impact of very frequent elements.

Source corpora We collect co-occurrence counts from the concatenation of two corpora, ukWaC and Wackypedia (size: 1.9B and 820M running words, or tokens, respectively). ukWaC is a collection of Web pages based on a linguistically-controlled crawl of the `.uk` domain conducted in the mid 2000s. Wackypedia was built from a mid-2009 dump of the English Wikipedia. Both corpora have been automatically annotated with lemma (dictionary form) and part-of-speech (POS) category information using the TreeTagger,[1] they are freely and publicly available,[2] and they are widely used in linguistic research.

Target and context elements Since our source corpora are annotated with lemma and part-of-speech information, we take both into account when extracting target and context words (e.g., the string *sang* is treated as an instance of the verb lemma *sing*). We collect semantic vectors for a set of 30K target words (lemmas), namely the top 20K most frequent nouns, 5K most frequent adjectives and 5K most frequent verbs in the combined corpora. The same 30K lemmas are also employed as contextual elements (consequently, our text-based semantic models are encoded in a 30K×30K matrix). Note that when we combine the text matrices with the image-based ones, we preserve only those rows (target words) for which we also have an image-based vector, trimming the matrix to size 20,525×30K.

Context We define context in terms of words that co-occur within a window of fixed width, in the tradition of the popular HAL model [Lund and Burgess, 1996]. Window-based models are attractive for their simplicity and the fact that they do not require resource-intensive advanced linguistic annotation. They have moreover been reported to be at the state of the art in various semantic tasks [Rapp, 2003; Sahlgren, 2008], and in Bruni et al. [2012b] we show that the window-based methods we use here outperform both a document-as-context model and a sophisticated syntax- and lexical-pattern-based model on the MEN and WordSim test sets introduced in Section 5.2 below (see also the post-hoc analysis using the document-based model discussed at the end of Section 5.2.2 below). We consider two variants, **Window2** and **Window20** (we chose these particular variants arbitrarily, as representatives of narrow and wide windows, respectively). Window2 records sentence-internal co-occurrence with the nearest 2 content words to the left and right of each target word (function words such as articles and prepositions

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

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

are ignored). Window20 considers a larger window of 20 content words to the left and right of the target. A narrower window is expected to capture a narrower kind of semantic similarity, such as the one that exists between terms that are closely taxonomically related, for example coordinate concepts (*dog* and *cat*) or pairs of superordinate and subordinate concepts (*animal* and *dog*). The rationale behind this expectation is that terms will share many narrow-window collocates only if they are *very* similar, both semantically and syntactically. On the other hand, a broader window will capture a broader kind of "topical" similarity, such as one would expect of words that tend to occur in the same paragraphs (for example, *war* and *oil*, that are rather distant concepts in a taxonomic sense, but might easily occur in the same discourse). See Sahlgren [2006] for further discussion of the effects of context width on distributional semantic models.

Association score We transform raw co-occurrence counts into nonnegative Local Mutual Information (**LMI**) association scores. LMI scores are obtained by multiplying raw counts by Pointwise Mutual Information, and in the nonnegative case they are a close approximation to Log-Likelihood Ratio scores, that are one of the most widely used weighting schemes in computational linguistics [Evert, 2005]. The nonnegative LMI of target element t and context element c is defined as:

$$LMI(t,c) = \max\left(\text{Count}(t,c) \times \log \frac{\text{P}(t,c)}{\text{P}(t)\text{P}(c)}, 0\right)$$

It is worth observing that, in an extensive study of how parameters affect the quality of semantic vectors, Bullinaria and Levy [2007] and Bullinaria and Levy [2012] found that a model not unlike our Window2 (co-occurrence statistics from ukWaC, narrow window, lemmatized content word collocates, nonnegative pointwise mutual information instead of LMI) performs at or near the top in a variety of semantic tasks. Thus, we have independent grounds to claim that we are using a state-of-the-art text-based model.

Chapter 4

A framework for multimodal distributional semantics

In this Chapter, a general and flexible architecture for multimodal semantics is presented. The architecture makes use of distributional semantic models based on textual and visual information to build a multimodal representation of meaning. To merge the two sources, it uses a parametrizable pipeline which is able to capture both a straightforward concatenation method which has been used in most of the experiments presented below, and a more experimental fusion pipeline which constitutes itself an independent line of research and for which some preliminary results are also presented.

Section 4.1 describes a first naive method, in which the feature vectors are normalized and directly concatenated. Section 4.2 introduces a more advanced fusion scheme which encourages some further interaction of the visual and textual features while merging. Section 4.3 introduces an extension of the framework which takes into account the level of concreteness or imageability of the represented concepts, modeling fusion on a word-by-word level.

4.1 The unweighted concatenation approach

In the unweighted concatenation combination method combination we row-normalize, linearly weight and concatenate the text- and image-based vectors. Given a word that is present both in the text-based model and (as a tag) in the image-based model, we separately normalize the two vectors representing the word to length 1 (so that the text and image components will have equal weight), and we concatenate them to obtain the multimodal distributional semantic vector representing the word. The matrix of concatenated text- and image-based vectors is our multimodal distributional semantic model **UnweightedConcatFL** (the FL suffix is used to distinguish this model from another approach based on unweighted concatenation, but which concatenates the similarity scores instead of the feature vectors, see 4.2.3 for a detailed explanation).

4.2 Multimodal fusion

In a nutshell, the multimodal fusion approach consists in projecting concatenated text and image-based vectors onto a lower dimensionality latent space, in order to promote the formation of new connections within the components from each modality taking into account information and connections present in the other modality (see Caicedo et al. [2012] for similar ideas applied to image annotation and retrieval tasks).

The pipeline is based on two main steps:

(1) Latent Multimodal Mixing: The text and vision matrices are concatenated, obtaining a single matrix whose row vectors are projected onto a single, common space to make them interact.

(2) Multimodal Similarity Estimation: Information in the text- and image-based matrices is combined in two ways to obtain similarity estimates for pairs of target words: at the Feature Level and at the Scoring Level.

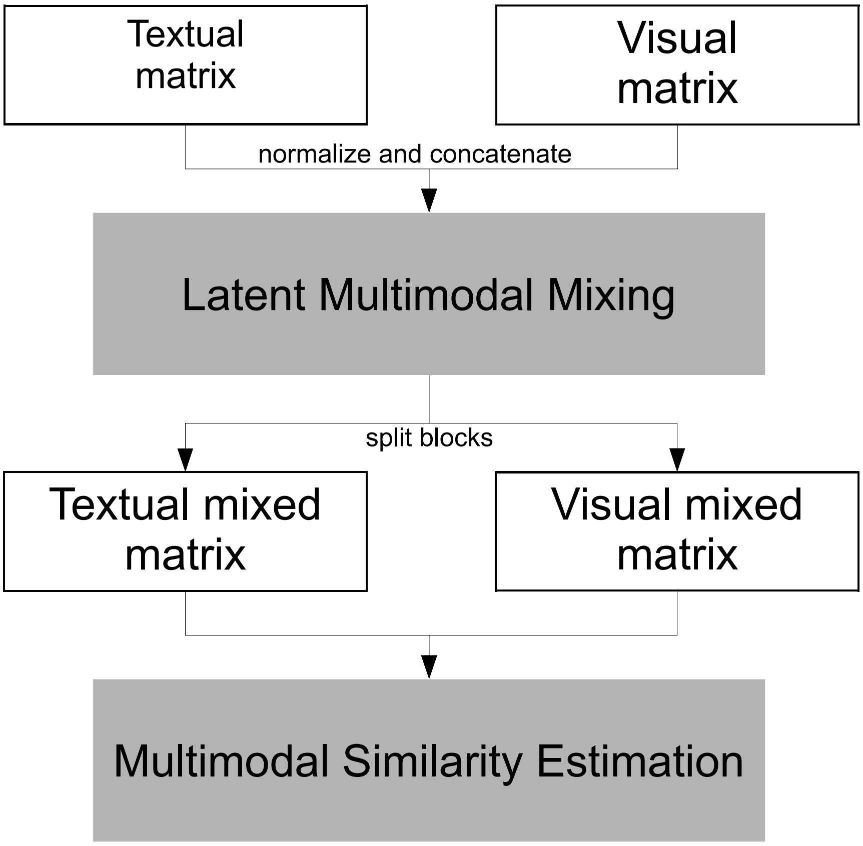


Figure 4.1: Multimodal fusion for combining textual and visual information in a semantic model.

Figure 4.1 describes the infrastructure we propose for fusion. First, we introduce a mixing phase to promote the interaction between modalities that we call Latent Multimodal Mixing. While this step is part of what other approaches would consider Feature Level fusion (see below), we keep it separated as it might benefit the Scoring Level fusion as well.

Once the mixing is performed, we proceed to integrate the textual and visual features. As reviewed in Section 2.2 above, in the literature fusion is performed at two main levels, the Feature Level and the Scoring Level. In the first case features are first combined and considered as a single input for operations, in the second case a task is performed separately with different sets of features and the separate results are then combined. Each approach has its own advantages and limitations and this is why both of them are incorporated into the multimodal infrastructure and together constitute what we call Multimodal Similarity Estimation. A Feature Level approach requires only one learning step (i.e., determining the parameters of the feature vector combination) and offers a richer vector-based representation of the combined information, that can also be used for other purposes (e.g., image and text features could be used together to train a classifier). Benefits of a Scoring Level approach include the possibility to have different representations (in principle, not even vectorial) and different similarity scores for different modalities and the ease of increasing (or decreasing) the number of different modalities used in the representation.

4.2.1 Latent multimodal mixing

This is a preparatory step in which the textual and the visual components are projected onto a common representation of lower dimensionality to discover correlated latent factors. The result is that new connections are made in each source matrix taking into account information and connections present in the other matrix, originating from patterns of co-variance that overlap. Importantly, we assume that mixing is done via a dimensionality reduction technique that has the following characteristics: a parameter k that determines the dimensionality of the reduced space and the fact that when k equals the rank of the original matrix the reduced matrix is identical or can be considered a good approximation of the original one. The commonly used Singular Value Decomposition reduction method that we adopt here for the mixing step satisfies these constraints.

As a toy example of why mixing might be beneficial, consider the concepts *pizza* and *coin*, that we could use as features in our text-based semantic vectors (i.e., record the co-occurrences of target words with these concepts as part of the vector dimensions). While these words are not likely to occur in similar contexts in text, they are obviously visually similar. So, the original text features *pizza* and *coin* might not be highly correlated. However, after mixing in multimodal space, they might both be associated with (have high weights on) the same reduced space component, if they both have similar distributions to visual features that cue roundness. Consequently, two textual features that were originally uncorrelated might be drawn closer to each other by multimodal mixing, if the corresponding concepts are visually similar, resulting in mixed textual features that are, in a sense, visually enriched, and *vice versa* for mixed visual features (interestingly, psychologists have shown that, under certain conditions, words such as *pizza* and *coin*, that are not strongly associated but perceptually similar, can prime each other; e.g., Pecher et al. [1998]).

Note that the matrices obtained by splitting the reduced-rank matrix back into the original textual and visual blocks have the same number of feature columns as the original textual and visual blocks, but the values in them have been smoothed by dimensionality reduction (we explain the details of how this is achieved in our specific implementation in the next paragraph). These matrices are then used to calculate a similarity score for a word pair by (re-)merging information at the feature and scoring levels.

4.2.2 Mixing with SVD

In our implementation, we perform mixing across text- and image-based features by applying the Singular Value Decomposition (**SVD**)[1] to the matrix obtained by concatenating the two feature types row-wise (so that each row of the concatenated matrix describes a target word in textual and visual space). SVD is a widely used technique to find the best approximation of the original data points in a space of lower underlying dimensionality whose basis vectors ("principal components" or "latent dimensions") are selected to capture as much of the variance in the original space as possible [Manning et al., 2008, Ch. 18]. By performing SVD on the concatenated textual and visual matrices, we project the two types of information into the same space, where they are described as linear combinations of principal components. Following the description in Pham et al. [2007], the SVD of a matrix M of rank r is a factorization of the form

$$M = U\Sigma V^t$$

[1] Computed with SVDLIBC: http://tedlab.mit.edu/~dr/SVDLIBC/

where

$$\begin{cases} U : \text{matrix of eigenvectors derived from } MM^t \\ \Sigma : r \times r \text{ diagonal matrix of singular values } \sigma \\ \sigma : \text{square roots of the eigenvalues of } MM^t \\ V^t : \text{matrix of eigenvectors derived from } M^tM \end{cases}$$

In our context, the matrix M is given by normalizing two feature matrices separately and then concatenating. By selecting the k largest values from matrix Σ and keeping the corresponding columns in matrices U and V, the reduced matrix M_k is given by

$$M_k = U_k \Sigma_k V_k^t$$

where $k < r$ is the dimensionality of the latent space. While M_k keeps the same number of columns/dimensions as M, its rank is now k. k is a free parameter that we tune on the development sets. Note that when k equals the rank of the original matrix, then trivially $M_k = M$. Thus we can consider not performing any SVD reduction as a special case of SVD, which helps when searching for the optimal parameters.

Note also that, if M has n columns, then V_k^t is a $k \times n$ matrix, so that M_k has the same number of columns of M. If the first j columns of M contain textual features, and columns from $j + 1$ to n contain visual features, the same will hold for M_k, although in the latter the values of the features will have been affected by global SVD smoothing. Thus, in the current implementation of the pipeline in Figure 4.1, block splitting is attained simply by dividing M_k into a textual mixed matrix containing its first j columns, and a visual mixed matrix containing the remaining columns.

4.2.3 General form and special cases

Given fixed and normalized text- and image-based matrices, our multimodal approach is parametrized by k (dimensionality of latent space), FL vs. SL, α (weight of text component in FL similarity estimation) and β (weight of text component in SL).

Note that when $k{=}r$, with r the rank of the original combined matrix, Latent Multimodal Mixing returns the original combined matrix (no actual mixing). Picking SL with $\beta{=}1$ or $\beta{=}0$ corresponds to using the textual or visual matrix only, respectively. We thus derive as special cases the models in which only text ($k{=}r$, SL, $\beta{=}1$) or only images ($k{=}r$, SL, $\beta{=}0$) are used (called **Text** and **Image** models in the Results section below). The linear approach presented in Section 4.1, in which the two matrices are concatenated without mixing, is the parametrization $k{=}r$, FL, $\alpha{=}0.5$ (called **UnweightedConcatFL** model, below). The summing approach of Leong and Mihalcea [2011] corresponds to $k{=}r$, SL, $\beta{=}0.5$ (**UnweightedConcatSL**, below). Picking $k{<}r$, SL, $\beta{=}1$ amounts to performing latent multimodal mixing, but then using textual features only; and the reverse with mixed image features only for $\beta{=}0$ (**Text**$_{mixed}$ and **Image**$_{mixed}$, respectively). Reducing these and other models to the same parametrized approach means that, given a development set for a specific task that requires similarity measurements, we can discover in a data-driven way which of the various models is best for the task at hand (for example, for a certain task we might discover that we are better off using text only, for another mixed text features, for yet another both text and image features, and so on).

Formally, given the set $k_1, ..., k_n \in R^1$ of n dimensionalities of the latent space (with k_n equal to the original dimensionality, and arbitrary steps between the chosen values), the sets $\alpha_1, ..., \alpha_m \in R^1$ of m potential weights of the text block in FL (with $\alpha_1 = 0$ and $\alpha_m = 1$) and $\beta_1, ..., \beta_l \in R^1$ of l weights of the text block in SL (with $\beta_1 = 0$ and $\beta_l = 1$), we can calculate the number of possible configurations to explore by $tot_c = n(m{+}l)$. Unless n, m and l are very large (i.e., we consider very small intervals between the values to be tested), it is completely feasible to perform a full search for the best parameters for a certain task without approximate optimization methods.

4.2.4 Multimodal similarity estimation

Similarity function Following the distributional hypothesis, DSMs describe a word in terms of the contexts in which it occurs. Therefore, to measure the similarity of two words DSMs need a function capable of determining the similarity of two such descriptions (i.e., of two semantic vectors). In the literature, there are many different similarity functions used to compare two semantic vectors, including cosine similarity, Euclidean distance, L_1 norm, Jaccard's coefficient, Jensen-Shannon divergence, Lin's similarity. For an extensive evaluation of different similarity measures, see Weeds [2003].

Here we focus on **cosine** similarity since it has been shown to be a very effective measure on many semantic benchmarks [Bullinaria and Levy, 2007; Padó and Lapata, 2007]. Also, given that our system is based on geometric principles, the cosine, together with Euclidean distance, is the most principled choice to measure similarity. For example, some of the measures listed above, having been developed from probabilistic considerations, will only be applicable to vectors that encode well-formed probability distributions, which is typically not the case (for example, after multimodal mixing, our vectors might contain negative values).

The cosine of two semantic vectors **a** and **b** is their dot product divided by the product of their lengths:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{i=n} a_i \times b_i}{\sqrt{\sum_{i=1}^{i=n} a_i^2} \times \sqrt{\sum_{i=1}^{i=n} b_i^2}}$$

The cosine ranges from 0 (orthogonal vectors) to $|1|$ (parallel vectors pointing in the same or opposite directions have cosine values of 1 and -1, respectively).

Feature Level fusion In Feature Level fusion (**FL**), we use the linear weighted fusion method to combine text- and image-based feature vectors of words into a single representation and then we use the latter to estimate the similarity of pairs. The linear weighted combination function is defined as

$$F = \alpha \times F_t \oplus (1 - \alpha) \times F_v$$

where $\oplus$ is the vector-concatenate operator.

Scoring Level fusion In Scoring Level fusion (**SL**), text- and image-based matrices are used to estimate similarity of pairs independently. The scores are then combined to obtain the final estimate by using a linear weighted scoring function:

$$S = \beta \times S_t + (1 - \beta) \times S_v$$

4.3 The concreteness factor to model fusion

One of the strongest limitations of the current version of our multimodal architecture as presented above, is the fact that every target word is assumed to be equally perceptually salient and consequently uniformly enriched with visual information. Intuitively, we might want to distinguish instead between concrete words, such as *chair* or *cat*, that require an integration of perceptual information for their representation, and abstract words, such as *consequence* or *absurd*, that can be represented on a purely symbolic basis.

This observation leads us to the idea of local weighting: Find a way to measure the concreteness/imageability of every word, and then use it during the multimodal fusion. In particular, we use the abstractness score automatically computed by the algorithm recently introduced by Turney et al. [2011], and the concreteness and imageability scores directly computed by us via the re-implementation of Turney's algorithm.[1] In both the original algorithm and our re-implementation, scores are calculated by computing the difference between the sum of text-based semantic similarities of a target word with a set of concrete paradigm words and the sum of its semantic similarities with a set of abstract paradigm words. In its original version, all words (i.e., both the paradigm words and the words for which an abstractness score is computed) are represented in a co-occurrence based matrix gathered from a large corpus of university websites. Co-occurrence counts are then transformed into Positive Pointwise Mutual Information scores [Church and Hanks, 1990] and the resulting matrix is smoothed with SVD. Pairwise semantic similarity is measured by cosines. In our re-implementation, we use in turn our co-occurrence matrix Window20 (see Section 3.3 for details about its construction). Both in the original algorithm and in our re-implementation, the paradigm words are selected with a supervised learning method trained on subject-rated words from the MRC Psycholinguis-

[1] It is worth noticing that having an automated way to retrieve these scores is crucial if we want to maintain the entire pipeline unsupervised.

tic Database Machine Usable Dictionary [Coltheart, 1981]. The MRC Database contains 150,837 words with 26 linguistic and psycholinguistic attributes. It is compiled from a number of different sources, and then normalized to have values from 100 to 700. Examples of the attributes are *age of acquisition*, *imagery*, *concreteness*, *familiarity*, and *ambiguity*. In our experiments, we considered the two attributes “concreteness" and “imagery". Words with a high concreteness score typically refer to objects, materials or persons, while words referring to abstract concepts that could not be experienced by the senses have a low concreteness score. Words with a high imagery score are those which tend to arouse images readily and vice versa for low scored words. In particular, from the concreteness attribute we obtain the concreteness weight, while the imagery attribute gives us the word’s imageability weight. In total, there are 8,228 words with concreteness scores and 9,240 words with imagery scores.

Examples of highly abstract words in the automatically rated list by the original Turney’s algorithm are *purvey*: 1.00, *sense*: 0.96 and *improbable*: 0.92, while examples of highly concrete words (i.e., words with a very low abstractness score) are *donut*: 0.00, *bullet*: 0.07 and *shoe*: 0.10.

Example of highly concrete words in the automatically rated list generated by our re-implementation of Turney’s algorithms are *weapon*: 0.72, *crocodile*: 0.96 and *pig*: 0.91, while examples of highly abstract words (i.e., words with a very low concreteness score) are *likelihood*: 0.17, *expectation*: 0.05 and *absolute*: 0.00. Examples of highly imageable words are *carnival*: 0.81, *toddler*: 0.81, *baby*: 0.86, while examples of words with very low imageability score are *await*: 0.32, *honesty*: 0.18 and *exist*: 0.00.

These data are given as input to the algorithm that provides abstractness, concreteness and imageability scores for the word list.

4.3.1 Local fusion scheme

Once *concreteness*, *abstractness* and *imageability* scores are computed as explained in the previous section, we proceed to incorporate the scores in a similarity measure. We utilize formula 4.1 for concreteness and imageability and formula 4.2 for abstractness.

$$\begin{aligned} sim(w_1, w_2) = imageSim(w_1, w_2) \times \frac{F(score(w_1), score(w_2))}{\beta} + \\ textSim(w_1, w_2) \times (1 - \frac{F(score(w_1), score(w_2))}{\beta}) \end{aligned} \tag{4.1}$$

$$\begin{aligned} sim(w_1, w_2) = imageSim(w_1, w_2) \times (1 - \frac{F(score(w_1), score(w_2))}{\beta}) + \\ textSim(w_1, w_2) \times \frac{F(score(w_1), score(w_2))}{\beta} \end{aligned} \tag{4.2}$$

Where w_1 and w_2 stand for the words being compared; *imageSim* measures the similarity score between visual feature vectors; *textSim* measures the similarity score between textual feature vectors; $score(w_i)$ stands for the property rating (concreteness, imageability or abstractness) of the word w_i; $F(x, y)$ is a function that combines the property ratings of two words - F can be either min, max or *mean*; β is a value between 1 and $+\infty$ that models the influence of the property rating - for the experiments we used a value of β in the interval [1, 2] with increasing steps of .01 (we used also the three extra values 10, 100, 1000).

The motivation behind the use of F is that we need a single property score which represents the level of abstractness/concreteness/imageability of the whole word pair. To obtain that, we apply an F function that, given as input the property scores of two words, outputs a combined property score. In addition, the parameter β weights the impact of the combined property score: In the case of concreteness and imageability, β measures the impact that the similarity between visual vectors has on the resulting score; in the case of abstractness, β decides the impact that the similarity between textual vectors has on the resulting score. This is based on the idea that for words with high concreteness/imageability scores, our multimodal system should rely more on the visual data than the textual; and vice-versa in the case of words with high abstractness scores.

Importantly, the reason why it is worth using both concreteness and abstractness in the experiments presented in the next Chapter is that the concreteness scores were computed directly by us in the re-implementation of Turney's algorithm, while for abstractness we directly rely on the original scores. Therefore, applying one or the other measure to our fusion function F results in different, non-complementary weighting scores.

Chapter 5

Evaluation of the framework

In this Chapter three evaluations of the framework are conducted. The first investigates how well the models reproduce different kinds of semantic relations. The second and third tasks look at the framework from a more quantitative angle, testing word relatedness and word clustering respectively.

Importantly, to compare our multimodal model to an external (still multimodal) approach, we re-implement Feng and Lapata's approach (discussed in Section 2.2) in a comparable setting to ours. Therefore, we treat the ESP-Game data set as a mixed-media corpus where each image together with the associated tags constitutes a document. For each image, we extract the image-based features with the procedure described above in 3.2.2 and use the words labeling that image to obtain the text-based features. These features are then stored in a term-by-document matrix, in which each image is treated as a document and a term can be either a textual tag or a visual word extracted from that image. We obtain a matrix of size 90K×100K, with 10K textual words (the word list resulting from the intersection of all the words used in our experimental data sets), 80K visual words and 100K documents (images). The Latent Dirichlet Allocation (MixLDA) model is trained on this matrix and tuned on the MEN development set by varying the number of topics K_t.[1] The optimal value we find is $K_t = 128$. Under MixLDA, each target word in an evaluation set is represented by the vector giving its distribution over the 128 latent topics.

[1] LDA was computed with Gensim: http://radimrehurek.com/gensim/

5.1 Differentation between semantic relations

To acquire a qualitative insight into how well our text- and image-based models are capturing word meaning, we test them on BLESS (Baroni-Lenci Evaluation of Semantic Similarity), a benchmark recently introduced by Baroni and Lenci [2011] to analyze specific aspects of lexico-semantic knowledge. Rather than focusing on a point estimate of quality of a model on a specific semantic task, BLESS allows us to assess the overall pattern of semantic relations that the model tends to capture. We run the BLESS evaluation before combining the textual and the visual channels together as a sanity check on the semantic meaningfulness of the image-based vectors, looking for potential complementary information with respect to text which can further motivate fusion. Note that since we are not combining the textual and visual sources, there are no tuning parameters to report.

5.1.1 Benchmark and method

BLESS contains a set of 200 **pivot** words denoting concrete concepts (we use 184 pivots, since for the remaining 16 we do not have a sufficiently large set of related words covered by our models). For each of the pivots, the data set contains a number of related words, or **relata**, instantiating the following 8 common **semantic relations** with the pivots: COORD: the relatum is a noun that is a co-hyponym (coordinate) of the pivot (*alligator-lizard*); HYPER: the relatum is a noun that is a hypernym (superordinate) of the pivot (*alligator-reptile*); MERO: the relatum is a noun referring to a meronym, that is, a part or material of the pivot (*alligator-teeth*); ATTRI: the relatum is an adjective expressing an attribute of the pivot (*alligator-ferocious*); EVENT: the relatum is a verb referring to an action or event involving the concept (*alligator-swim*); RAN.N, RAN.J and RAN.V, finally, are control cases where the pivot is matched to a set of random nouns (*alligator-trombone*), adjectives (*alligator-electronic*) and verbs (*alligator-conclude*), respectively.

For each pivot, BLESS contains a set of relata of each category (ranging from 7 hypernyms to 33 random nouns per pivot on average). In this way, BLESS can highlight the broader semantic properties of a model independently of its more specific preferences. For example, both a model that assigns a high score to *alligator-ferocious* and a model that assigns a high score to *alligator-green* will be correctly treated as models that have picked a relevant attribute of *alligators*. At the same time, the comparison of the specific relata selected by the models allows a more granular qualitative analysis of their differences.

Following the guidelines of Baroni and Lenci [2011], we analyze a semantic model as follows. We compute the cosine between the model vectors representing each of the 184 pivots and each of its relata, picking the relatum with the highest cosine for each of the 8 relations (the nearest hypernym, the nearest random noun, etc.). We then transform the 8 similarity scores collected in this way for each pivot onto standardized z scores (to get rid of pivot-specific effects), and produce a boxplot summarizing the distribution of scores per relation across the 184 pivots (for example, the leftmost box in the first panel of Figure 5.1 reports the distribution of 184 standardized cosines of nearest coordinate relata with the respective pivots). Besides analyzing the distributions qualitatively, we also discuss significant differences between the cosines of different relation types that were obtained via Tukey's Honestly Significance tests, thus correcting for multiple pairwise comparisons Abdi and Williams [2010].

5.1.2 Results

In Fig. 5.1, we report BLESS nearest relata distributions for the purely textual model Window20 (the Window2 distribution shows an even stronger skew in favour of coordinate neighbours) and the purely visual model we call Image in the next sections. The patterns produced by the text-based model (left panel) illustrate how a sensible word meaning profile should look like: coordinates are the most similar terms (an *alligator* is maximally similar to a *crocodile*), followed by superordinates (*reptile*) and parts (*teeth*). Semantically related adjectives (ATTRI: *ferocious*) and verbs (EVENT: *swim*) are less close to the pivots, but still more so than any random item.

The right panel shows the distribution of relata in the image-based semantic vectors. The overall pattern is quite similar to the one observed with the text-based vectors: there is a clear preference for coordinates, followed by hypernyms

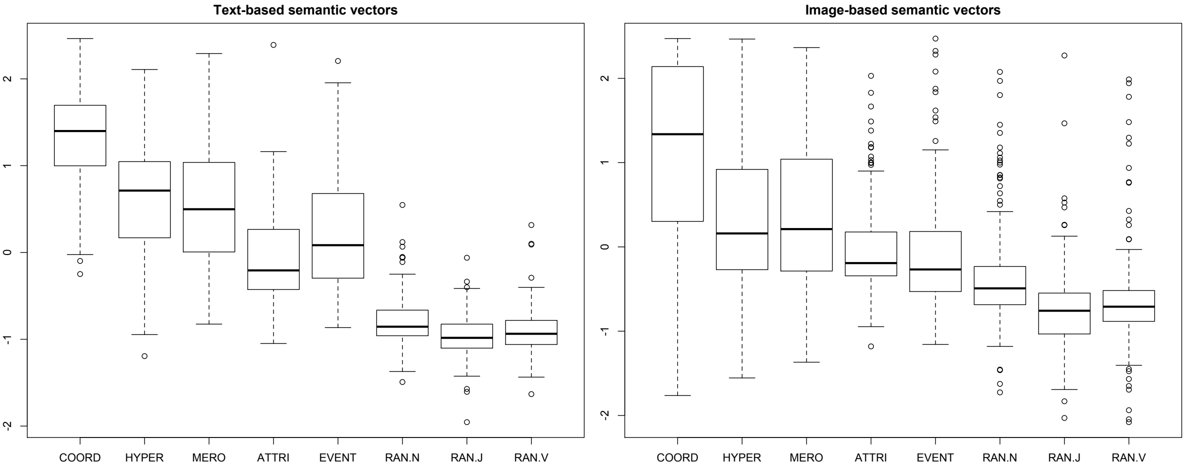


Figure 5.1: Distribution of z-normalized cosines of words instantiating various relations across BLESS pivots. Text-based vectors from the Window20 model.

and parts, then attributes and events, with all random relata further away from the pivots than the semantically meaningful categories. For both models, coordinates are significantly closer to the relata than hypernyms and meronyms, that are significantly closer than attributes and events, that are in turn significantly closer than any random category. Although the difference between hypernyms and parts is not significant with either representation, intriguingly the image-based vectors show a slight preference for the more imageable parts (*teeth*) than the more abstract hypernyms (*reptile*). The only difference of statistical import is the one between events and attributes, where the text-based model shows a significant preference for events, whereas the two categories are statistically indistinguishable in the image-based model (as we will see shortly, the relative preference of the latter for attributes is probably due to its tendency to pick perceptual adjectives denoting color and size).

Looking more closely at the specific relata picked by the text- and image-based models, the most striking differences pertain, again, to attributes. The text- and image-based models picked the same attribute for a pivot in just 20% of the cases (compare to 40% overlap across all non-random relation types). Table 5.1 reports the attributes picked by the text- vs. image-based models for 20 random cases where the two mismatch.

It is immediately clear from the table that, despite the fact that the pivots

pivot	*text*	*image*	*pivot*	*text*	*image*
cabbage	leafy	white	helicopter	heavy	old
carrot	fresh	orange	onion	fresh	white
cherry	ripe	red	oven	electric	new
deer	wild	brown	plum	juicy	red
dishwasher	electric	white	sofa	comfortable	old
elephant	wild	white	sparrow	wild	little
glider	heavy	white	stove	electric	hot
gorilla	wild	black	tanker	heavy	grey
hat	white	old	toaster	electric	new
hatchet	sharp	short	trout	fresh	old

Table 5.1: Attributes preferred by text- (Window20) vs. image-based models.

are nouns denoting concrete concepts, the text-based model almost never picks adjectives denoting salient perceptual properties (and in particular visual properties: just *white* for *hat* and *leafy* for *cabbage*). The text-based model focuses instead on encyclopedic properties such as *fresh, ripe, wild, electric* and *comfortable.* This is in line with earlier analyses of the "ungrounded" semantics provided by text-based models Andrews et al. [2009]; Baroni and Lenci [2008]; Baroni et al. [2010]; Riordan and Jones [2011], and differs greatly from the trend found in the image-based model. In 12/20 cases, the closest attribute for the latter model is a color. In the remaining cases, we have size (*short*, *little*), one instance of *hot* and, surprisingly, four of *old.*

To conclude, the analysis we presented confirms, on the one hand, our hypothesis that image-based distributional vectors contain sufficient information to capture a network of sensible word meaning relations. On the other, there are intriguing differences in the relations picked by the text- and image-based models, pointing to their complementarity.

5.2 Word relatedness

As is standard in the distributional semantics literature Budanitsky and Hirst [2006]; Sahlgren [2006], we assess the performance of our models on the task of predicting the degree of semantic relatedness between two words as rated by human judges. We test the models on the WS and MEN benchmarks.

5.2.1 Benchmarks and method

WS, that is, WordSim353[1] (see also Section 2.1) is a widely used benchmark constructed by asking 13 subjects to rate a set of 353 word pairs on an 11-point meaning similarity scale and averaging their ratings (e.g., *dollar/buck* gets a very high average rating, *professor/cucumber* a very low one). Our target words cover 252 WS pairs (thus, the correlations reported below are not directly comparable to those reported in other studies that used WS). However, our text-based models have much higher WS coverage (96%). When evaluated on the larger WS set they cover, Window2 and Window20 achieve 0.64 and 0.68 correlations, respectively. We are thus comparing the multimodal approach with purely textual models that are at the state of the art for WS (see results reported in Section 5.2.2 below).

The second benchmark we use, **MEN** (for Marco, Elia and Nam, the resource creators) was developed by us, specifically for the purpose of testing multimodal models. We created a large data set that, while comparable to WS and other benchmarks commonly used by the computational semantics community, contains only words that appear as image labels in the ESP-Game and MIRFLICKR-1M[2] collections, thus ensuring full coverage to researchers that train visual models from these resources. MEN consists of 3,000 word pairs with $[0, 1]$-normalized semantic relatedness ratings provided by Amazon Mechanical Turk workers (via the CrowdFlower[3] interface). For example, *beach/sand* has a MEN score of 0.96, *bakery/zebra* received a 0 score.

Compared to WS, MEN is sufficiently large to allow us to separate development and test data, avoiding issues of overfitting. We use indeed 2,000 MEN pairs (development set) for model tuning and 1,000 pairs for evaluation (test set). Importantly, the development set has been used to find the best configuration once for both the MEN test set and WS. Thus, the WS evaluation illustrates how well the parameters learned on training data from a specific data set generalize when applied to the same semantic task but on a different data set.

Models are evaluated as follows. For each pair in a data set, we compute the cosine of the model vectors representing the words in the pair, and then calculate

[1] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/
[2] http://press.liacs.nl/mirflickr/
[3] http://crowdflower.com/

the Spearman correlation of these cosines with the (pooled) human ratings of the same pairs, the idea being that the higher the correlation the better the model can simulate the relatedness scores.

MEN construction An earlier version of MEN has been used for the first time by the authors in Bruni et al. [2012a] but since the current article is the first major publication in which we focus specifically on it, and we have recently improved the benchmark by extending the ratings, we provide here further details on how it was constructed.

The word pairs that constitute MEN were randomly selected from words that occur at least 700 times in the concatenated ukWaC and Wackypedia text corpora and at least 50 times as tags in the ESP-Game and MIRFLICKR-1M tagged image collections. In order to avoid picking only pairs that were weakly related, as would happen if we were to sample random word pairs from a list, we ranked all possible pairs by their cosines according to our text-based model Window20. To gather the 3000 word pairs needed for the construction of MEN, we subsequently picked the first 1000 word pairs, another 1000 was sampled from pairs placed between 1001 and 3000 in the cosine-ranked list and the last block of 1000 pairs from the remaining items.

To acquire human semantic relatedness judgments, we decided to ask for comparative judgments on two pair exemplars at a time rather than absolute scores for single pairs, as was done by the creators of WS. This should constitute a more natural way to evaluate the target pairs, since human judgments are comparative in nature. When a person evaluates a given target, she does not do so in a vacuum, but in relation with a certain context. Moreover, binary choices were preferred because they make the construction of “right” and “wrong” control items straightforward (see Footnote 1). Operationally, each word pair was randomly matched with a comparison pair coming from the same set of 3000 items and rated by a single Turker as either more or less related than the comparison item. The validity of this approach is confirmed by the high annotation accuracy we observe in the control set,[1] and by the high correlation of the MEN scores

[1] The control items are correct annotations created prior to running the job on Amazon Mechanical Turk, which act as hidden tests that are randomly shown to Turkers as they complete

with ratings collected on a Likert scale we report below.

In the instructions, annotators were warned that sometimes both candidate pairs could contain words related in meaning and in such cases we asked them to pick the pair with the more strongly related words (e.g., both *wheels-car* and *dog-race* are somewhat related pairs, but the first one should be preferred as every car has wheels but not every dog is involved in a race). In other cases, annotators could find that neither pair contains closely related words, and in such cases they were instructed to pick the pair that contained slightly more related words (e.g., neither *auction-car* nor *cup-asphalt* are closely related words, but the first pair should be picked because fancy vintage cars are sold at auctions). We requested participants to be native speakers and only accepted those connecting from an English speaking country. We cannot guarantee that non-natives did not take part in the study, but our subject filtering techniques based on control pairs (see Footnote 1) ensures that only the data of speakers with a good command of English were retained.

To transform binary preference data to relatedness scores about the retrieved pairs, each of them was evaluated against 50 randomly picked comparison pairs, thus it received a score on a 50-point scale (given by the number of times out of 50 the pair was picked as the most related of the two). The score was subsequently normalized between 0 and 1 by dividing the number of times the pair was picked as the most related by 50. For example, *fun-night* was chosen as more related than the comparison pair 20 times, thus its normalized score is given by $20 \div 50 = 0.4$. Note that, in each comparison, we only recorded the preference assigned to one of the two pairs, to avoid dependencies between the final scores assigned to different pairs (that is, the times a pair was selected as a random comparison item for another pair were not counted as ratings of that pair).

the job. In this way, we can calculate the quality of a contributor's performance and reject their annotations if the accuracy drops below a certain percentage (we set a required minimum precision equal to 70%, but we obtained almost 100% average accuracy overall). Control items are also of great help to train quickly new workers to perform the required task. To create our control items we harvested two equally-sized sets of word pairs from WS, one containing only pairs with a high relatedness score, one containing only pairs with a low relatedness score. Each control item was then obtained by juxtaposing a high score pair with a low score pair and by treating the pair with the higher score as the one that should be selected by the annotators as the most related. All control items were manually checked. Examples of control items are *hotel-word* vs. *psychology-depression*, *telephone-communication* vs. *face-locomotive*.

Because raters saw the MEN pairs matched to different random items, with the number of pairs also varying from rater to rater, it is not possible to compute annotator agreement scores for MEN. However, to get a sense of human agreement, the first and third author rated all 3,000 pairs (presented in different random orders) on a standard 1-7 Likert scale. The Spearman correlation of the two authors is at 0.68, the correlation of their average ratings with the MEN scores is at 0.84. On the one hand, this high correlation suggests that MEN contains meaningful semantic ratings. On the other, it can also be taken as an upper bound on what computational models can realistically achieve when simulating the human MEN judgments.

The high-score MEN pairs include not only pairs of terms that are strictly taxonomically close (*cathedral-church*: 0.94) but also terms that are connected by broader semantic relations, such as whole-part (*flower-petal*: 0.92), item and related event (*boat-fishing*: 0.9), etc. For this reason, we prefer to refer to MEN as a semantic *relatedness* rather than *similarity* score data set. Note that WS is also capturing a broader notion of relatedness Agirre et al. [2009]. MEN is publicly available and it can be downloaded from: http://clic.cimec.unitn.it/~elia.bruni/MEN.

5.2.2 Results

Table 5.2 reports the correlations on the MEN testing and WS data sets when using either Window2 or Window20 as textual model. Our automated tuning method selected $k = 2^9$ (when textual information comes from Window2) and $k = 2^{10}$ (with Window20) as optimal, and Feature Level (FL) similarity estimation with $\alpha = 0.5$ in both cases (since the input matrices are row-normalized, the latter setting assigns equal weights to the textual and visual components). These are the models called TunedFL in the table. The Scoring Level (SL) strategy (again with similar weights assigned to the two channels, and same k values as TunedFL) performed only slightly worse than TunedFL, and we report the results for the best SL-based models as tuned on the development MEN data as well (TunedSL). In all other models reported in the table (UnweightedConcatFL, UnweightedConcatSL, MixLDA, Text$_{mixed}$ and Image$_{mixed}$), some parameters were

	Window2		Window20	
Model	*MEN*	*WS*	*MEN*	*WS*
Text	0.73	0.70	0.68	0.70
Image	0.43	0.36	0.43	0.36
UnweightedConcatFL	0.75	0.67	0.73	0.67
UnweightedConcatSL	0.76	0.69	0.74	0.64
MixLDA	0.30	0.23	0.30	0.23
Text$_{mixed}$	0.77	**0.73**	0.74	**0.75**
Image$_{mixed}$	0.55	0.52	0.57	0.51
TunedFL	**0.78**	0.72	0.76	**0.75**
TunedSL	**0.78**	0.71	**0.77**	0.72

Table 5.2: Spearman correlation of the models on MEN and WordSim (all coefficients significant with $p < 0.001$). TunedFL is the model automatically selected on the MEN development data.

tuned manually in order to gain insights on combination strategies representing ideas from the earlier literature.[1]

The first two rows of the table show results of the text- and image-based models, before any mixing. Text shows comparable performances on both data sets. Image correlates significantly better with MEN than WS but the correlations are lower than those of Text, in accordance with what was found in earlier studies. In the next three rows we find the results of the earlier multimodal approaches we took into consideration Bruni et al. [2011]; Feng and Lapata [2010]; Leong and Mihalcea [2011]. While the UnweightedConcatFL approach (analogous to Bruni et al.'s method), in which textual and visual matrices are concatenated without mixing, performs slightly better than Text on MEN, it attains lower performance on WS. Also UnweightedConcatSL (equivalent to Leong and Mihalcea's summing approach), where text and image sources are combined at the scoring level, obtains improvements only on MEN, loosing several correlation points on WS compared to Text.

Our implementation of MixLDA achieves very poor results both on MEN and WS. One might attribute this to the fact that Feng and Lapata's approach is

[1] For Text$_{mixed}$ and Image$_{mixed}$, the best k values were found on the development data. They were both set to 2^{10} with both textual sources.

	Window2	Window20
Text$_{mixed}$	0.47	0.49
TunedFL	0.46	0.49
TunedSL	0.46	0.47

Table 5.3: Pearson correlation of some of our best multimodal combinations on the WordSim subset covered by Feng and Lapata [2010] (all coefficients significant with $p < 0.001$; Pearson used instead of Spearman for full comparability with Feng and Lapata). The models assigned 0 similarity to the 71/253 pairs for which they were missing a vector. Feng and Lapata [2010] report 0.32 correlation for MixLDA.

constrained to using the same source for the textual and the visual model and our image data set is a poor source of textual data. Our approach is however also outperforming the original MixLDA by a large margin on the latter WS test set, where we are strongly disfavoured. In particular, Feng and Lapata [2010] report a correlation of 0.32 for the subset of 253 WS pairs covered by their model. We tested our system on the same subset, despite the fact that we are missing one or both vectors for 71 of the pairs (almost one third), so that our models are forced to assign 0 cosines to all these cases. Despite this huge handicap, our models are still attaining much higher correlations than the original MixLDA on the Feng and Lapata pairs, as illustrated for the most interesting fusion strategies in Table 5.3.

Analyzing now the effects of our fusion strategies, we can first see a uniform enhancement on both MEN and WS for Text$_{mixed}$ and Image$_{mixed}$ (the models obtained by first performing latent multimodal mixing on the combined matrix, but then using textual features only for Text$_{mixed}$ and visual features only for Image$_{mixed}$). Text$_{mixed}$ reaches the best performance overall on WS with both source textual models, and it is significantly better than Text on MEN according to a two-tailed paired permutation test Moore and McCabe [2005]. Looking then at the automatically selected TunedFL model, it reaches the best performance overall. Not only it significantly outperforms Text models on both data sets, but it is significantly better than Text$_{mixed}$ on MEN with Window20 (the difference is approaching significance with Window2 as well: $p = 0.06$). TunedSL is also very competitive. It is also significantly better than Text with both window sizes and Text$_{mixed}$ for Window20. It is noticeably worse than TunedFL on WS with Window20 only, and it is actually having a slight advantage on MEN with Window20 (the difference between TunedFL and TunedSL is never significant).

It is worth remarking that while Text$_{mixed}$ is a bit worse than the full fusion models, it still achieves high correlations with the human judgments and it has an extremely high correlation with the TunedFL best model ($\rho = 0.98$). This suggests that most of the benefits of multimodality are already captured by latent mixing. Text$_{mixed}$ is an attractive model because it has less parameters than the whole pipeline and it is more compact than TunedFL, since it discards the visual features after using them for mixing.

Validating the results While we have shown significant improvements when visual features are added to distributional models, one could object that improvements are due to the fact that we are using more information: a larger number of features (higher-dimensional vectors) for Feature Level fusion, and a more complex model (two similarity scores as independent variables to predict human judgments) for Scoring Level fusion. Further experiments provide evidence to respond to this objection.

First, we built purely textual models with the same number of features as our multimodal models – that is, instead of collecting co-occurrence of the target terms with the 30K most frequent content lemmas in our corpus (see Section 3.3 above), we extended the list of context items to the 110K most frequent content lemmas. The results with this larger textual models were virtually identical to those with 30K-dimensional vectors reported in Table 5.2 (correlation for the Window20 model on MEN was 0.69 instead of 0.68). Thus, at least when using our large corpus and a window-based approach, with 30K features we have pretty much exhausted the useful textual information, and it's the *nature*, not simply the quantity of the extra visual features we add that matters.

To answer the objection that the Scoring Level approach is using a more complex model, with two independent variables (text- and image-base similarities) instead of one, we casted the problem in standard inferential statistical terms (see e.g. Baayen [2008, Ch. 6]). Specifically, we fitted ordinary linear regression models to predict the MEN and WS ratings with only text-based similarities vs. text-

Text	TunedFL
dawn/dusk	pet/puppy
sunrise/sunset	candy/chocolate
canine/dog	paw/pet
grape/wine	bicycle/bike
foliage/plant	apple/cherry
foliage/petal	copper/metal
skyscraper/tall	military/soldier
cat/feline	paws/whiskers
pregnancy/pregnant	stream/waterfall
misty/rain	cheetah/lion

Table 5.4: Top 10 pairs whose relatedness is better captured by Text (Window20) vs. TunedFL.

and image-based similarities (for comparability with the Spearman correlation results reported above, the analyses were also replicated after transforming ratings and similarities into ranks). Both variables were highly significant in all experiments, and, more importantly, sequential F-tests over the nested models revealed that in all cases adding image-based similarities explains significantly more variance than what would be expected by chance given the extra parameter ($p < 0.01$).

Qualitative analysis To acquire qualitative insights into how multimodality is contributing to meaning representation, we first picked the top 200 most related pairs from the combined MEN and WS norms, so that we would be confident that they are indeed highly related pairs for humans, and then we looked, within this subset, at those pairs with the most pronounced difference in cosines between Text and TunedFL, using Window20 as our textual source. That is, the first column of Table 5.4 presents pairs that are considered very related by humans and where relatedness was better captured by Text, the second column pairs where relatedness was better captured by TunedFL.

Notice that 7/10 of the relations better captured by TunedFL are between coordinates or synonyms pertaining to concrete objects (*candy/chocolate*, *bicycle/bike*, *apple/cherry*, *military/soldier*, *paws/whiskers*, *stream/waterfall* and *cheetah/lion*), that should indeed be maximally visually similar (either the objects themselves or, in a case such as *paws/whiskers*, their surrounds). The purely text-based model, on the other hand, captures relations between times of the day, that, while imageable, are not well-delimited concrete objects (*dawn/dusk*, *sunrise/sunset*). It captures properties of concepts expressed by adjectives (*dog/canine*, *skyscraper/tall*, *cat/feline*, *pregnancy/pregnant*, *rain/misty*), and at least one case where spotting the relation requires encyclopedic knowledge (*grape/wine*). We thus hypothesize that the added value of the multimodally-enhanced model derives from the power of vision in finding relations between concrete objects at the same taxonomic level, that results in detecting particularly "tight" forms of relatedness, such as synonymy and coordination.

As observed by one reviewer, given the taxonomic nature of the information captured by the multimodal approach, it will be interesting to compare it in future work with features directly extracted from a linguistic taxonomy, such as WordNet. We observe in passing that such a manually-constructed resource, unlike those extracted from textual corpora, is likely to reflect both the linguistic and the perceptual knowledge of the lexicographers who built it.

Going in the opposite direction, another reviewer observed that we might get more mileage by combining visual features with textual models that are less taxonomic in nature. This hypothesis is partially confirmed by the fact that we obtain a larger relative improvement by mixing vision with Window20 than with Window2 (look back at Table 5.2, and see Section 3.3 above on why we think that the narrower window mainly captures taxonomic relations, the larger one broader topical themes). To further explore this conjecture, we re-ran the MEN and WS experiments combining the visual vectors with a document-based textual model (i.e., a semantic space whose dimensions record the number of occurrences of words in documents). Such a space is expected to capture mostly topical information, as it estimates relatedness on the basis of the tendency of words to occur in the same documents Sahlgren [2006]. The document-based model alone was not as good ad the window-based models, and combining it with image-based models led to relative improvements comparable or inferior to those attained with Window20. We conclude that, while looking for textual models that are more complementary with respect to visual information seems a reasonable direction to develop multimodal systems that cover a broader range of semantic phenomena, simply emphasizing the topical side of textual models evidently does not suffice.

5.3 Concept categorization

To verify if the conclusions reached on WS and MEN extend to different semantic tasks and, in particular, to assess whether our multimodal approach is able to capture and organize meaning as humans do, we use two existing **concept categorization** benchmarks that we call **Battig** and Almuhareb-Poesio (**AP**), respectively, where the goal is to cluster a set of (nominal) concepts into broader categories, as already discussed in Section 2.1.

In particular, we use Battig exclusively for tuning (in the same way we used the MEN development set in the previous section) and AP for testing. Only results on AP are reported. While in the word relatedness task the tuning and testing sets were quite similar (MEN development and MEN testing are two subsets of the same data set and the words in WS are similar to those in MEN), here the task is more challenging since Battig and AP are two independent data sets which were built following different strategies and populated with different kinds of concepts, namely very concrete and unambiguous concepts for Battig, vs. a mixture of concrete and abstract, possibly ambiguous concepts in AP. We adopted the present challenging training and testing regime because we felt that neither data set was of sufficient size to allow a split between development and testing data. More details follow.

5.3.1 Benchmarks and method

The Battig benchmark was introduced by Baroni et al. [2010] and it is based on the Battig and Montague norms of Van Overschelde et al. [2004]. It consists of 83 highly prototypical concepts from 10 common concrete categories (up to 10 concepts per class). Battig contains basic-level concepts belonging to categories such as *bird* (*eagle, owl...*), *kitchenware* (*bowl, spoon...*) or *vegetable* (*broccoli, potato...*). In the version we cover there are 77 concepts from 10 different classes.

AP was introduced by Almuhareb and Poesio [2005] and it is made of 402 nouns from 21 different WordNet classes. In the version we cover, AP contains 231 concepts to be clustered into 21 classes such as *vehicle* (*airplane, car...*), *time* (*aeon, future...*) or *social unit* (*brigade, nation*). The data set contains many difficult cases of unusual or ambiguous instances of a class, such as *casuarina* and *samba* as trees.

For both sets, following the original proponents and others, we cluster the words based on their pairwise cosines in the semantic space defined by a model using the CLUTO toolkit [Karypis, 2003]. We use CLUTO's built-in *repeated bisections with global optimization* method, accepting all of CLUTO's default values. Cluster quality is often evaluated by percentage purity [Zhao and Karypis, 2003]. If n_r^i is the number of items from the i-th true (gold standard) class that were assigned to the r-th cluster, n the total number of items, and k the number of clusters, then

$$purity = \frac{1}{\mathrm{n}} \sum_{i=1}^{i=n} \max\left(n_i^r\right)$$

In words, the number of items belonging to the majority true class (i.e., the most represented class in the cluster) are summed up across clusters and divided by the total number of items. In the best scenario purity will be 1 and it will approach 0 as cluster quality deteriorates.

Since we lack full AP coverage, the results we report below are not directly comparable with other studies that used it. However, our text-based models do have perfect coverage, and when evaluated on the full set achieve purities of 0.67 (Window2) and 0.61 (Window2), that are at state-of-the-art levels for comparable models, as reported in Section 2.1 above. So, again, we can confidently claim that the improvements achieved with multimodality are obtained by comparing our approach to competitive purely textual models.

5.3.2 Results

Table 5.5 reports percentage purities in the AP clustering task. Also here the best automatically selected model (TunedFL) uses FL similarity estimation as in the previous task, and has similar SVD k (2^7 for Window2 and 2^9 for Window20)

	Window2	Window20
Model	*AP*	*AP*
Text	0.73	0.65
Image	0.26	0.26
UnweightedConcatFL	0.74	0.64
UnweightedConcatSL	0.65	0.66
MixLDA	0.14	0.14
Text$_{mixed}$	0.74	0.67
Image$_{mixed}$	0.35	0.29
TunedFL	0.74	**0.69**
TunedSL	**0.75**	**0.69**

Table 5.5: Percentage purities of the models on AP. TunedFL is the model automatically selected on the Battig data

and α (0.5) parameters to the ones found for relatedness, suggesting that this particular parameter choice is robust and could be used out-of-the-box in other tasks as well. TunedSL is the best SL-based method on the tuning Battig set (same ks as TunedFL, $\alpha = 0.5$ for Window20 but $\alpha = 0.9$ on Window2).

Analogously to the previous semantic task, we see that the Image model alone is not at the level of the text models, although its AP purities are significantly above chance ($p < 0.05$ based on simulated distributions for random cluster assignment). Thus, we have a further confirmation of the fact that image-based vectors do capture important aspects of meaning. As in the previous task, MixLDA achieves very poor results.

Looking at the text-based models enhanced with visual information, we can see a general improvement in performance in almost all the multimodal combination strategies, except for UnweightedConcatFL with Window20 and UnweightedConcatSL with Window2. Even if Text$_{mixed}$ benefits from visual smoothing in both cases, it is again outperformed by TunedFL, whose performance is here very similar to that of TunedSL, that actually is slightly better on Window2. Interestingly, TunedSL outperforms Text on Window2 despite the fact this is the single combination strongly unbalanced towards textual similarity ($\alpha = 0.9$), indicating that visual information can be beneficial even when textual information accounts for the lion's share of the composed estimate.

Like in the relatedness task, adding an equal amount of further textual features instead of image-based ones does not help with Window20 (0.66 purity with 110K textual features) and even lowers performance with Window2 (0.69 purity). Thus, the improvement brought about by visual features must be attributed to their quality, not just quantity.

According to a two-tailed permutation test, even the largest difference between TunedFL and Text on Window20 is not significant. This might be due to the brittleness of the purity statistics leading to high variance in the permutations, and possibly to suboptimal tuning. Recall, in this respect, that the tuning phase was performed on a rather different data set (Battig) compared to the data set on which we eventually evaluated the models (AP). However, the overall trends are very encouraging, and in line with what we found in the relatedness study.

5.4 Evaluation of the local fusion scheme

We evaluate our scheme for local fusion on the two standard tasks of word relatedness and concept categorization. The choice was a natural consequence of the fact that we have already conducted an extensive evaluation of our original (global) framework on the same two tasks, granting a complete comparability of the old and the new schemes. We combined with the new scheme two sets of models introduced in Section 4.2.3: Text with Image (we call the resulting multimodal models **PlainLocal** models) and Text$_{mixed}$ with Image$_{mixed}$ (we call the resulting multimodal models **MixedLocal**).

The results are shown in Table 5.6 (word relatedness) and Table 5.7 (concept clustering).

	Window2			Window20		
Model	*Parameters*	*MEN*	*WS*	*Parameters*	*MEN*	*WS*
PlainLocal						
ABST	F=max, b=1.00	0.73	0.66	F=mean, b=1.03	0.71	0.65
IMAG	F=mean, b=1.31	0.75	0.69	F=max, b=1.03	0.73	0.67
CONC	F=max, b=1.83	0.75	0.69	F=max, b=1.35	0.72	0.67
MixedLocal						
ABST	F=max, b=1.00	0.75	0.67	F=mean, b=1.00	0.75	**0.71**
IMAG	F=min, b=2.00	**0.78**	0.71	F=max, b=1.23	**0.76**	0.67
CONC	F=max, b=10.00	**0.78**	**0.72**	F=max, b=1.76	**0.76**	0.67

Table 5.6: Spearman correlation of the new models on MEN and WordSim353 (all coefficients significant with $p < 0.001$).

	Window2		Window20	
Model	*Parameters*	*AP*	*Parameters*	*AP*
PlainLocal				
ABST	F=max, b=1.18	0.66	F=mean, b=1.00	0.63
IMAG	F=mean, b=1.01	**0.75**	F=min, b=10.00	0.65
CONC	F=min, b=1.23	0.70	F=min, b=10.00	**0.68**
MixedLocal				
ABST	F=max, b=1.00	0.69	F=min, b=1.00	0.60
IMAG	F=min, b=1.11	0.72	F=min, b=1.00	0.67
CONC	F=mean, b=1.70	0.73	F=min, b=1.00	**0.68**

Table 5.7: Purity values of the new models on AP.

In both tables, there are two blocks of models, the PlainLocal and the MixedLocal models. In each block each line corresponds to the usage of a particular property score: ABST for Turney's abstractness score, IMAG and CONC for the two measures extracted by us. On the second and third set of columns of the tables the experimental results for Window2 and Window20 are reported. Column "Parameters" shows the best learned values of the F function and β.

Word relatedness Comparing the results of the proposed models with the results of Section 5.2.2, we can find some enhancement. First of all, it is worth noticing that the best results in both cases are shown by the models that performed latent multimodal mixing ($Text_{mixed}$, TunedFusion and MixedLocal models). This observation supports the idea of the importance of SVD in improving the data representation. The best performance on the WordSim353 dataset was achieved with the $Text_{mixed}$ model and the MixedLocal-CONC model. On the MEN dataset the best performance was shown by the TunedFusion model and the MixedLocal-IMAG and the MixedLocal-CONC model both for Window2 and Window20. This shows that modeling the visual data with the concreteness or imageability score is reasonable, and can possibly improve the results.

Concept categorization The models that achieved the best purity value on the Almuhareb-Poesio dataset are the TunedFusion (from Section 5.3.2) and the PlainLocal-IMAG. Models with the new fusion scheme achieved results that are at least as good as the previous ones. Although there is a slight preference for latent multimodal mixing, the difference between PlainLocal and MixedLocal is not as clear as in the word relatedness task.

5.4.1 Analysis

In order to gain some qualitative insights about the differences between the globally and the locally fused multimodal models, we compare one representative model for each of the two combination techniques. In particular, for the word relatedness task, we pick $Text_{mixed}$ for global fusion and MixedLocal for local fusion, both with the textual model Window2. We proceed then to analyze the top 10 nearest neighbors from the concatenation of the MEN and the WS datasets of 3 concrete words (*map*, *toy*, and *ink*) and 3 abstract words (*practice*, *interest*, and *situation*).

Concrete-Global					
map		*toy*		*ink*	
map	1.0000	toy	1.0000	ink	1.0000
aerial	0.4666	doll	0.6647	pencil	0.7024
figure	0.4525	baby	0.5824	paper	0.3757
picture	0.4219	lego	0.3751	stencil	0.3658
illustration	0.4071	colorful	0.3432	white	0.3568
image	0.3967	cute	0.3276	dye	0.3433
photo	0.3458	lingerie	0.3179	handwriting	0.3287
sketch	0.3139	christmas	0.3164	doodle	0.3284
panorama	0.2901	play	0.2994	origami	0.2705
evidence	0.2740	child	0.2981	fountain	0.2671

Table 5.8: Top 10 closest words to the concrete words provided by the global weighting model.

Abstract-Global					
practice		*interest*		*situation*	
practice	1.0000	interest	1.0000	situation	1.0000
idea	0.7519	fertility	0.5178	kind	0.5197
example	0.5301	currency	0.4218	crisis	0.5120
news	0.4956	percent	0.3259	attitude	0.5104
performance	0.4761	shpere	0.2858	possibility	0.4624
reason	0.4742	activity	0.2350	reason	0.4250
guy	0.3773	focus	0.2249	pattern	0.4221
friend	0.3596	heart	0.2235	type	0.4087
cop	0.2578	recovery	0.2164	challenge	0.3810
chance	0.2429	development	0.2083	**issue**	0.3735

Table 5.9: Top 10 closest words to the abstract words provided by the global weighting model.

Concrete-Locall					
map		*toy*		*ink*	
map	1.0000	toy	1.0000	ink	1.0000
aerial	0.4574	doll	0.6375	pencil	0.6613
figure	0.4353	baby	0.5290	paper	0.3796
picture	0.4344	lego	0.3695	white	0.3659
image	0.3857	colorful	0.3314	fountain	0.3327
illustration	0.3752	cute	0.3314	dye	0.3235
photo	0.3575	play	0.3207	doodle	0.3204
sketch	0.3243	lingerie	0.3186	handwriting	0.3135
direction	0.2875	christmas	0.3171	stencil	0.2878
silhouette	0.2795	child	0.3043	origami	0.2479

Table 5.10: Top 10 closest words to the concrete words provided by the local weighting model.

Abstract-Locall					
practice		*interest*		*situation*	
practice	1.0000	interest	1.0000	situation	1.0000
idea	0.7201	fertility	0.5068	kind	0.5027
example	0.5020	currency	0.4100	crisis	0.4973
news	0.4663	percent	0.3131	attitude	0.4938
performance	0.4475	sphere	0.2847	possibility	0.4600
reason	0.4473	activity	0.2266	pattern	0.4096
guy	0.3598	heart	0.2245	reason	0.4048
friend	0.3421	focus	0.2211	type	0.3931
cop	0.2468	recovery	0.2126	challenge	0.3693
chance	0.2283	development	0.2084	**problem**	0.3636

Table 5.11: Top 10 closest words to the abstract words provided by the local weighting model.

The resulting lists of words for the global models are reported in Tables 5.8 and 5.9, and for the local models in Tables 5.10 and 5.11.

The first thing to notice is that there is a big overlap in the results of the examined models: 28 of 30 words from the concrete set results, and 29 of 30 words from the abstract set results, are the same (words that are different appear highlighted). For the examined lists of words, the similarity scores are almost the same for both models. The majority of the proposed words are indeed similar in meaning to the target words. Therefore, we have also a qualitative evidence that the new local fusion scheme doesn't bring a significant contribution to the multimodal framework.

As for the clustering experiments, one possible reason why no major differences can be found between the global and the local fusion schemes is probably given by the fact that both models are reaching the performance ceiling within the training set Battig (the purity values are indeed very high, between 0.93 and 0.97). Moreover, some error analysis we do not report here showed that the errors made by the systems are reasonable and explainable both in terms of benchmark coverage and, in some cases, ambiguity. This of course shouldn't prevent us to learn more sophisticated methods to train the local fusion parameters in future work.

5.5 Discussion

Still, by looking at the numerical results, we cannot deny that the improvement in performance attained when including visual information is not dramatic. Indeed, a pessimistic interpretation of the experiments could be that they confirm the hypothesis by Louwerse and others [Louwerse, 2011; Louwerse and Connell, 2011; Tillman et al., 2012] that perceptual information is already encoded, to a sufficient degree, into linguistic data, so direct visual features don't bring much to the table. However, we showed through various statistical and validation tests that our most important result, namely that adding visual information improves over using text alone, is robust and reliable. We think a more realistic take-home message is that the experiments we reported, while establishing the basic result we just mentioned, had some drawbacks we should overcome in further work.

First of all, we deliberately used general semantic benchmarks and state-of-the-art text models, so that the performance of computational methods might be getting close to the ceiling. At 0.78 correlation, our best models still have a few percentage points to go on MEN (estimated upper bound based on raters' agreement: 0.84, see Section 5.2.1), but the improvements are bound to be quite small. Concerning the AP benchmark, consider how difficult it would be even for humans to categorize *casuarina* and *samba* among the trees. Indeed, an error analysis of the TunedFL clustering results suggests that factors that might lead to better performance have little to do with vision. For example, the model "wrongly" clusters *branch* (a social unit according to AP) with the trees, and merges concepts such as *melon* and *peach* (fruit in AP) with *mandarin* and *lime* (trees). In lack of further contextual information, it's hard to dispute the model choices. Similarly, TunedFL splits the AP animal class into a cluster of small domestic mammals (*cats*, *dogs*, *kittens*, *mice*, *puppies* and *rats*) and a cluster containing everything else (mostly larger mammals such as *cows* and *elephants*). Again, the clustering procedure had no information about the classes we were searching for (e.g., animals in general, and not small animals), and so it is hard to see how performance could have improved thanks to better semantic features, visual or of other kinds. Moreover, all data sets include abstract terms, and are not specifically designed to test the more grounded aspects of meaning, where visual features might help most. We think it made sense to start our investigation with these general benchmarks of semantics, as opposed to *ad hoc* test sets, to show the viability of the multimodal approach.

Another factor to take into account is that both large-scale image data sets and the techniques to extract features from them are in their infancy, and we might be able to improve performance further by developing better image-based models. Regarding the data sets, we explained in Section 3.2.1 above why we chose ESP-Game, but obviously it is sub-optimal in many respects, as we also discuss there. Regarding the features, as we mentioned at the beginning of Section 3.2.2, recent advances in image processing, such as Fisher encoding, might lead to better ways to extract the information contained in images.

In the experiments, we also compared our automatically tuned multimodal model to other settings, showing its overall stability and superiority, with two important *caveats*. First, in both experiments good results are already obtained by using visual information to smooth text features, without using the visual features directly (what we called the Text$_{mixed}$ approach). Note that this is already a multimodal approach, in that visual information is crucially used to improve the quality of the textual dimensions, and indeed we've seen that it consistently outperforms using non-multimodally-smoothed text features. While Text$_{mixed}$ is *not* as good as our full tuned model, its simplicity makes it a very attractive approach.

Second, although automated tuning led us to prefer Feature Level over Scoring Level fusion on the development sets, TunedSL was clearly worse than TunedFL in just one case (with Window20 on WS), suggesting that, at least for the evaluation settings we considered, the difference between the two fusion strategies is not crucial. However, when comparing the "naive" versions of both strategies to the tuned ones across the results, it is clear that tuning is important to obtain consistently good performance, confirming the usefulness of our general fusion architecture.

To conclude, we observe that local fusion as implemented here doesn't seem to have a significant effect on fusion. An important direction for future work include finding new local weighting schemes.

Chapter 6

Exploring different visual spaces: The case of color

In this chapter we extend the evaluation of the proposed multimodal framework to two tasks which are based on color information and which therefore require models that are very effective in capturing visual information. The first is a color guessing task, in which a model needs to spot the correct colors of a list of concepts. The second task is slightly more sophisticated, and asks to distinguish between literal and nonliteral usages of color terms. Since for these experiments we use a larger number of visual semantic spaces, as a first sanity check, we assess the general quality of the spaces by repeating the semantic relatedness task. As in Section 5.2, we use WordSim and MEN as semantic relatedness datasets.

Our results show that the visual models are as good or better models of the meaning of words with visual correlates such as color terms, even in a nontrivial task that involves nonliteral uses of such words. Moreover, we show that visual and textual information are tapping on different aspects of meaning, such that they are complementary sources of information, and indeed combining them in multimodal models often improves performance. We also show that "hybrid" models exploiting the patterns of co-occurrence of words as tags of the same images can be a powerful surrogate of visual information under certain circumstances.

The rest of the chapter is structured as follows. Section 2 introduces the textual, visual, multimodal, and hybrid models we use for our experiments. We present our experiments in sections 3 to 5. Section 6 reviews related work, and section 7 finishes with conclusions and future work.

6.1 Distributional semantic models

6.1.1 Textual models

For the current project, we constructed a set of textual distributional models that implement various standard ways to extract them from a corpus, chosen to be representative of the state of the art. In all cases, occurrence and co-occurrence statistics are extracted from the freely available ukWaC and Wackypedia corpora combined.

In addition to the two window-based models **Window2** and **Window20** already introduced in Section 3.3, for these particular experiments we consider also a "topic-based" approach that we call **Document**, in which words are represented as distributions over documents. It is based on a word-by-document matrix, recording the distribution of the 30K target words across the 30K documents in the concatenated corpus that have the largest cumulative LMI mass. This model is thus akin to traditional Latent Semantic Analysis [Landauer and Dumais, 1997], without dimensionality reduction.

We add to the models we constructed also the freely available Distributional Memory (**DM**) model,[1] that has been shown to reach state-of-the-art performance in many semantic tasks [Baroni and Lenci, 2010]. DM is an example of a more complex text-based model that exploits lexico-syntactic and dependency relations between words (see Baroni and Lenci's article for details), and we use it as an instance of a grammar-based model. DM is based on the same corpora we used plus the 100M-word British National Corpus,[2] and it also uses LMI scores.

[1] http://clic.cimec.unitn.it/dm

[2] http://www.natcorp.ox.ac.uk/

6.1.2 Visual models

We extract descriptor features of two types. First, the standard SIFT feature vectors, good at characterizing parts of objects. For these experiments, we varied the number k of visual words between 500 and 2,500 in steps of 500 (information about the effect of this parameter in the next section). Second, **LAB** features [Fairchild, 2005], which encode only color information. The LAB color space plots image data in 3 dimensions along 3 independent (orthogonal) axes, one for brightness (luminance) and two for color (chrominance). Luminance corresponds closely to brightness as recorded by the brain-eye system; the chrominance (red-green and yellow-blue) axes mimic the oppositional color sensations the retina reports to the brain [Szeliski, 2010]. LAB features are densely sampled for each pixel. Also here we use the k-means algorithm to build the descriptor space. We varied the number of k visual words between 128 and 1,024 in steps of 128.

We also experimented with other visual features, such as those focusing on edges [Canny, 1986], texture [Zhu et al., 2002], and shapes [Oliva and Torralba, 2001], but they were not useful for the color tasks. Moreover, we experimented also with different color scales, such as LUV, HSV and RGB, obtaining significantly worse performance compared to LAB.

6.1.3 Multimodal models

To assemble the textual and visual representations in multimodal semantic spaces, we use the unweighted combination function introduced in Section 4.1. The choice is justified by the fact that we had to deal here with a lot of different textual and visual models, so that a straightforward to compute combination function became necessary.

6.1.4 Hybrid models

We further introduce hybrid models that exploit the patterns of co-occurrence of words as tags of the same images. Like textual models, these models are based on word co-occurrence; like visual models, they consider co-occurrence in images (image labels). In one model (**ESP-Win**, analogous to window-based models), words tagging an image were represented in terms of co-occurrence with the other tags in the image label (Baroni and Lenci Baroni and Lenci [2008] are a precedent for the use of ESP-Win). The other (**ESP-Doc**, analogous to document-based models) represented words in terms of their co-occurrence with images, using each image as a different dimension. This information is very easy to extract, as it does not require the sophisticated techniques used in computer vision. We expected these models to perform very bad; however, as we will show, they perform relatively well in all but one of the tasks tested.

6.2 Textual and visual models as general semantic models

We test the models just presented in two different ways: First, as general models of word meaning, testing their correlation to human judgements on word similarity and relatedness (this section). Second, as models of the meaning of color terms (sections 6.3 and 6.4). As in Chapter 5, we use the standard dataset WordSim353 (**WS**) and our new dataset **MEN** to test word similarity and relatedness.

Columns WS and MEN in Table 6.1 report correlations with the WordSim and MEN ratings, respectively. As expected and already verified in Section 5.2, because they are more mature and capture a broader range of semantic information, textual models perform much better than purely visual models. Also as expected, SIFT features outperform the simpler LAB features for this task.

The overall performance on MEN and WordSim seems to confirm that visual information helps, in particular for MEN, where multimodal models perform best.

Surprisingly, the newly introduced hybrid models perform quite well: They are around 10 points worse than textual and multimodal models for WordSim, and only slightly worse than multimodal models for MEN.

Model	*WS*	*MEN*	*E1*	*E2*
DM	.44	.42	3 (09)	.14
Document	.63	.62	3 (07)	.06
Window2	**.70**	.66	5 (13)	.49***
Window20	**.70**	.62	3 (11)	.53***
LAB$_{128}$	.21	.41	**1** (27)	.25*
LAB$_{256}$	.21	.41	2 (24)	.24*
LAB$_{1024}$	.19	.41	2 (24)	.28**
SIFT$_{2.5K}$	.33	.44	3 (15)	.57***
W2-LAB$_{128}$	.40	.59	**1** (27)	.40***
W2-LAB$_{256}$	.41	.60	2 (23)	.40***
W2-LAB$_{1024}$	.39	.61	2 (24)	.44***
W20-LAB$_{128}$	.40	.60	**1** (27)	.36***
W20-LAB$_{256}$	.41	.60	2 (23)	.36***
W20-LAB$_{1024}$	.39	.62	2 (24)	.40***
W2-SIFT$_{2.5K}$	.64	**.69**	2.5 (19)	.68***
W20-SIFT$_{2.5K}$	.64	.68	2 (17)	**.73*****
ESP-Doc	.52	.66	**1** (37)	.29*
ESP-Win	.55	.68	4 (15)	.16

Table 6.1: Results of the textual, visual, multimodal, and hybrid models on the general semantic tasks (first two columns, section 6.2; Pearson ρ) and Experiments 1 (E1, section 6.3) and 2 (E2, section 6.4). E1 reports the median rank of the correct color and the number of top matches (in parentheses), and E2 the average difference in normalized cosines between literal and nonliteral adjective-noun phrases, with the significance of a t-test (*** for p$<$ 0.001, ** $<$ 0.01, * $<$ 0.05).

6.3 Experiment 1: Discovering the color of concrete objects

In Experiment 1, we test the hypothesis that the relation between words denoting concrete things and words denoting their typical color is reflected by the distance of the corresponding vectors better when the models are sensitive to visual information.

6.3.1 Method

Two authors labeled by consensus a list of concrete nouns (extracted from the BLESS dataset[1] and the nouns in the BNC occurring with color terms more than 100 times) with one of the 11 colors from the basic set proposed by Berlin and Key [1969]: *black*, *blue*, *brown*, *green*, *grey*, *orange*, *pink*, *purple*, *red*, *white*, *yellow*. Objects that do not have an obvious characteristic color (*computer*) and those with more than one characteristic color (*zebra*, *bear*) were eliminated. Moreover, only nouns covered by all the models were preserved. The final list contains 52 nouns.[2] Some random examples are *fog–grey*, *crow–black*, *wood–brown*, *parsley–green*, and *grass–green*.

For evaluation, we measured the cosine of each noun with the 11 basic color words in the space produced by each model, and recorded the rank of the correct color in the resulting ordered list.

6.3.2 Results

Column E1 in Table 6.1 reports the median rank for each model (the smaller the rank, the better the model), as well as the number of exact matches (that is, number of nouns for which the model ranks the correct color first).

Discovering knowledge such that grass is green is arguably a simple task but Experiment 1 shows that textual models fail this simple task, with median ranks around 3.[3] This is consistent with the findings in Baroni and Lenci [2008] that standard distributional models do not capture the association between concrete concepts and their typical attributes. Visual models, as expected, are better at capturing the association between concepts and visual attributes. In fact, all models that are sensitive to visual information achieve median rank 1.

Multimodal models do not increase performance with respect to visual models: For instance, both W2-LAB$_{128}$ and W20-LAB$_{128}$ have the same median rank and

[1] http://sites.google.com/site/geometricalmodels/shared-evaluation

[2] Dataset available from the second author's webpage, under `resources`.

[3] We also experimented with a model based on direct co-occurrence of adjectives and nouns, obtaining promising results in a preliminary version of Exp. 1. We abandoned this approach because such a model inherently lacks scalability, as it will not generalize behind cases where the training data contain direct examples of co-occurrences of the target pairs.

number of exact matches as LAB_{128} alone. Textual information in this case is not complementary to visual information, but simply poorer.

Also note that LAB features do better than SIFT features. This is probably due to the fact that Experiment 1 is basically about identifying a large patch of color. The SIFT features we are using are also sensitive to color, but they seem to be misguided by the other cues that they extract from images. For example, pigs are pink in LAB space but brown in SIFT space, perhaps because SIFT focused on the color of the typical environment of a pig. We can thus confirm that, by limiting multimodal spaces to SIFT features, as has been done until now in the literature, we are missing important semantic information, such as the color information that we can mine with LAB.

Again we find that hybrid models do very well, in fact in this case they have the top performance, as they perform better than LAB_{128} (the difference, which can be noticed in the number of exact matches, is highly significant according to a paired Mann-Whitney test, with $p<0.001$).

6.4 Experiment 2: Discriminating between literal and nonliteral uses of color terms

Experiment 2 requires more sophisticated information than Experiment 1, as it involves distinguishing between literal and nonliteral uses of color terms.

6.4.1 Method

We test the performance of the different models with a dataset consisting of color adjective-noun phrases, randomly drawn from the most frequent 8K nouns and 4K adjectives in the concatenated ukWaC, Wackypedia, and BNC corpora (four color terms are not among these, so the dataset includes phrases for *black*, *blue*, *brown*, *green*, *red*, *white*, and *yellow* only). These were tagged by consensus by two human judges as **literal** (*white towel*, *black feather*) or **nonliteral** (*white wine*, *white musician*, *green future*). Some phrases had both literal and nonliteral uses, such as *blue book* in “book that is blue” vs. “automobile price guide”. In these cases, only the most common sense (according to the judges) was taken into account for the present experiment. The dataset consists of 370 phrases, of which our models cover 342, 227 literal and 115 nonliteral.[1]

The prediction is that, in good semantic models, literal uses will in general result in a higher similarity between the noun and color term vectors: A white towel is white, while wine or musicians are not white in the same manner. We test this prediction by comparing the average cosine between the color term and the nouns across the literal and nonliteral pairs (similar results were obtained in an evaluation in terms of prediction accuracy of a simple classifier).

6.4.2 Results

Column E2 in Table 6.1 summarizes the results of the experiment, reporting the mean difference between the normalized cosines (that is, how large the difference is between the literal and nonliteral uses of color terms), as well as the significance of the differences according to a t-test. Window-based models perform best among textual models, particularly Window20, while the rest can't discriminate between the two uses. This is particularly striking for the Document model, which performs quite well in general semantic tasks but bad in visual tasks.

Visual models are all able to discriminate between the two uses, suggesting that indeed visual information can capture nonliteral aspects of meaning. However, in this case SIFT features perform much better than LAB features, as Experiment 2 involves tackling much more sophisticated information than Experiment 1. This is consistent with the fact that, for LAB, a lower k (lower granularity of the information) performs better for Experiment 1 and a higher k (higher granularity) for Experiment 2.

One crucial question to ask, given the goals of our research, is whether textual and visual models are doing essentially the same job, only using different types of information. Note that, in this case, multimodal models increase performance over the individual modalities, and are the best models for this task. This suggests that the information used in the individual models is complementary, and indeed there is no correlation between the cosines obtained with the best textual and visual models (Pearson's $\rho = .09$, $p = .11$).

[1] Dataset available upon request to the second author.

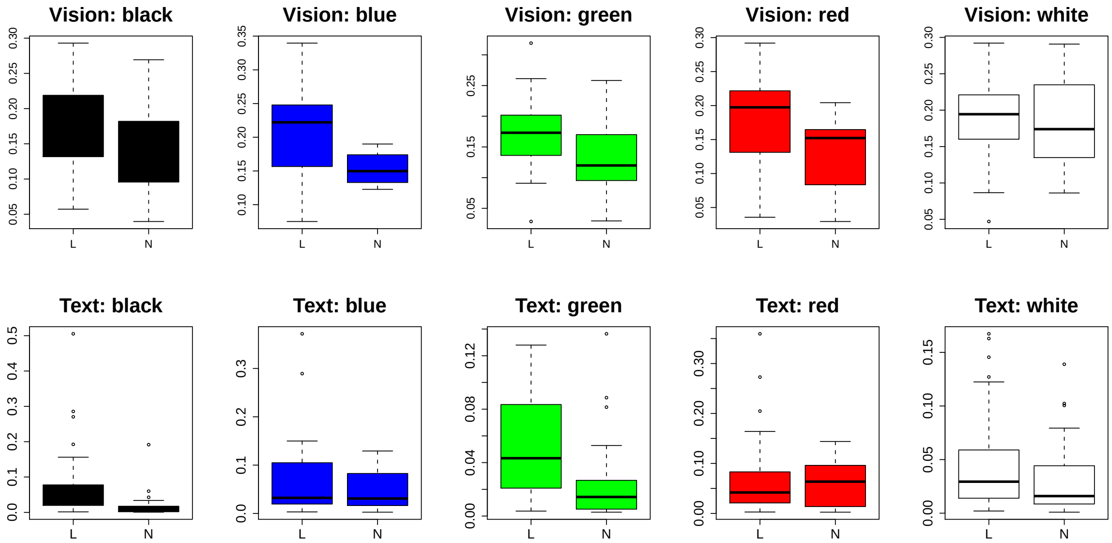


Figure 6.1: Discrimination of literal (L) vs. nonliteral (N) uses by the best visual and textual models.

Figure 6.1 depicts the results broken down by color.[1] Both modalities can capture the differences for *black* and *green*, probably because nonliteral uses of these color terms have also clear textual correlates (more concretely, topical correlates, as they are related to race and ecology, respectively).[2] Significantly, however, vision can capture nonliteral uses of *blue* and *red*, while text can't. Note that these uses (*blue note, shark, shield*, *red meat, district, face*) do not have a clear topical correlate, and thus it makes sense that vision does a better job.

Finally, note that for this more sophisticated task, hybrid models perform quite bad, which shows their limitations as models of word meaning.[3] Overall, our

[1] *Yellow* and *brown* are excluded because the dataset contains only one and two instances of nonliteral cases for these terms, respectively. The significance of the differences as explained in the text has been tested via t-tests.

[2] It's not entirely clear why neither modality can capture the differences for *white*; for text, it may be because the nonliteral cases are not so tied to race as is the cases for *black*, but they also contain many other types of nonliteral uses, such as type-referring (*white wine/rice/cell*) or metonymical ones (*white smile*).

[3] The hybrid model that performs best in the color tasks is ESP-Doc. This model can only detect a relation between an adjective and a noun if they directly co-occur in the label of at least one image (a "document" in this setting). The more direct co-occurrences there are, the more related the words will be for the model. This works for Exp. 1: Since the ESP labels are lists of what subjects saw in a picture, and the adjectives of Exp. 1 are typical colors of objects,

results suggest that co-occurrence in an image label can be used as a surrogate of true visual information to some extent, but the behavior of hybrid models depends on ad-hoc aspects of the labeled dataset, and, from an empirical perspective, they are more limited than truly multimodal models, because they require large amounts of rich verbal picture descriptions to reach good coverage.

6.5 Discussion

We have presented evidence that distributional semantic models based on text, while providing a good general semantic representation of word meaning, can be outperformed by models using visual information for semantic aspects of words where vision is relevant. We have also shown that different types of visual features (LAB, SIFT) are appropriate for different tasks. Future research should investigate automated methods to discover which (if any) kind of visual information should be highlighted in which task, more sophisticated multimodal models, visual properties other than color, and larger color datasets, such as the one recently introduced by Mohammad [2011].

there is a high co-occurrence, as all but one adjective-noun pairs co-occur in at least one ESP label. For the model to perform well in Exp. 2 too, literal phrases should occur in the same labels and non-literal pairs should not. We find no such difference (89% of adjective-noun pairs co-occur in at least one image in the literal set, 86% in the nonliteral set), because many of the relevant pairs describe concrete concepts that, while not necessarily of the "right" literal colour, are perfectly fit to be depicted in images ("blue shark", "black boy", "white wine").

Chapter 7

Using information about object location

In this Chapter we investigate how spatial information can induce better image-based distributional models. As explained in Section 3.1.2, the bag-of-visual-words pipeline suffers of a serious limitation which is the complete absence of spatial information. A first way to restore some geometry within the representation is by using spatial binning (see 3.2.1). Here we introduce a second, richer way that is object localization, and evaluate it on two tasks. In Section 7.1 we show that the object location is an important information which leads to more effective image-based semantic representations for word relatedness approximation. In Section 7.2, we continue investigating the impact of object locations by comparing image-based distributional models with conceptual representations in the brain.

7.1 Using location in multimodal distributional semantics

A natural way to advance multimodal distributional semantic models is to look at what computer vision experts are doing to improve visual feature extraction in general. Bag of Visual Words collects local image details into a global description of the image. More recent work tries to capture the location of the object itself

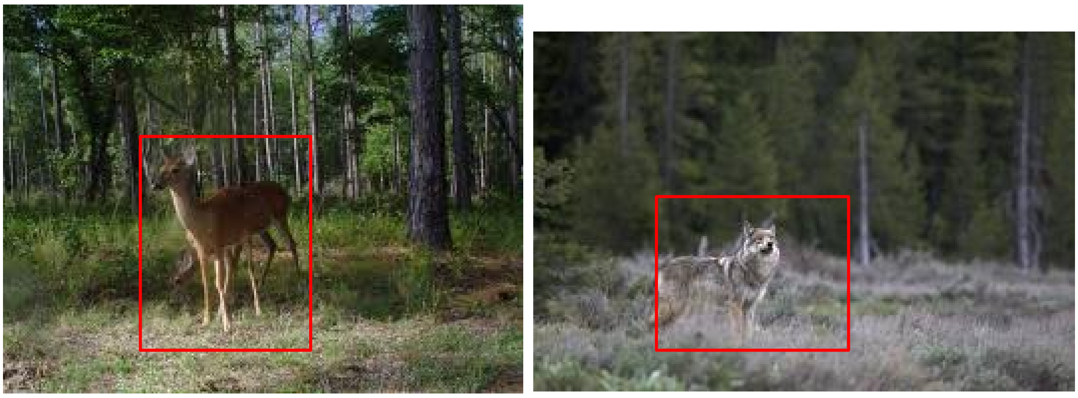


Figure 7.1: It is easier to distinguish the deer from the wolf once we localize them, but the surroundings tell us that they live in a similar environment, and are thus somewhat related concepts.

and uses features from this area only [Felzenszwalb et al., 2010; Harzallah et al., 2009; van de Sande et al., 2011]. Suppose we learned our visual model of deers from pictures such as the one on the left of Figure 7.1. We would extract features pertaining to trees and other forest elements as part of our deer representation, and we might end up wrongly classifying other wild animals living in similar surroundings (such as the wolf on the right) as deers. Clearly, a deer model learned from the bounding box containing the deer is better at distinguishing true deers from related creatures.

In this Chapter, we explore how object localization can improve the extraction of visual features to represent word meaning distributionally. However, we add a twist to the localization story: Consider again Figure 7.1. If our goal, as in the computer vision literature, is to find out that the left and right pictures contain different objects, it is a good idea to focus on features extracted from the localized objects (the two animals have different colors, their furs have different textures, their ears look different, etc). If on the other hand our goal (as in distributional semantics) is to estimate the semantic similarity or relatedness of words (and the concepts they denote), then the visual surroundings in which the localized objects occur (that is, what lies *outside* the object boxes) might be as informative, or even more so, than the object boxes themselves. In the example, the fact that the two pictures contain similar surrounds tells us that deers and wolves live in similar environments, and it is thus likely that they are somewhat related concepts. This is the distributional hypothesis transposed to images: objects that are semantically similar occur in similar visual contexts!

We use both ground-truth annotations and a state-of-the-art localization algorithm [van de Sande et al., 2011] to segment images into location boxes and surroundings, in order to check if localization improves the quality of visual features for semantic similarity measurement. Moreover, we verify whether the distributional hypothesis transfers to visual features (object vs. surrounds). Finally, we test if it is advantageous to combine object and surround features for predicting semantic similarity.

7.1.1 Semantic model construction

Visual features The visual features we use are all based on Bags of Visual Words [Csurka et al., 2004; Sivic and Zisserman, 2003], which we extract using the publicly available software of Uijlings et al. [2010]. Specifically we take local patches of 16-by-16 pixels which are sampled at every single pixel. From these patches we extract SIFT [Lowe, 2004] descriptors and two colour variants, Opponent-SIFT and RGB-SIFT, as recommended by van de Sande et al. [2010]. The visual vocabulary is created using a Random Forest [Moosmann et al., 2006] with 4 binary trees of depth 10, resulting in 4096 visual words.

Localization We use visual words in three different types of representation: **global, object**, and **surround**. The global representation uses the visual words from the whole image. The object representation uses visual words from the object location only. The surround representation uses visual words from outside the object location.

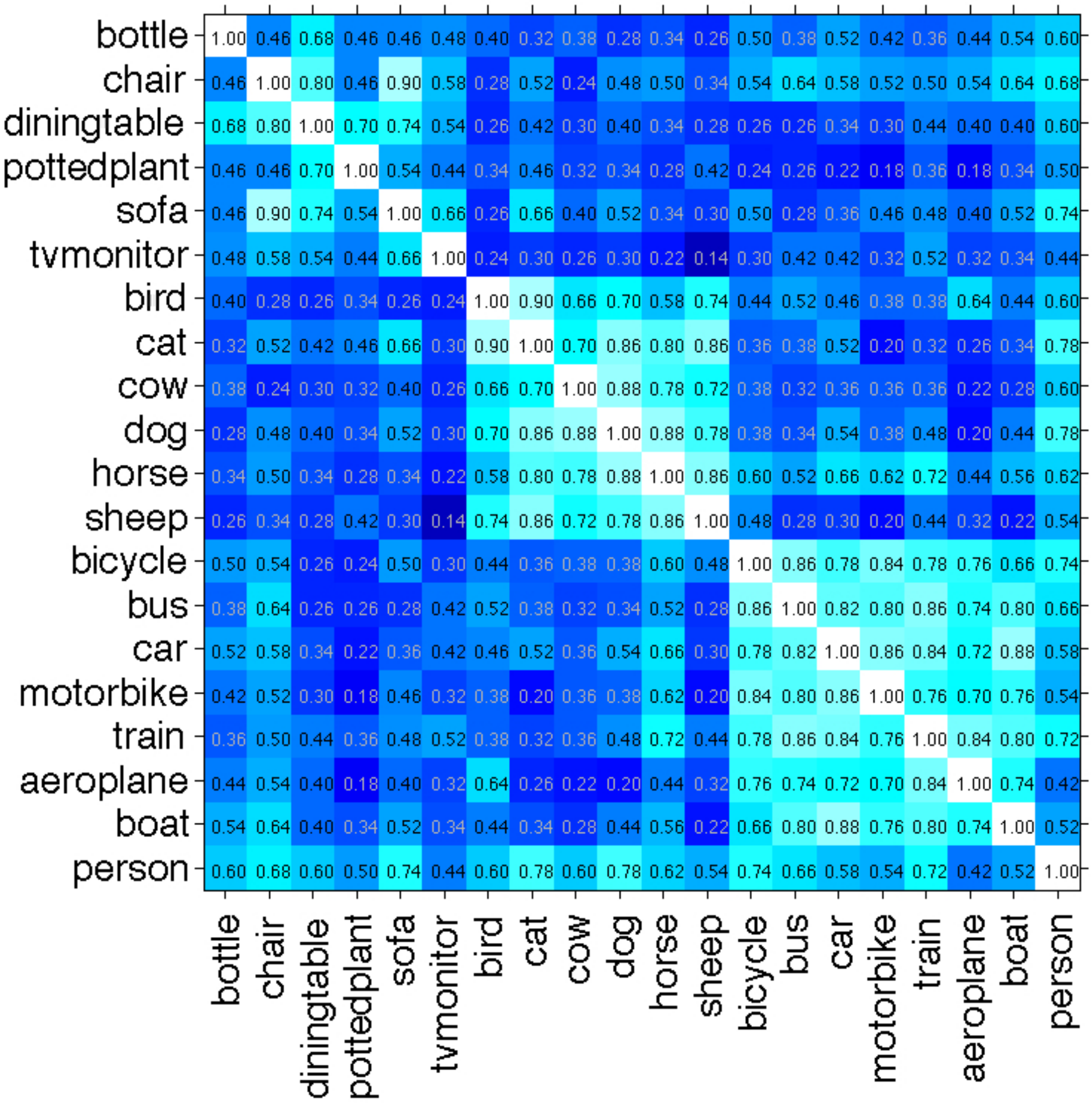
We explore two methods to distinguish object and surround: Ground-truth object location (**GL**) as given by the Pascal dataset we use [Everingham et al., 2010], and a localization method which automatically determines the object location within an image (**AL**), namely the localization algorithm of van de Sande et al. [2011]. We use annotated object locations of the target object as positive training examples and annotated object locations of non-target objects as negative training examples. We mine “hard negatives” using a single retraining step in which we test our classifiers on the training set and add, for each negative image, the highest scored location [Felzenszwalb et al., 2010; Laptev, 2009; van de Sande et al., 2011]. Most localization methods evaluate all possible locations within an image using a sliding window technique [Felzenszwalb et al., 2010; Harzallah et al., 2009; Moosmann et al., 2006; Viola and Jones, 2001]. We use instead the selective search strategy introduced in van de Sande et al. [2011]. Basically this method uses multiple complementary segmentations to generate plausible object locations, reducing the number of locations from 10^5-10^6 for sliding windows to only 1.5k, while still capturing most object locations. In the testing phase, the location with the highest classification score is considered to contain the target object. To extract locations we use a 2-by-2 Spatial Pyramid [Lazebnik et al., 2006].

Semantic vectors Each target (textual) word is associated to the list of images depicting the corresponding concept in the Pascal dataset, and the visual word occurrences across this list are summed to obtain the overall co-occurrence counts for the target. Then, raw counts are transformed into nonnegative Local Mutual Information (LMI) scores [Evert, 2005]. In this way we obtain a semantic space represented as a matrix, where the Pascal words/concepts are the rows and their visual word scores are the columns/dimensions.

7.1.2 Data

Image data set We use the Pascal VOC 2007 dataset [Everingham et al., 2010], a widely used dataset in Computer Vision with 5011 training images and 4952 test images, containing 20 concepts including *person* and subclasses of *animal, vehicle*, and *indoor* (e.g., house decoration/furniture). In all experiments we use representations extracted from test images only. We train the localization method on the training set.

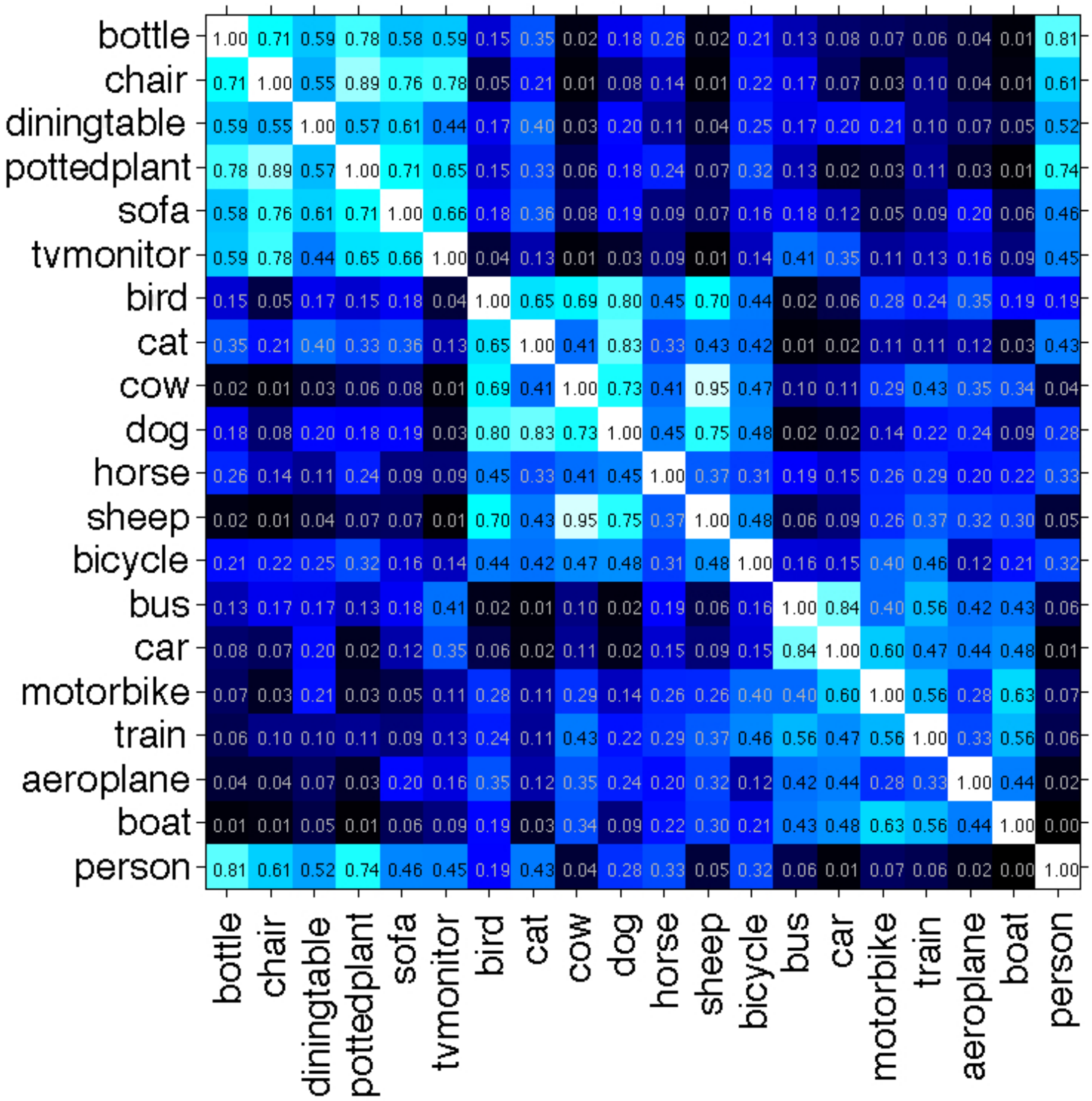
Human semantic relatedness ratings We test the models by measuring their correlation to human judgments on word similarity. We created a new evaluation benchmark for this purpose as follows. We first formed every possible pairing of



	bottle	chair	diningtable	pottedplant	sofa	tvmonitor	bird	cat	cow	dog	horse	sheep	bicycle	bus	car	motorbike	train	aeroplane	boat	person
bottle	1.00	0.46	0.68	0.46	0.46	0.48	0.40	0.32	0.38	0.28	0.34	0.26	0.50	0.38	0.52	0.42	0.36	0.44	0.54	0.60
chair	0.46	1.00	0.80	0.46	0.90	0.58	0.28	0.52	0.24	0.48	0.50	0.34	0.54	0.64	0.58	0.52	0.50	0.54	0.64	0.68
diningtable	0.68	0.80	1.00	0.70	0.74	0.54	0.26	0.42	0.30	0.40	0.34	0.28	0.26	0.26	0.34	0.30	0.44	0.40	0.40	0.60
pottedplant	0.46	0.46	0.70	1.00	0.54	0.44	0.34	0.46	0.32	0.34	0.28	0.42	0.24	0.26	0.22	0.18	0.36	0.18	0.34	0.50
sofa	0.46	0.90	0.74	0.54	1.00	0.66	0.26	0.66	0.40	0.52	0.34	0.30	0.50	0.28	0.36	0.46	0.48	0.40	0.52	0.74
tvmonitor	0.48	0.58	0.54	0.44	0.66	1.00	0.24	0.30	0.26	0.30	0.22	0.14	0.30	0.42	0.42	0.32	0.52	0.32	0.34	0.44
bird	0.40	0.28	0.26	0.34	0.26	0.24	1.00	0.90	0.66	0.70	0.58	0.74	0.44	0.52	0.46	0.38	0.38	0.64	0.44	0.60
cat	0.32	0.52	0.42	0.46	0.66	0.30	0.90	1.00	0.70	0.86	0.80	0.86	0.36	0.38	0.52	0.20	0.32	0.26	0.34	0.78
cow	0.38	0.24	0.30	0.32	0.40	0.26	0.66	0.70	1.00	0.88	0.78	0.72	0.38	0.32	0.36	0.36	0.36	0.22	0.28	0.60
dog	0.28	0.48	0.40	0.34	0.52	0.30	0.70	0.86	0.88	1.00	0.88	0.78	0.38	0.34	0.54	0.38	0.48	0.20	0.44	0.78
horse	0.34	0.50	0.34	0.28	0.34	0.22	0.58	0.80	0.78	0.88	1.00	0.86	0.60	0.52	0.66	0.62	0.72	0.44	0.56	0.62
sheep	0.26	0.34	0.28	0.42	0.30	0.14	0.74	0.86	0.72	0.78	0.86	1.00	0.48	0.28	0.30	0.20	0.44	0.32	0.22	0.54
bicycle	0.50	0.54	0.26	0.24	0.50	0.30	0.44	0.36	0.38	0.38	0.60	0.48	1.00	0.86	0.78	0.84	0.78	0.76	0.66	0.74
bus	0.38	0.64	0.26	0.26	0.28	0.42	0.52	0.38	0.32	0.34	0.52	0.28	0.86	1.00	0.82	0.80	0.86	0.74	0.80	0.66
car	0.52	0.58	0.34	0.22	0.36	0.42	0.46	0.52	0.36	0.54	0.66	0.30	0.78	0.82	1.00	0.86	0.84	0.72	0.88	0.58
motorbike	0.42	0.52	0.30	0.18	0.46	0.32	0.38	0.20	0.36	0.38	0.62	0.20	0.84	0.80	0.86	1.00	0.76	0.70	0.76	0.54
train	0.36	0.50	0.44	0.36	0.48	0.52	0.38	0.32	0.36	0.48	0.72	0.44	0.78	0.86	0.84	0.76	1.00	0.84	0.80	0.72
aeroplane	0.44	0.54	0.40	0.18	0.40	0.32	0.64	0.26	0.22	0.20	0.44	0.32	0.76	0.74	0.72	0.70	0.84	1.00	0.74	0.42
boat	0.54	0.64	0.40	0.34	0.52	0.34	0.44	0.34	0.28	0.44	0.56	0.22	0.66	0.80	0.88	0.76	0.80	0.74	1.00	0.52
person	0.60	0.68	0.60	0.50	0.74	0.44	0.60	0.78	0.60	0.78	0.62	0.54	0.74	0.66	0.58	0.54	0.72	0.42	0.52	1.00

Figure 7.2: Similarity matrices for the human subjects. Lighter color cues higher similarity.

the 20 Pascal concept words, obtaining 190 pairs in total. Then we obtained relatedness ratings for the pairs by crowdsourcing using Amazon Mechanical Turk. We presented Turkers with a list of two candidate word pairs, each pair randomly matched with a comparison pair sampled without replacement from the same list and rated in this setting (as either more or less related than the comparison point) by a single Turker. In total, each pair was rated in this way against 50 comparison



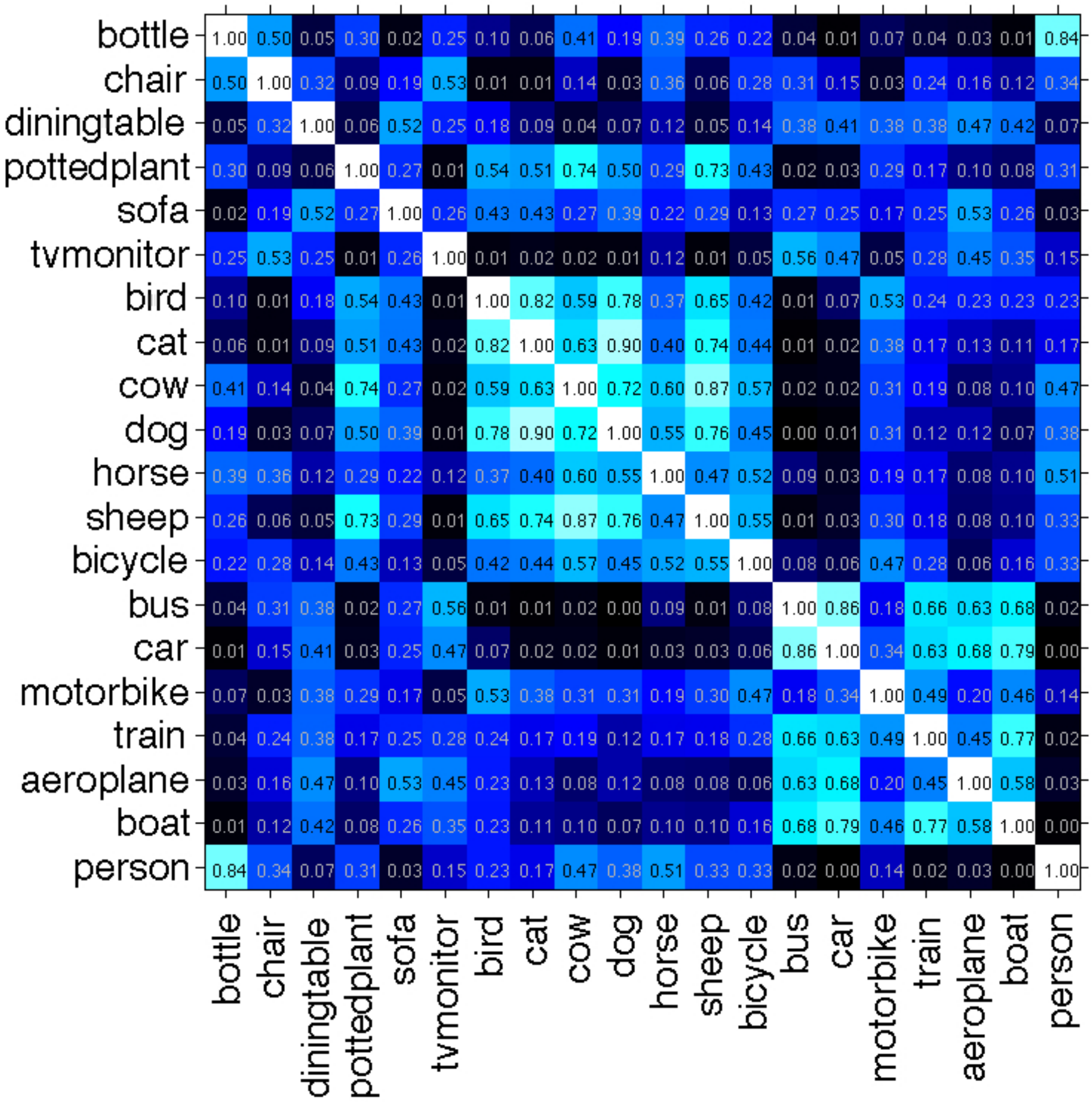
	bottle	chair	diningtable	pottedplant	sofa	tvmonitor	bird	cat	cow	dog	horse	sheep	bicycle	bus	car	motorbike	train	aeroplane	boat	person
bottle	1.00	0.71	0.59	0.78	0.58	0.59	0.15	0.35	0.02	0.18	0.26	0.02	0.21	0.13	0.08	0.07	0.06	0.04	0.01	0.81
chair	0.71	1.00	0.55	0.89	0.76	0.78	0.05	0.21	0.01	0.08	0.14	0.01	0.22	0.17	0.07	0.03	0.10	0.04	0.01	0.61
diningtable	0.59	0.55	1.00	0.57	0.61	0.44	0.17	0.40	0.03	0.20	0.11	0.04	0.25	0.17	0.20	0.21	0.10	0.07	0.05	0.52
pottedplant	0.78	0.89	0.57	1.00	0.71	0.65	0.15	0.33	0.06	0.18	0.24	0.07	0.32	0.13	0.02	0.03	0.11	0.03	0.01	0.74
sofa	0.58	0.76	0.61	0.71	1.00	0.66	0.18	0.36	0.08	0.19	0.09	0.07	0.16	0.18	0.12	0.05	0.09	0.20	0.06	0.46
tvmonitor	0.59	0.78	0.44	0.65	0.66	1.00	0.04	0.13	0.01	0.03	0.09	0.01	0.14	0.41	0.35	0.11	0.13	0.16	0.09	0.45
bird	0.15	0.05	0.17	0.15	0.18	0.04	1.00	0.65	0.69	0.80	0.45	0.70	0.44	0.02	0.06	0.28	0.24	0.35	0.19	0.19
cat	0.35	0.21	0.40	0.33	0.36	0.13	0.65	1.00	0.41	0.83	0.33	0.43	0.42	0.01	0.02	0.11	0.11	0.12	0.03	0.43
cow	0.02	0.01	0.03	0.06	0.08	0.01	0.69	0.41	1.00	0.73	0.41	0.95	0.47	0.10	0.11	0.29	0.43	0.35	0.34	0.04
dog	0.18	0.08	0.20	0.18	0.19	0.03	0.80	0.83	0.73	1.00	0.45	0.75	0.48	0.02	0.02	0.14	0.22	0.24	0.09	0.28
horse	0.26	0.14	0.11	0.24	0.09	0.09	0.45	0.33	0.41	0.45	1.00	0.37	0.31	0.19	0.15	0.26	0.29	0.20	0.22	0.33
sheep	0.02	0.01	0.04	0.07	0.07	0.01	0.70	0.43	0.95	0.75	0.37	1.00	0.48	0.06	0.09	0.26	0.37	0.32	0.30	0.05
bicycle	0.21	0.22	0.25	0.32	0.16	0.14	0.44	0.42	0.47	0.48	0.31	0.48	1.00	0.16	0.15	0.40	0.46	0.12	0.21	0.32
bus	0.13	0.17	0.17	0.13	0.18	0.41	0.02	0.01	0.10	0.02	0.19	0.06	0.16	1.00	0.84	0.40	0.56	0.42	0.43	0.06
car	0.08	0.07	0.20	0.02	0.12	0.35	0.06	0.02	0.11	0.02	0.15	0.09	0.15	0.84	1.00	0.60	0.47	0.44	0.48	0.01
motorbike	0.07	0.03	0.21	0.03	0.05	0.11	0.28	0.11	0.29	0.14	0.26	0.26	0.40	0.40	0.60	1.00	0.56	0.28	0.63	0.07
train	0.06	0.10	0.10	0.11	0.09	0.13	0.24	0.11	0.43	0.22	0.29	0.37	0.46	0.56	0.47	0.56	1.00	0.33	0.56	0.06
aeroplane	0.04	0.04	0.07	0.03	0.20	0.16	0.35	0.12	0.35	0.24	0.20	0.32	0.12	0.42	0.44	0.28	0.33	1.00	0.44	0.02
boat	0.01	0.01	0.05	0.01	0.06	0.09	0.19	0.03	0.34	0.09	0.22	0.30	0.21	0.43	0.48	0.63	0.56	0.44	1.00	0.00
person	0.81	0.61	0.52	0.74	0.46	0.45	0.19	0.43	0.04	0.28	0.33	0.05	0.32	0.06	0.01	0.07	0.06	0.02	0.00	1.00

Figure 7.3: Similarity matrices for Global. Lighter color cues higher similarity.

pairs, thus obtaining a final score on a 50-point scale (then normalized between 0 and 1), although the Turkers' had to make simple binary choices.

7.1.3 Results

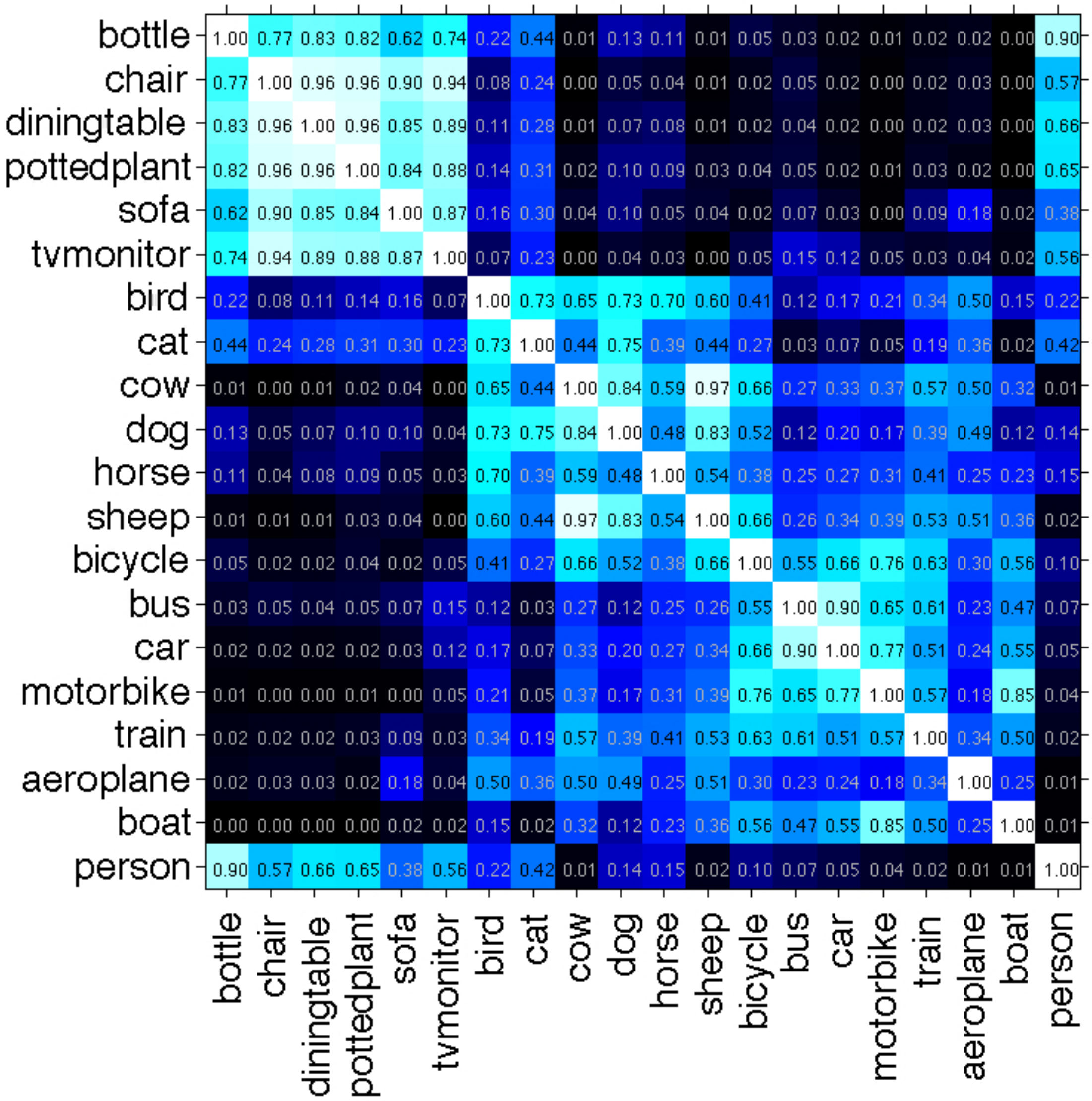
Table 7.1 reports the correlations of our models with human similarity ratings (p<0.0001 for all correlations). The model based on the global approach (Global)



	bottle	chair	diningtable	pottedplant	sofa	tvmonitor	bird	cat	cow	dog	horse	sheep	bicycle	bus	car	motorbike	train	aeroplane	boat	person
bottle	1.00	0.50	0.05	0.30	0.02	0.25	0.10	0.06	0.41	0.19	0.39	0.26	0.22	0.04	0.01	0.07	0.04	0.03	0.01	0.84
chair	0.50	1.00	0.32	0.09	0.19	0.53	0.01	0.01	0.14	0.03	0.36	0.06	0.28	0.31	0.15	0.03	0.24	0.16	0.12	0.34
diningtable	0.05	0.32	1.00	0.06	0.52	0.25	0.18	0.09	0.04	0.07	0.12	0.05	0.14	0.38	0.41	0.38	0.38	0.47	0.42	0.07
pottedplant	0.30	0.09	0.06	1.00	0.27	0.01	0.54	0.51	0.74	0.50	0.29	0.73	0.43	0.02	0.03	0.29	0.17	0.10	0.08	0.31
sofa	0.02	0.19	0.52	0.27	1.00	0.26	0.43	0.43	0.27	0.39	0.22	0.29	0.13	0.27	0.25	0.17	0.25	0.53	0.26	0.03
tvmonitor	0.25	0.53	0.25	0.01	0.26	1.00	0.01	0.02	0.02	0.01	0.12	0.01	0.05	0.56	0.47	0.05	0.28	0.45	0.35	0.15
bird	0.10	0.01	0.18	0.54	0.43	0.01	1.00	0.82	0.59	0.78	0.37	0.65	0.42	0.01	0.07	0.53	0.24	0.23	0.23	0.23
cat	0.06	0.01	0.09	0.51	0.43	0.02	0.82	1.00	0.63	0.90	0.40	0.74	0.44	0.01	0.02	0.38	0.17	0.13	0.11	0.17
cow	0.41	0.14	0.04	0.74	0.27	0.02	0.59	0.63	1.00	0.72	0.60	0.87	0.57	0.02	0.02	0.31	0.19	0.08	0.10	0.47
dog	0.19	0.03	0.07	0.50	0.39	0.01	0.78	0.90	0.72	1.00	0.55	0.76	0.45	0.00	0.01	0.31	0.12	0.12	0.07	0.38
horse	0.39	0.36	0.12	0.29	0.22	0.12	0.37	0.40	0.60	0.55	1.00	0.47	0.52	0.09	0.03	0.19	0.17	0.08	0.10	0.51
sheep	0.26	0.06	0.05	0.73	0.29	0.01	0.65	0.74	0.87	0.76	0.47	1.00	0.55	0.01	0.03	0.30	0.18	0.08	0.10	0.33
bicycle	0.22	0.28	0.14	0.43	0.13	0.05	0.42	0.44	0.57	0.45	0.52	0.55	1.00	0.08	0.06	0.47	0.28	0.06	0.16	0.33
bus	0.04	0.31	0.38	0.02	0.27	0.56	0.01	0.01	0.02	0.00	0.09	0.01	0.08	1.00	0.86	0.18	0.66	0.63	0.68	0.02
car	0.01	0.15	0.41	0.03	0.25	0.47	0.07	0.02	0.02	0.01	0.03	0.03	0.06	0.86	1.00	0.34	0.63	0.68	0.79	0.00
motorbike	0.07	0.03	0.38	0.29	0.17	0.05	0.53	0.38	0.31	0.31	0.19	0.30	0.47	0.18	0.34	1.00	0.49	0.20	0.46	0.14
train	0.04	0.24	0.38	0.17	0.25	0.28	0.24	0.17	0.19	0.12	0.17	0.18	0.28	0.66	0.63	0.49	1.00	0.45	0.77	0.02
aeroplane	0.03	0.16	0.47	0.10	0.53	0.45	0.23	0.13	0.08	0.12	0.08	0.08	0.06	0.63	0.68	0.20	0.45	1.00	0.58	0.03
boat	0.01	0.12	0.42	0.08	0.26	0.35	0.23	0.11	0.10	0.07	0.10	0.10	0.16	0.68	0.79	0.46	0.77	0.58	1.00	0.00
person	0.84	0.34	0.07	0.31	0.03	0.15	0.23	0.17	0.47	0.38	0.51	0.33	0.33	0.02	0.00	0.14	0.02	0.03	0.00	1.00

Figure 7.4: Similarity matrices for AL-Object. Lighter color cues higher similarity.

already achieves a good correlation. Under both ground-truth (GL) and automated (AL) localization we observe the same intriguing pattern: The models outperform Global when they rely on features extracted from the surroundings, but not when they rely on features coming from the localized objects. This shows how the distributional hypothesis transfers to images: The context (surround) of an object is a good indicator of its semantics, even more so than the appearance



	bottle	chair	diningtable	pottedplant	sofa	tvmonitor	bird	cat	cow	dog	horse	sheep	bicycle	bus	car	motorbike	train	aeroplane	boat	person
bottle	1.00	0.77	0.83	0.82	0.62	0.74	0.22	0.44	0.01	0.13	0.11	0.01	0.05	0.03	0.02	0.01	0.02	0.02	0.00	0.90
chair	0.77	1.00	0.96	0.96	0.90	0.94	0.08	0.24	0.00	0.05	0.04	0.01	0.02	0.05	0.02	0.00	0.02	0.03	0.00	0.57
diningtable	0.83	0.96	1.00	0.96	0.85	0.89	0.11	0.28	0.01	0.07	0.08	0.01	0.02	0.04	0.02	0.00	0.02	0.03	0.00	0.66
pottedplant	0.82	0.96	0.96	1.00	0.84	0.88	0.14	0.31	0.02	0.10	0.09	0.03	0.04	0.05	0.02	0.01	0.03	0.02	0.00	0.65
sofa	0.62	0.90	0.85	0.84	1.00	0.87	0.16	0.30	0.04	0.10	0.05	0.04	0.02	0.07	0.03	0.00	0.09	0.18	0.02	0.38
tvmonitor	0.74	0.94	0.89	0.88	0.87	1.00	0.07	0.23	0.00	0.04	0.03	0.00	0.05	0.15	0.12	0.05	0.03	0.04	0.02	0.56
bird	0.22	0.08	0.11	0.14	0.16	0.07	1.00	0.73	0.65	0.73	0.70	0.60	0.41	0.12	0.17	0.21	0.34	0.50	0.15	0.22
cat	0.44	0.24	0.28	0.31	0.30	0.23	0.73	1.00	0.44	0.75	0.39	0.44	0.27	0.03	0.07	0.05	0.19	0.36	0.02	0.42
cow	0.01	0.00	0.01	0.02	0.04	0.00	0.65	0.44	1.00	0.84	0.59	0.97	0.66	0.27	0.33	0.37	0.57	0.50	0.32	0.01
dog	0.13	0.05	0.07	0.10	0.10	0.04	0.73	0.75	0.84	1.00	0.48	0.83	0.52	0.12	0.20	0.17	0.39	0.49	0.12	0.14
horse	0.11	0.04	0.08	0.09	0.05	0.03	0.70	0.39	0.59	0.48	1.00	0.54	0.38	0.25	0.27	0.31	0.41	0.25	0.23	0.15
sheep	0.01	0.01	0.01	0.03	0.04	0.00	0.60	0.44	0.97	0.83	0.54	1.00	0.66	0.26	0.34	0.39	0.53	0.51	0.36	0.02
bicycle	0.05	0.02	0.02	0.04	0.02	0.05	0.41	0.27	0.66	0.52	0.38	0.66	1.00	0.55	0.66	0.76	0.63	0.30	0.56	0.10
bus	0.03	0.05	0.04	0.05	0.07	0.15	0.12	0.03	0.27	0.12	0.25	0.26	0.55	1.00	0.90	0.65	0.61	0.23	0.47	0.07
car	0.02	0.02	0.02	0.02	0.03	0.12	0.17	0.07	0.33	0.20	0.27	0.34	0.66	0.90	1.00	0.77	0.51	0.24	0.55	0.05
motorbike	0.01	0.00	0.00	0.01	0.00	0.05	0.21	0.05	0.37	0.17	0.31	0.39	0.76	0.65	0.77	1.00	0.57	0.18	0.85	0.04
train	0.02	0.02	0.02	0.03	0.09	0.03	0.34	0.19	0.57	0.39	0.41	0.53	0.63	0.61	0.51	0.57	1.00	0.34	0.50	0.02
aeroplane	0.02	0.03	0.03	0.02	0.18	0.04	0.50	0.36	0.50	0.49	0.25	0.51	0.30	0.23	0.24	0.18	0.34	1.00	0.25	0.01
boat	0.00	0.00	0.00	0.00	0.02	0.02	0.15	0.02	0.32	0.12	0.23	0.36	0.56	0.47	0.55	0.85	0.50	0.25	1.00	0.01
person	0.90	0.57	0.66	0.65	0.38	0.56	0.22	0.42	0.01	0.14	0.15	0.02	0.10	0.07	0.05	0.04	0.02	0.01	0.01	1.00

Figure 7.5: Similarity matrices for AL-Context. Lighter color cues higher similarity.

of the object itself. Concatenating the two localized feature channels (object and surround) results in the highest overall correlation. This confirms the intuition that object and context appearance should be distinguished, but both taken into account. Finally, the automated segmentation model is as good as the manual one.

To gain qualitative insights into what the models are doing, we looked at the

Model	ρ	*Model*	ρ	*Model*	ρ
Global	47	GL-Object	39	AL-Context	36
		GL-Context	50	AL-Context	51
		GL-Object&Context	**54**	AL-Object&Context	**54**

Table 7.1: Percentage Spearman correlations of the models with human semantic relatedness intuitions for the Pascal concepts.

similarity matrices obtained by comparing concepts with each other in terms of cosine similarity, also checking how the overall model-based patterns compare to human intuitions (Figure 7.2, 7.3, 7.4 and 7.5). We focus on AL because it has comparable performance to ground-truth segmentation, but it is more generally useful.

By looking first at the human similarity matrix, we notice blocks corresponding to the three classes of *indoor objects, animals* and *vehicles.* While *animals* and *vehicles* are clearly discernible, the *indoor* class does not emerge with the same clarity, due to phenomena such as *chair* being also semantically related to *bus* and *boat* (buses and boats typically contain chairs), or *cat* being related to *sofa* (one of the most likely place where you can find one). More in general, the *indoor* class shares some semantic properties with both *animals* and *vehicles*, which makes its borders somewhat fuzzy. Looking now at the model matrices, Global is capturing the same division into classes, albeit with the opposite pattern compared to humans: much more fuzziness in *animals* and *vehicles* than in the *indoor* class. AL-Object is better at grouping together *animals* and *vehicles* (except for *bicycle* and *motorbike*, not entirely unreasonably misplaced into *animals*), but there isn't any clear grouping of *indoor objects.* AL-Context is capturing the three classes very clearly, with just a bit more confusion between *animals* and *vehicles* (*aeroplane* being nearer to *animals* than to most *vehicles*). This model captures the *indoor* cluster particularly well, probably because "indoorness" is mostly a property of the surroundings of an object.

7.1.4 Discussion

Our results suggest that localization techniques from computer vision are mature enough to help visual feature extraction for multimodal distributional semantic models. Moreover, we showed how the distributional hypothesis transfers to images: The appearance of the surroundings of an object carry more semantic information than the appearance of the object itself (but the best results are obtained by capturing the object and surround separately and using their combination). Interestingly, for object recognition opposite results were obtained. Uijlings et al. [2011] found that an object's appearance is most important for its recognition, and once its location is known the surrounds contain little extra information. Hence the informative parts of the image are markedly different for object recognition and semantic similarity.

7.2 Correlating image-based distributional semantic models with neural representations of concepts

In the previous Section we have shown that the object location within an image induces more effective image-based distributional models. In this Section, we test weather image-based models capture the semantic patterns that emerge from fMRI recordings of the neural signal. We consider this as interesting test for localization since, as we will discuss below, it is natural to hypothesize that different lobes of the brain might process information that is more pertinent to the visual object with respect to the context and vice versa.

7.2.1 Background

Many recent neuroscientific studies have brought support to the view that concepts are represented in terms of patterns of neural activation over broad areas, naturally encoded as vectors in a neural semantic space [Haxby et al., 2001; Huth et al., 2012]. Similar representations are also widely used in computational linguistics, and in particular in distributional semantics [Clark, 2013; Erk, 2012;

Turney and Pantel, 2010], that captures meaning in terms of vectors recording the patterns of co-occurrence of words in large corpora, under the hypothesis that words that occur in similar contexts are similar in meaning.

Since the seminal work of Mitchell et al. [2008], there has thus being interest in investigating whether corpus-harvested semantic representations can contribute to the study of concepts in the brain. The relation is mutually beneficial: From the point of view of brain activity decoding, a strong correlation between corpus-based and brain-derived conceptual representations would mean that we could use the former (much easier to construct on a very large scale) to make inferences about the second: e.g., using corpus-based representations to reconstruct the likely neural signal associated to words we have no direct brain data for. From the point of view of computational linguistics, neural data provide the ultimate testing ground for models that strive to capture important aspects of human semantic memory (much more so than the commonly used explicit semantic rating benchmarks). If we found that a corpus-based model of meaning can make non-trivial predictions about the structure of the semantic space in the brain, that would make a pretty strong case for the intriguing idea that the model is approximating, in interesting ways, the way in which humans acquire and represent semantic knowledge.

We take as our starting point the extensive experiments reported in Murphy et al. [2012], who showed that purely corpus-based distributional models are at least as good at brain signal prediction tasks as earlier models that made use of manually-generated or controlled knowledge sources [Chang et al., 2011; Palatucci et al., 2009; Pereira et al., 2011], and we extend the analysis to our multimodal distributional models, with the usual hypothesis that our new generation of distributional models provides a more realistic view of meaning.

The first question that we ask, thus, is whether the more "grounded" image-based models can help us in interpreting conceptual representations in the brain. More specifically, we will compare the performance of different image-based representations, and we will test whether text- and image-based representations are complementary, so that when used together they can better account for patterns in neural data. Finally, we will check for differences between anatomical regions in the degree to which text and/or image models are effective, as one might expect given the well-known functional specializations of different anatomical regions.

7.2.2 Brain data

We use the data that were recorded and preprocessed by Mitchell et al. [2008], available for download in their supporting online material.[1] Full details of the experimental protocol, data acquisition and preprocessing can be found in Mitchell et al. [2008] and the supporting material. Key points are that there were nine right-handed adult participants (5 female, age between 18 and 32). The experimental task was to actively think about the properties of sixty objects that were presented visually, each as a line drawing in combination with a text label. The entire set of objects was presented in a random order in six sessions, each object remained on screen for 3 seconds with a seven second fixation gap between presentations.

Mitchell and colleagues examined 12 categories, five objects per category, for a total of 60 concepts (words). Due to coverage limitations, we use 51/60 words representing 11/12 categories. Table 7.2 contains the full list of 51 words organized by category.

Animals	Bear, Cat, Cow, Dog Horse
Building	Apartment, Barn, Church, House
Building parts	Arch, Chimney, Closet, Door, Window
Clothing	Coat, Dress, Pants, Shirt, Skirt
Furniture	Bed, Chair, Desk, Dresser, Table
Insect	Ant, Bee, Beetle, Butterfly, Fly
Kitchen utensils	Bottle, Cup, Glass, Knife, Spoon
Man made objects	Bell, Key, Refrigerator, Telephone, Watch
Tool	Chisel, Hammer, Screwdriver
Vegetable	Celery, Corn, Lettuce, Tomato
Vehicle	Airplane, Bicycle, Car, Train, Truck

Table 7.2: The 51 words represented by the brain and the distributional models, organized by category.

[1] http://www.cs.cmu.edu/~tom/science2008/

fMRI acquisition and preprocessing Mitchell et al. [2008] acquired functional images on a Siemens Allegra 3.0T scanner using a gradient echo EPI pulse sequence with TR=1000 ms, TE=30 ms and a 60° angle. Seventeen 5-mm thick oblique-axial slices were imaged with a gap of 1-mm between slices. The acquisition matrix was 64×64 with 3.125×3.125×5-mm voxels. They subsequently corrected data for slice timing, motion, linear trend, and performed temporal smoothing with a high-pass filter at 190s cutoff. The data were normalized to the MNI template brain image, spatially normalized into MNI space and resampled to 3×3×6 mm^3 voxels. The voxel-wise percent signal change relative to the fixation condition was computed for each object presentation. The mean of the four images acquired 4s post stimulus presentation was used for analysis.

To create a single representation per object per participant, we took the voxel-wise mean of the six presentations of each word. Likewise to create a single representation per category per participant, we took the voxel-wise mean of all word models per category, per participant.

Anatomical parcellation Analysis was conducted on the whole brain, and to address the question of whether there are differences in models' effectiveness between anatomical regions, brains were further partitioned into frontal, parietal, temporal and occipital lobes. This partitioning is coarse (each lobe is large and serves many diverse functions), but, for an initial test, appropriate, given that each lobe has specialisms that on face value are amenable to interpretation by our different distributional models and the exact nature of specialist processing in localised areas is often subject to debate (so being overly restrictive may be risky). Formulation of the distributional models is described in detail in the Section 7.2.3, but for now it is sufficient to know that the Object model is derived from image statistics of the object depicted in images, Context from image statistics of the background scene, Object&Context is a combination of the two, and Window2 is a text-based model.

The *occipital* lobe houses the primary visual processing system and consequently it is reasonable to expect some bias toward image-based semantic models. Furthermore, given that experimental stimuli incorporated line drawings of the object,and the visual cortex has a well-established role in processing low-level visual statistics including edge detection [Bruce et al., 2003], we naturally expected a good performance from Object (formulated from edge orientation histograms of similar objects).

Following Goodale and Milner [1992]'s influential perception-action model (see McIntosh and Schenk [2009] for recent discussion), visual information is channeled from the occipital lobe in two streams: a perceptual stream, serving object identification and recognition; and an action stream, specialist in processing egocentric spatial relationships and ultimately supporting interaction with the world.

The perceptual stream leads to the *temporal lobe*. Here the fusiform gyrus (shared with the occipital lobe) plays a general role in object categorisation (e.g., animals and tools [Chao et al., 1999], faces [Kanwisher and Yovel, 2006], body parts [Peelen and Downing, 2005] and even word form perception [McCandliss et al., 2003]). As the parahippocampus is strongly associated with scene representation [Epstein, 2008], we expect both the Object and Context models to capture variability in the temporal lobe. Of wider relevance to semantic processing, the medial temporal gyrus, inferior temporal gyrus and ventral temporal lobe have generally been implicated to have roles in supramodal integration and concept retrieval [Binder et al., 2009]. Given this, we expected that incorporating text would also be valuable and that the Window2&Object&Context combination would be a good model.

The visual action stream leads from the occipital lobe to the *parietal* lobe to support spatial cognition tasks and action control [Sack, 2009]. In that there seems to be an egocentric frame of reference, placing actor in environment, it is tempting to speculate that the Context model is more appropriate than the Object model here. As the parietal lobe also contains the angular gyrus, thought to be involved in complex, supra-modal information integration and knowledge retrieval [Binder et al., 2009], we might again forecast that integrating text and image information would boost performance, so Window2&Context was earmarked as a strong candidate.

The *frontal lobe*, is traditionally associated with high-level processing and manipulation of abstract knowledge and rules and controlled behaviour [Miller et al., 2002]. Regarding semantics, the dorsomedial prefrontal cortex has been implicated in self-guided retrieval of semantic information (e.g., uncued speech production), the ventromedial prefrontal cortex in motivation and emotional processing, the inferior frontal gyrus in phonological and syntactic processing, [Binder et al., 2009] and integration of lexical information [Hagoort, 2005]. Given the association with linguistic processing we anticipated a bias in favour of Window2.

The four lobes were identified and partitioned using Tzourio-Mazoyer et al. [2002]'s automatic anatomical labelling scheme.

Voxel selection The set of 500 most stable voxels, both within the whole brain and from within each region of interest were identified for analysis. The most stable voxels were those showing consistent variation across the different stimuli between scanning sessions. Specifically, and following a similar strategy to Mitchell et al. [2008], for each voxel, the set of 51 words from each unique pair of scanning sessions were correlated using Pearson's correlation (6 sessions and therefore 15 unique pairs), and the mean of the 15 resulting correlation coefficients was taken as the measure of stability. The 500 voxels with highest mean correlations were selected.

7.2.3 Distributional models

7.2.3.1 Textual models

Verb We experiment with the original text-based semantic model used to predict fMRI patterns by Mitchell et al. [2008]. Each object stimulus word is represented as a 25-dimensional vector, with each value corresponding to the normalized sentence-wide co-occurrence of that word with one of 25 manually-picked sensorimotor verbs (such as *see*, *hear*, *eat*, . . .) in a trillion word text corpus.

Window2 We experiment also with Window2 (see Section 3.3 for its description). In Murphy et al. [2012] a window-based model very similar to ours was not significantly worse than their best model for brain decoding. We tried also a few variations, e.g., using a larger window or different transformations on the raw co-occurrences from those presented below, but with little, insignificant changes in performance. Given that our focus here is on visual information, we only report results for Window2 and its combination with visual models.

7.2.3.2 Visual models

As in Section 7.1, to build our visual models we exploit also the information about object location. Therefore, the visual features (visual words) are extracted from the object bounding box (in our experiments, the **Object** model) or from only outside the object box (the **Context** model). A combined model is obtained by concatenating the two feature vectors (the **Object&Context** model). Note that here we do not exploit a global model, namely a model which is not based on localization, having already shown the superiority of Object&Context in the previous Section.

Visual model construction pipeline Differently from the previous Section and also all other reported experiments, for these experiments we use images from ImageNet [Deng et al., 2009],[1] a very large image database organized on top of the WordNet hierarchy [Fellbaum, 1998]. ImageNet has more than 14 million images, covering 21K WordNet nominal synsets. The previous experiment with the Pascal dataset showed the usefulness of localization, while for these experiments we switch to ImageNet because it guarantees a better coverage of the Mitchell fMRI data in terms of images annotated with object bounding boxes.

To build visual distributional models, we utilize the visual pipeline described in Section 3.1. More in the detail, as low-level features we use SIFT. To construct the visual vocabulary, we cluster the SIFT features into 25K different clusters.[2] We add also spatial information by dividing the image into several subregions, representing each of them in terms of BoVW and then stacking the resulting histograms [Lazebnik et al., 2006]. We use in total 8 different regions, obtaining a final vector of 200K dimensions (25K visual words × 8 regions). Since each concept in our dataset is represented by multiple images, we pool the visual word occurrences across images by summing them up into a single vector.

To perform the entire visual pipeline we use VSEM (see Appendix 9.1).

[1] http://www.image-net.org/

[2] We use k-means, the most commonly employed clustering algorithm for this task.

7.2.3.3 Model transformations and combination

Once both the textual and the visual models are built, we perform two different transformations on the raw co-occurrence counts. First, we transform them into nonnegative Pointwise Mutual Information (PMI) association scores [Church and Hanks, 1990]. As a second transformation, we apply dimensionality reduction to the two matrices. In particular, we adopt the Singular Value Decomposition (SVD), one of the most effective methods to approximate the original data in lower dimensionality space [Schütze, 1997], and reduce the vectors to 50 dimensions. Having the feature data in a lower dimensionality helps us in comparing them with the fMRI data.

To combine text- and image-based semantic models in a joint representation, we use the unweighted combination function of Section 4.1. We chose a straightforward combination method in order to maintain a lower level of complexity and to better estimate the impact of each modality on the correlation scores.

Finally, to represent the 11 categories we experiment with (see Table 7.2), we average the vectors of the concepts they include.

7.2.4 Experiments

A question is posed over how to evaluate the relationship between the different distributional models and brain data. Comparing each model's predictive performance using the same strategy as Mitchell et al. [2008] (also followed by Murphy et al. [2012]) is one possibility: they used multiple regression to relate distributional codes to individual voxel activations, thus allowing brain states to be estimated from previously unseen distributional codes. Regression models were trained on 58/60 words and in testing the regression models estimated the brain state associated with the 2 unseen distributional codes. The predicted brain states were compared with the actual fMRI data, and the process repeated for each permutation of left-out words, to build a metric of prediction accuracy. For our purposes, a fair comparison of models using this strategy is complicated by differences in dimensionality between both semantic models and lobes (which we compare to other lobes) in association with the comparatively small number of words in the fMRI data set. Large dimensionality models risk overfitting the data, and it is a nuisance to try to reliably correct for the effects of overfitting in performance comparisons. Not least, to thoroughly evaluate all possible cross-validation permutations is demanding in processing time, and we have many models to compare.

An alternative approach, and that which we have adopted, is representational similarity analysis [Kriegeskorte et al., 2008]. Representational similarity analysis circumvents the previous problems by abstracting each fMRI/distributional data source to a common structure capturing the interrelationships between each pair of data items (e.g., words). Specifically, for each model/participant's fMRI data/anatomical region, the similarity structure was evaluated by taking the pairwise correlation (Pearson's correlation coefficient) between all unique category or word combinations. This produced a list of 55 category pair correlations and 121 word pair correlations for each data source. For all brain data, correlation lists were averaged across the nine participants to produce a single list of mean word pair correlations and a single list of mean category pair correlations for each anatomical region and the whole brain. Then to provide a measure of similarity between models and brain data, the correlation lists for respective data sources were themselves correlated using Spearman's rank correlation. Statistical significance was tested using a permutation test: The word-pair (or category-pair) labels were randomly shuffled 10,000 times to estimate a null distribution when the two similarity lists are not correlated. The p-value is calculated as the proportion of random correlation coefficients that are greater than or equal to the observed coefficient.

7.2.5 Results

7.2.5.1 Category-level analyses

Do image models correlate with brain data? Table 2 displays results of Spearman's correlations between the per-category similarity structure of distributional models and brain data. There is a significant correlation between every purely image-based model and the occipital, parietal and temporal lobes, and also the whole brain (.38$\leq \rho \geq$.51, all $p \leq$.01). The frontal lobe is less well described. Still, whilst not significant, correlations are only marginally above the

	Frontal	Parietal	Occipital	Temporal	Whole-Brain
Verb	0.00 (0.51)	0.06 (0.37)	0.24 (0.10)	0.07 (0.35)	0.17 (0.17)
Window2	0.34 (0.06)	0.49 (0.00)	0.47 (0.01)	0.47 (0.00)	0.44 (0.00)
Object	0.27 (0.07)	0.38 (0.02)	0.45 (0.00)	0.47 (0.00)	0.43 (0.01)
Context	0.33 (0.06)	0.50 (0.00)	0.44 (0.00)	0.44 (0.01)	0.44 (0.01)
Object&Context	0.32 (0.05)	0.48 (0.00)	0.51 (0.00)	0.49 (0.00)	0.49 (0.00)
Window2&Object	0.32 (0.06)	0.45 (0.00)	0.52 (0.00)	0.53 (0.00)	0.49 (0.00)
Window2&Context	**0.39 (0.04)**	**0.57 (0.00)**	0.53 (0.00)	**0.55 (0.00)**	0.51 (0.00)
Window2&Object&Context	0.37 (0.04)	0.52 (0.00)	**0.55 (0.00)**	**0.55 (0.00)**	**0.53 (0.00)**

Table 7.3: Matrix of correlations between each pairwise combination of distributional semantic models and brain data. Correlations correspond to the pairwise similarity between the 11 categories. In each column the first value corresponds to Spearman's rank correlation coefficient and the value in parenthesis is the p-value.

	Frontal	Parietal	Occipital	Temporal	Whole-Brain
Verb	-0.04 (0.72)	0.09 (0.06)	0.07 (0.20)	0.03 (0.31)	0.07 (0.18)
Window2	0.07 (0.13)	0.19 (0.00)	0.12 (0.06)	0.21 (0.00)	0.13 (0.04)
Object	0.01 (0.40)	0.08 (0.07)	**0.17 (0.01)**	0.18 (0.00)	0.17 (0.01)
Context	0.04 (0.24)	0.14 (0.01)	0.01 (0.44)	0.12 (0.02)	0.02 (0.38)
Object&Context	0.03 (0.31)	0.13 (0.01)	0.10 (0.07)	0.17 (0.00)	0.11 (0.06)
Window2&Object	0.04 (0.24)	0.16 (0.00)	0.16 (0.01)	**0.23 (0.00)**	**0.17 (0.00)**
Window2&Context	**0.07 (0.12)**	**0.20 (0.00)**	0.09 (0.11)	0.22 (0.00)	0.11 (0.07)
Window2&Object&Context	0.05 (0.18)	0.18 (0.00)	0.12 (0.05)	**0.23 (0.00)**	0.13 (0.02)

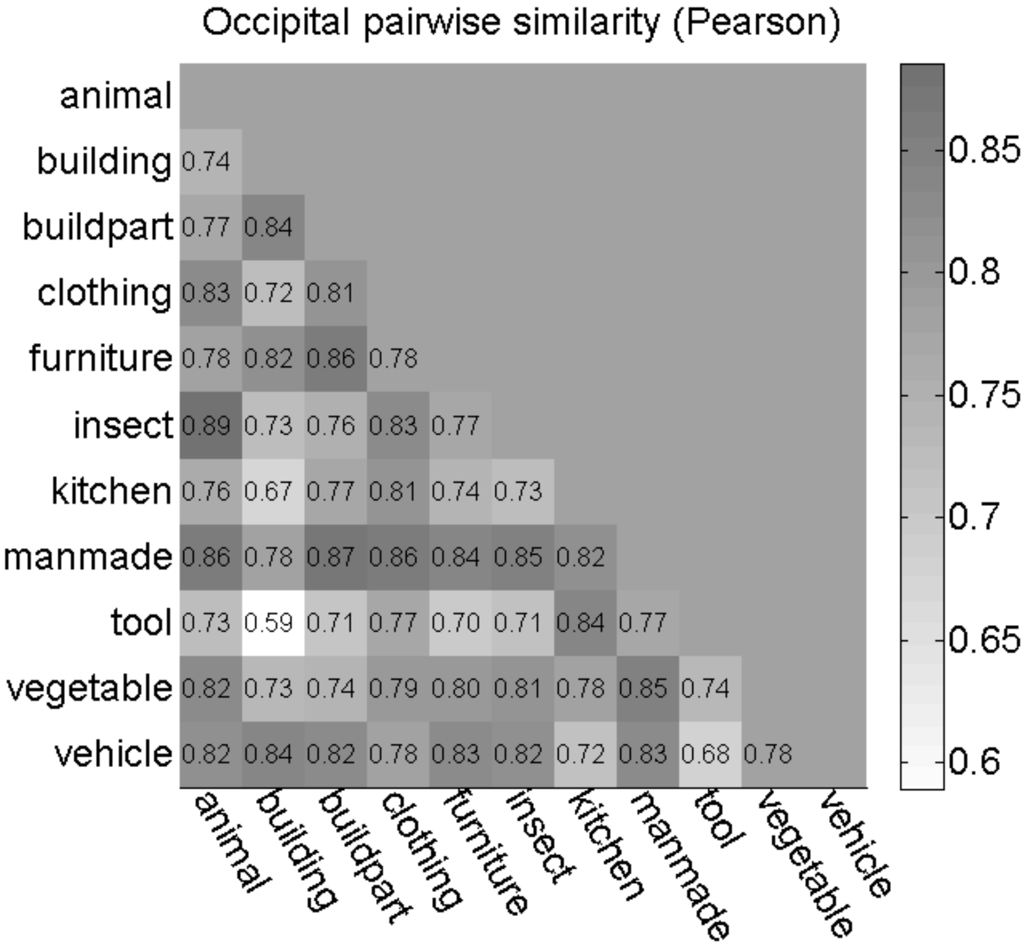
Table 7.4: Matrix of correlations between each pairwise combination of distributional semantic models and brain data. Correlations correspond to the pairwise similarity between the 51 words. In each column the first value corresponds to Spearman's rank correlation coefficient and the value in parenthesis is the p-value.

conventional $p = .05$ cutoff (all are less than $p = .064$). This strongly suggests that the answer to our first question is yes: *distributional models derived from images can be used to explain concept fMRI data.* Otherwise Window2 significantly correlates with the whole brain and all anatomical regions except for the frontal lobe where ρ=.34, $p = .07$. In contrast Verb (the original, partially hand-crafted model used by Mitchell and colleagues) captures inter-relationships poorly and neither correlates with the whole brain or any lobe.

Do different models correlate with different anatomical regions? 2-way ANOVA without replication was used to test for differences in correlation coefficients between the five pure-modality models (Verb, Window2, Object, Context and Object&Context), and the four brain lobes. This revealed a highly significant difference between models F(4,12)=45.2, p<.001. Post-hoc 2-tailed t-tests comparing model pairs found that Verb differed significantly from all other models (correlations were lower). There was a clear difference even when Verb (mean±sd over lobes = .1±.1) was compared to the second weakest model, Object (mean±sd=.4±.09), where t =-7.7, p <.01, df=4. There were no other significant differences between models. However there was a highly significant difference between lobes F(3,12)=13.77, p <.001. Post-hoc 2-tailed t-tests comparing lobe pairs found that the frontal lobe yielded significantly different correlations (lower) than each other lobe. When the frontal lobe (mean±sd over models = .25±.14) was compared to the second weakest anatomical region, the parietal lobe (mean±sd=.38±.19), the difference was highly significant, t =-8, df=3, p <.01. This introduces the question of whether this difference in correlations is the result of differences in neural category organisation and representation, or differences in the quality of the signal, which we address next.

Category-level inter-correlations between lobes were all relatively strong and highly significant. The occipital lobe was found to be the most distinct, being similar to the temporal lobe (ρ=.71, p <.001), but less so to the parietal and frontal lobes (ρ=.53, p <.001 and ρ=.57, p <.001 respectively). The temporal lobe shows roughly similar levels of correlation to each other lobe (all .71$\leq \rho \geq$.73, all p <.001). The frontal and parietal lobes are related most strongly to each other (ρ=.77, p <.001), to a slightly lesser extent to the temporal lobe (in both cases ρ=.73, p <.001) and least so to the occipital lobe. These strong relationships are consistent with there being a broadly similar category organisation across lobes.

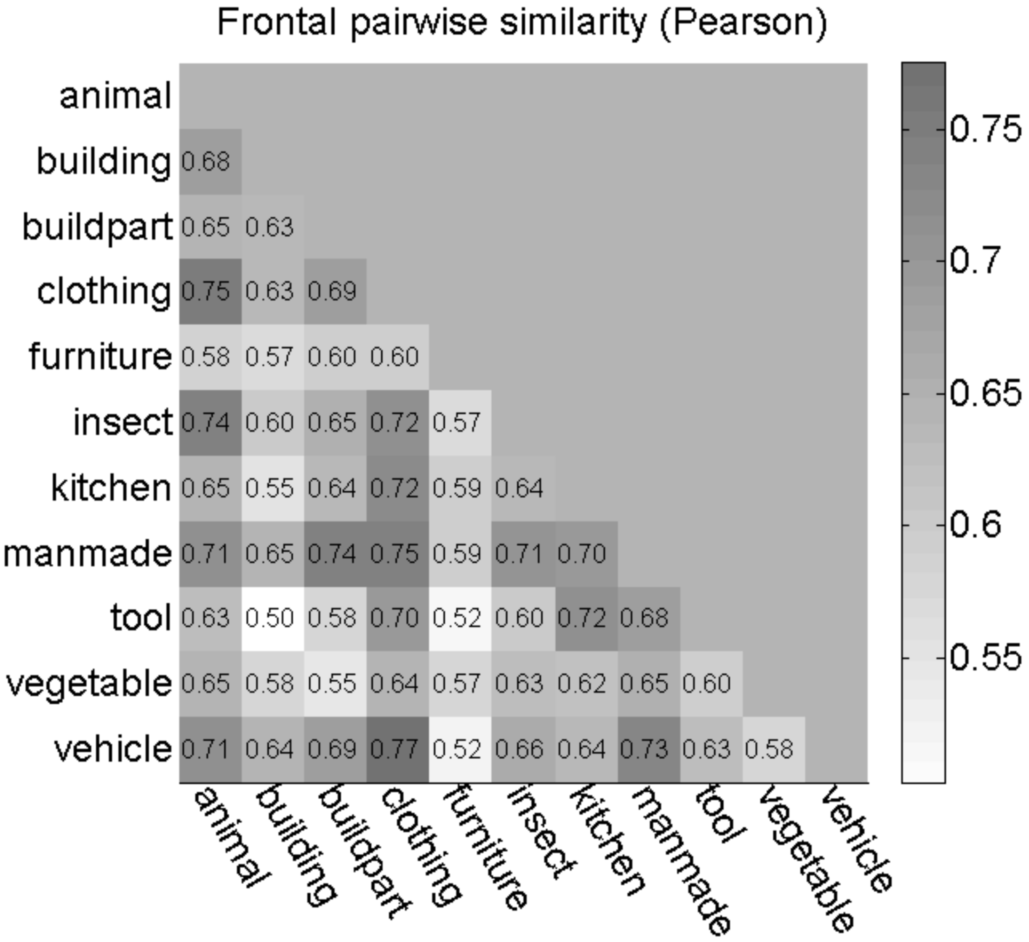
To appraise this assertion in the context of the previously detected difference between the frontal lobe and all other lobes, we examine the raw category pair similarity matrices derived from the occipital lobe and the frontal lobe (Figure 1). All the below observations are qualitative. Although it is difficult to have intuitions about the relative differences between all category pairs (e.g., whether tools or furniture should be more similar to animals), we might reasonably expect some obvious similarities. For instance, for animals to be visually similar to insects and clothing, because all have legs and arms and curves (of course we would not expect a strong relationship between insects and clothes in function or other modalities such as sound), buildings to be similar to building parts and



	animal	building	buildpart	clothing	furniture	insect	kitchen	manmade	tool	vegetable	vehicle
animal											
building	0.74										
buildpart	0.77	0.84									
clothing	0.83	0.72	0.81								
furniture	0.78	0.82	0.86	0.78							
insect	0.89	0.73	0.76	0.83	0.77						
kitchen	0.76	0.67	0.77	0.81	0.74	0.73					
manmade	0.86	0.78	0.87	0.86	0.84	0.85	0.82				
tool	0.73	0.59	0.71	0.77	0.70	0.71	0.84	0.77			
vegetable	0.82	0.73	0.74	0.79	0.80	0.81	0.78	0.85	0.74		
vehicle	0.82	0.84	0.82	0.78	0.83	0.82	0.72	0.83	0.68	0.78	

Figure 7.6: Similarity (Pearson correlation) between each category pair in occipital lobe.

vehicles (hard edges and windows), building parts to be similar to furniture (e.g., from Table 7.2 we see there is some overlap in category membership between these categories, such as closet and door) and tools to be similar to kitchen utensils. All of these relationships are maintained in the occipital lobe, and many are visible in the frontal lobe (including the similarity between insects and clothes), however there are exceptions that are difficult to explain e.g., within the frontal lobe, building parts are not similar to furniture, kitchen utensils are closer to clothing than to tools and vehicles are more similar to clothing than anything else. As such we conclude that *category-level representations were similar across lobes* with differences likely due to variation in signal quality between lobes.



	animal	building	buildpart	clothing	furniture	insect	kitchen	manmade	tool	vegetable	vehicle
animal											
building	0.68										
buildpart	0.65	0.63									
clothing	0.75	0.63	0.69								
furniture	0.58	0.57	0.60	0.60							
insect	0.74	0.60	0.65	0.72	0.57						
kitchen	0.65	0.55	0.64	0.72	0.59	0.64					
manmade	0.71	0.65	0.74	0.75	0.59	0.71	0.70				
tool	0.63	0.50	0.58	0.70	0.52	0.60	0.72	0.68			
vegetable	0.65	0.58	0.55	0.64	0.57	0.63	0.62	0.65	0.60		
vehicle	0.71	0.64	0.69	0.77	0.52	0.66	0.64	0.73	0.63	0.58	

Figure 7.7: Similarity (Pearson correlation) between each category pair in frontal lobe.

Are text- and image-based semantic models complementary? Turning to the question of whether text- and image-derived semantic information can be complementary, we observe from Table 2 that there is not a single instance of a joint model with a weaker correlation than its pure-image counterpart. The Window2 model showed a stronger correlation than the Window2&Object model for the frontal and parietal lobes, but was weaker than Window2&Object&Context and Window2&Context in all tests and was also weaker than any joint model in whole-brain comparisons. The mean±sd correlations for all purely image-based results pooled over lobes (3 models * 4 lobes) was .42±.08 in comparison to .49±.08 for the joint models. The relative performance of Object vs. Context vs. Object&Context on the four different lobes is preserved between image-based and joint models: correlating the 12 combinations using Spearman's correlation gives ρ=.85, p <.001. Differences can be statistically quantified by pooling all image related correlation coefficients for each anatomical region (3 models * 4 regions), as for the respective joint models, and comparing with a 2-tailed Wilcoxon signed rank test. Differences were highly significant (W=0, p <.001,n=12). This evidence accumulates to suggest that *text and image-derived semantic information can be complementary* in interpreting concept fMRI data.

7.2.5.2 Word-level analyses

Do image models capture word pair similarities? Per-word results generally corroborate the relationships observed in the previous section in the sense that Spearman's correlation between per-word and per-category results for the 40 combinations of models and lobes was ρ=.78, p <.001. There were differences, most obviously a dramatic drop in the strength of correlation coefficients for the per-word results, visible in Table 3. *Subsets of per-word image-based models correlated with three lobes and the whole brain.* Correlations corresponding to significance values of p <.05 were observed in the temporal and parietal lobes, for Context, Object&Context and Window2 whereas Object was correlated with the occipital and temporal lobes (p <.05). 2-way ANOVA without replication was used to test for differences between models and lobes. This revealed a significant difference between models (F(4,12)=4.05, p=.027). Post-hoc t-tests showed that the Window2 model significantly differed from (was stronger than) the Context (t=3.8, p =.03, df=3) and Object&Context models (t =4.5, p =.02, df=3). There were no other significant differences between models. There was again a significant difference between lobes (F(3,12)=7.89, p < .01), with the frontal lobe showing the weakest correlations. Post-hoc 2-tailed t-tests comparing lobe-pairs found that the frontal lobe differed significantly (correlations were weaker) from the parietal (t =-9, p <.001, df=4) and temporal lobes (t =-6.4, p <.01, df=4) but not from the occipital lobe (t =-2.18, p =.09, df=4). No other significant differences between lobes were observed.

Are there differences between models/lobes? Word-level inter-correlations between lobes were all significant and the pattern of differences in correlation strength largely resembled that of the category-level analyses. The occipital lobe was again most similar to the temporal lobe (ρ=.57, p <.001), but less so to the parietal and frontal lobes (ρ=.47, p <.001 and ρ=.34, p <.001 respectively). The temporal lobe this time showed stronger correlation to the parietal (ρ=.68, p <.001) and frontal lobes (ρ=.61, p <.001) than the occipital lobe. The frontal and parietal lobes were again strongly related to one another (ρ=.67, p <.001). These results echo the category-level findings, that *word-level brain activity is also organised in a similar way across lobes.* Consequently this diminishes our chances of uncovering neat interactions between models and brain areas (where for instance the Window2 model correlates with the frontal lobe and Object model matches the occipital lobe). It is however noteworthy that we can observe some interpretable selectivity in lobe*model combinations. In particular the Context model better matches the parietal lobe than the Object model, which in turn better captures the occipital and temporal lobes (Observations are qualitative). Also as we see next, adding text information boosts performance in both parietal and temporal lobes (see Section 7.2.2 on our expectations about information encoded in the lobes).

Does joining text and image models help word-level interpretation? As concerns the benefits of joining Text and Image information, *per-word joint models were generally stronger than the respective image-based models.* There was one exception: adding text to the Object model weakened correlation with the occipital lobe. Joint models were exclusively stronger than Window2 for the temporal and occipital lobes, and were stronger in 1/3 of cases for the frontal and parietal lobes. In an analogous comparison to the per-category analysis, a Wilcoxon signed rank test was used to examine the difference made by adding text information to image models (pooling 3 models over 4 anatomical areas for both image and joint models). The mean±sd of image models was .1±.06 whereas for Joint models it was .15±.07. The difference was highly significant (W=1, p <.001, n=12).

7.2.6 Discussion

This study brought together, for the first time, two recent research lines: The exploration of "semantic spaces" in the brain using distributional semantic models extracted from corpora, and the extension of the latter to image-based features. We showed that image-based distributional semantic measures significantly correlate with fMRI-based neural similarity patterns pertaining to categories of concrete concepts as well as concrete basic-level concepts expressed by specific words (although correlations, especially at the basic-concept level, are rather low, which might signify the need to develop still more performant distributional models and/or noise inherent to neural data). Moreover, image-based models complement a state-of-the-art text-based model, with the best performance achieved when the two modalities are combined. This not only presents an optimistic outlook for the future use of image-based models as an interpretative tool to explore issues of cognitive grounding, but also demonstrates that they are capturing useful additional aspects of meaning to the text models, which are likely relevant for computational semantic tasks.

The weak comparative performance of the original Mitchell et al.'s Verb model is perhaps surprising given its previous success in prediction [Mitchell et al., 2008], but a useful reminder that a good predictor does not necessarily have to capture the internal structure of the data it predicts.

The lack of finding organisational differences between anatomical regions differentially described by the various models seems to exclude localization as a major actor in modeling visual semantics. We have indeed no statistical support for differential performance of the different models in correlating with different brain areas. One possible reason for this might be ascribed to the strong correlation in representational similarity between different brain regions, which is evidence in line with a distributed neural representation of concepts. On the other hand, it is worth noticing that the Mitchell dataset was not originally designed to tease apart visual information from linguistic context. So our different models (object, context, text) are correlated for this dataset, and therefore are not ideal for segregating different aspects (e.g., modalities) of neural representation. Specifically animals are likely to appear in similar contexts as are tools as are kitchen, utensils, etc., which is the motivation for designing new stimulus sets. In future experiments it may prove valuable to configure a fMRI stimulus set where text-based and image-based interrelationships are maximally different. Collecting our own fMRI data will also allow us to move beyond exploratory analysis, to test sharper predictions about distributional models and their brain area correlates. There are also many opportunities for focusing analyses on different subsets of brain regions, with the semantic system identified by Binder et al. [2009] in particular presenting one interesting avenue for investigation.

Chapter 8

Conclusions

In this thesis I have provided an extensive introduction to a new approach to distributional semantics that we named Multimodal Distributional Semantics. A multimodal distributional semantic model integrates a traditional text-based representation of meaning with information coming from vision. In this way, it tries to answer the critique that distributional models lack grounding, since they base their representation of meaning entirely on the linguistic input, forgetting that humans have also access to rich sources of perceptual knowledge. Of course, a truly multimodal representation of meaning should account for the entire spectrum of human senses. On the other hand, this line of research is still in its embryonic stage and there is still a shortage of both perceptual data available and techniques to automatize their processing. This is why, in this thesis, we focused our analysis on the visual perceptual channel, for which we have at our disposal both large data sets and effective methods to analyze them.

In particular, we exploited the ESP-Game and the ImageNet datasets, where the image documents are tagged with words describing their content. To harvest visual information we adopted the bag-of-visual-words technique, which discretizes image content in ways that are analogous to standard text-based distributional representations.

To merge the textual and the visual information, we have proposed a multimodal framework organized towards increasingly sophisticated fusion methods. We proposed a first unweighted combination approach which has the advantage of being computationally light. A second approach introduces a weighted combination scheme and involves a dimensionality reduction step to induce the creation of new connections between the two sources of information. Finally, we described a first attempt to translate fusion from a global to a local process, in which each word is augmented by the visual channel according to its (visual) perceptual saliency.

We conducted a number of experiments to assess the quality of the obtained models. We first investigated the general semantic properties of a purely image-based model, to assess its overall quality as well as to look for information complementary to that present in text. We found systematic differences between the two modalities, such as the preference for encyclopedic properties of a text-based model and for perceptual properties in the case of the image-based model. We proceeded to test a selection of models obtained by the combination of the text- and image-based representations via our multimodal framework. We used two benchmarks for word relatedness and one benchmark for word categorization and in both cases we obtained a systematic improvement in performance with the multimodal models compared to models based on standalone channels.

In a second round of experiments, we tested the multimodal framework on two tasks which make vision explicitly relevant, namely two tasks involving color information. We showed how, in this case, semantic models based on visual features can complement or even outperform distributional semantic models based on text. In particular, we showed that even on an apparently trivial task such as discovering the color of an object, a purely textual model dramatically fails, while a relatively simple visual model achieves a significant performance. Also in a second task focused on processing both literal and metaphorical usages of color adjectives, the visual information induced a strong gain in performance when combined with text.

Finally, we investigated the role of object localization in the image-based semantic model construction. We show that it is possible to extract better image-based semantic vectors by first localizing the objects denoted by words and then extracting visual information from the object location and from its surround independently. We tested object localization on two tasks. First, on word relatedness. Interestingly, here we discovered that image-based semantic vectors extracted from the object surround are more effective than those based on the object location when tested on our word relatedness task. For example, the fact that pictures containing deers and wolves depict similar surrounds tells us that such creatures live in similar environments, and it is thus likely that they are somewhat related. This can be seen as the distributional hypothesis transposed to images: objects that are semantically similar occur in similar visual contexts. Nevertheless, the work has to be considered a proof of concept, since we experimented with 20 words only. In future studies we will test a larger number of words. In a second test, we tried to asses the impact of object localization on reproducing the semantic patterns captured by fMRI recordings where we compared the resulting visual models and a text-based model. Although we found an overall significant correlation between the multimodal models and the fMRI data, we couldn't confirm the intriguing hypothesis according to which different lobes of the brain process information more or less relevant with respect to the visual context or object. We don't exclude that the lack of results might be due the particular dataset we used, since it was not specifically designed to disentagle the object from its context.

There are several future directions that could be explored. First of all, there is big room to improve each of the steps of the multimodal framework. For example, the bag-of-visual-words pipeline could be replaced with more advanced representations such as the Fisher encoding; more recent techniques than SVD such as Stack Auto-Encoders could be used to merge vision and text in a single representation [Hinton and Salakhutdinov, 2006]; localization should be tested on a larger set of objects; a new stimulus set for fMRI recordings should be designed specifically for localization.

Another interesting direction would consist in developing new evaluation methods for multimodal distributional semantics. It is indeed clear that some of the tasks we used in this project have to be considered only partially relevant, since they were not designed with the specific goal of taking advantage of multimodality. Now that distributional semantics is gradually evolving into a multichannel system, we should try to encourage the different channels communicating and interacting with each other. Moreover, new tasks should involve the use of the multimodal features for more grounded tasks such as parsing navigation directions or mapping instructions to actions.

To conclude, we hope that the efforts to create the multimodal framework and to support it through a set of empirical evidences will bring a valid contribution to the fascinating line of research that is the development of human-like models of meaning.

Chapter 9

Appendix A

9.1 VSEM: An open library for visual semantics representation

In the last years we have seen great progress in the area of automated image analysis. Important advances, such as the introduction of local features for a robust description of the image content (see Mikolajczyk and Schmid [2005] for a systematic review) and the bag-of-visual-words method (**BoVW**)[1] for a standard representation across multiple images [Sivic and Zisserman, 2003], have contributed to make image analysis ubiquitous, with applications ranging from robotics to biology, from medicine to photography.

In addition, the development of these key ideas has been strongly accelerated by both the introduction of very well defined challenges such as the popular object recognition task Pascal Visual Object Classes Challenge [Everingham et al., 2010], which have been attracting also a wide community of "outsiders" specialized in a variety of disciplines (e.g., machine learning, neural networks, graphical models

[1] Bag-of-visual-words model is a popular technique for image classification inspired by the traditional bag-of-words model in Information Retrieval. It represents an image with a histogram of frequency of **visual words**. Visual words are identified by clustering a large corpus of training features. See Szeliski [2010] chapter 14 for more details.

and natural language processing), and by sharing effective, well documented implementations of cutting edge image analysis algorithms, such as OpenCV[1] and VLFeat.[2]

A comparable story can be told about automatic text analysis. The last decades have seen a long series of successes in the processing of large text corpora in order to extract more or less structured semantic knowledge. In particular, under the assumption that meaning can be captured by patterns of co-occurrences of words, distributional semantic models such as Latent Semantic Analysis [Landauer and Dumais, 1997] or Topic Models [Blei et al., 2003] have been shown to be very effective both in general semantic tasks such as approximating human intuitions about meaning, as well as in more application-driven tasks such as information retrieval, word disambiguation and query expansion [Turney and Pantel, 2010]. And also in the case of automated text analysis, a wide range of method implementations are at the disposal of the scientific community.[3]

Nowadays, given the parallel success of the two disciplines, there is growing interest in making the visual and textual channels interact for mutual benefit. If we look at the image analysis community, we discover that a surprisingly well established tradition of studies that exploit both channels of information have already been established. For example, there is a relatively extended amount of literature about enhancing the performance on visual tasks such as object recognition or image retrieval by replacing a purely image-based pipeline with hybrid methods augmented with textual information [Barnard et al., 2003; Berg et al., 2010; Farhadi et al., 2009, 2010; Kulkarni et al., 2011].

Unfortunately, the situation within automatic text analysis is not so rosy. Despite the huge potential that automatically induced visual features could represent as a new source of perceptually grounded semantic knowledge,[4] image-enhanced

[1] http://opencv.org/

[2] http://www.vlfeat.org/

[3] See for example the annotated list of corpus-based computational linguistics resources at http://www-nlp.stanford.edu/links/statnlp.html.

[4] Many studies have showed how semantic models derived from standard text corpora capture encyclopedic, functional and discourse-related properties of word meanings, but miss their concrete aspects [Andrews et al., 2009; Baroni and Lenci, 2008; Baroni et al., 2010; Riordan and Jones, 2011]. For example, we might gather from text the information that bananas are tropical and edible, but not that they are yellow (because it is highly unlikely that an author feels the need to write down into words that "bananas are yellow"). The same studies show

models of semantics developed so far [Bergsma and Goebel, 2011; Bruni et al., 2011, 2012a,b; Feng and Lapata, 2010; Leong and Mihalcea, 2011] have only scratched this great potential and are still considered as proof-of-concept studies only.

One possible reason of this delay with respect to the image analysis community might be ascribed to the high entry barriers that NLP researchers adopting image analysis methods have to face. Although many of the image analysis toolkits are open source and well documented, they mainly address users within the same community and therefore their use is not as intuitive for others. The final goal of libraries such VLFeat and OpenCV is the representation and classification of images. Therefore, they naturally lack of a series of complementary functionalities that are necessary to bring the visual representation to the level of semantic concepts.

For these reasons, we present here VSEM, a novel toolkit which makes the extraction of image-based representations of concepts an easy task. VSEM is equipped with state-of-the-art algorithms, from low-level feature detection and description up to the BoVW representation of images, together with a set of new routines necessary to move from an image-wise to a concept-wise representation of the image content. In a nutshell, VSEM extracts visual information in an analogous way to how it is done for automatic text analysis. Thanks to BoVW, the image content is indeed discretized and visual units somehow comparable to words in text are produced (the visual words). In this way, from a corpus of images annotated with a set of concepts, it is possible to derive semantic vectors of co-occurrence counts of concepts and visual words. The obtained semantic vectors can be then utilized for typical semantic tasks, such as approximating the semantic relatedness of two concepts by a similarity function over the visual words representing them.

Importantly, the part of VSEM concerning image analysis is based on VLFeat functionalities, offering enhanced VSEM image analysis pipeline makes use of VLFeat functionalities [Vedaldi and Fulkerson, 2010]. This guarantees that the image analysis underpinnings of the library be well maintained and state-of-the-

how, when humans are asked to describe concepts, the features they produce are instead mainly of a perceptual nature (bananas are yellow, tigers have stripes, and so on).

art.

The rest of this section is organized as follows. In Section 9.1.1 we introduce the procedure to obtain an image-based representation of a concept. Section 9.1.2 describes the VSEM structure. Section 9.1.3 shows how to install and run VSEM through an example that uses the Pascal VOC data set.

9.1.1 VSEM pipeline

More in detail, the pipeline encapsulating the whole process mentioned above takes as input a collection of images together with their associated tags and optionally object location annotations. Its output is a set of concept representation vectors for individual tags. The following steps are involved: (i) extraction of local image features, (ii) visual vocabulary construction, (iii) encoding the local features in a BoVW histogram, (iv) including spatial information with spatial binning, (v) aggregation of visual words on a per-concept basis in order to obtain the co-occurrence counts for each concept and (vi) transforming the counts into association scores and/or reducing the dimensionality of the data. A brief description of the individual steps follows.

Local features Local features are designed to find local image structures in a repeatable fashion and to represent them in robust ways that are invariant to typical image transformations, such as translation, rotation, scaling, and affine deformation. Local features constitute the basis of approaches developed to automatically recognize specific objects [Grauman and Leibe, 2011]. The most popular local feature extraction method is the Scale Invariant Feature Transform (SIFT), introduced by Lowe [2004]. VSEM uses the VLFeat implementation of SIFT.

Visual vocabulary To obtain a BoVW representation of the image content, a large set of local features extracted from a large corpus of images are clustered. In this way the local feature space is divided into informative regions (*visual words*) and the collection of the obtained visual words is called *visual vocabulary*. k-means is the most commonly used clustering algorithm [Grauman and Leibe, 2011]. In the special case of Fisher encoding (see below), the clustering of the features is performed with a *Gaussian mixture model* (GMM), see Perronnin et al. [2010]. Figure 9.1 exemplifies a visual vocabulary construction pipeline. VSEM contains both the k-means and the GMM implementations.

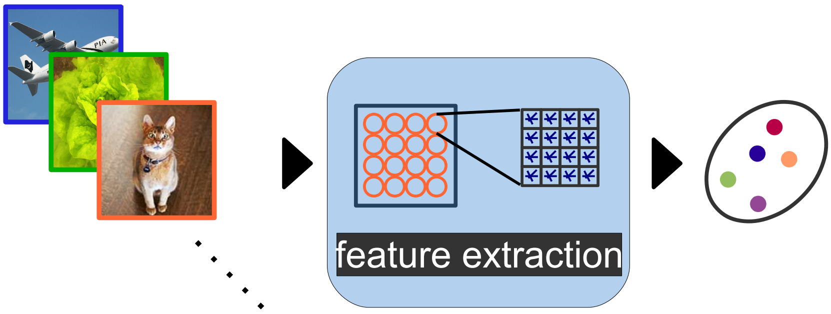


Figure 9.1: An example of a visual vocabulary creation pipeline. From a set of images, a larger set of features are extracted and clustered, forming the visual vocabulary.

Encoding The encoding step maps the local features extracted from an image to the corresponding visual words of the previously created vocabulary. The most common encoding strategy is called *hard quantization*, which assigns each feature to the nearest visual word's centroid (in Euclidean distance). Recently, more effective encoding methods have been introduced, among which the Fisher encoding [Perronnin et al., 2010] has been shown to outperform all the others [Chatfield et al., 2011]. VSEM uses both the hard quantization and the Fisher encoding.

Spatial binning A consolidated way of introducing spatial information in BoVW is the use of spatial histograms [Lazebnik et al., 2006]. The main idea is to divide the image into several (spatial) regions, compute the encoding for each region and stack the resulting histograms. This technique is referred to as *spatial binning* and it is implemented in VSEM. Figure 9.2 exemplifies the BoVW pipeline for a single image, involving local features extraction, encoding and spatial binning.

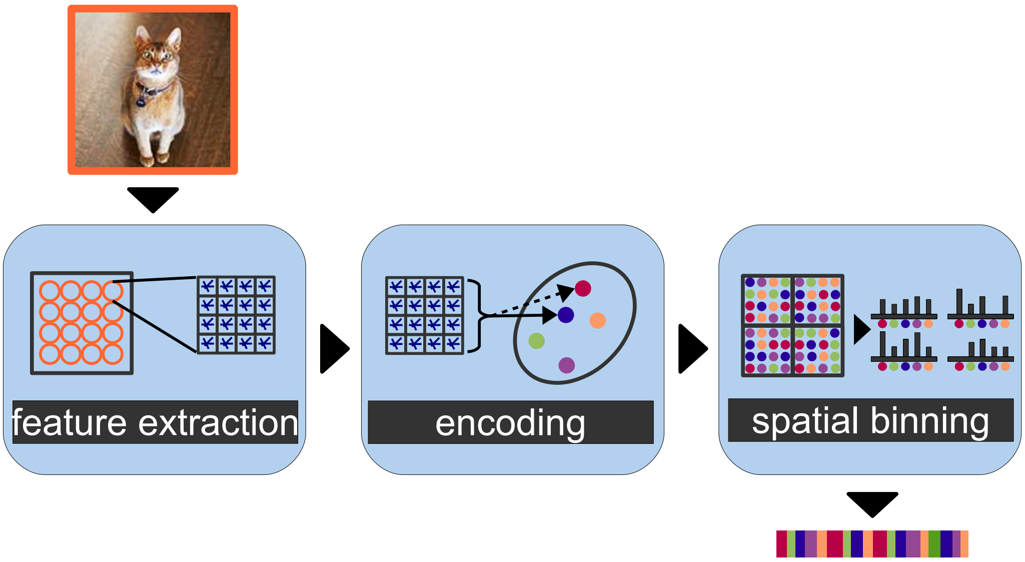


Figure 9.2: An example of a BoVW representation pipeline for an image. Figure inspired by Chatfield et al. [2011]. Each feature extracted from the target image is assigned to the corresponding visual word(s). Then, spatial binning is performed.

Moreover, the input of spatial binning can be further refined by introducing localization. Three different types of localization are typically used: global, object, and surrounding. Global extracts visual information from the whole image and it is also the default option when the localization information is missing. Object extracts visual information from the object location only and the surrounding extracts visual information from outside the object location. Localization itself can either be done by humans (or ground truth annotation) but also by existing localization methods [Uijlings et al., 2013].

For localization, VSEM uses annotated object locations (in the format of bounding boxes) of the target object.

Aggregation Since each concept is represented by multiple images, an aggregation function for pooling the visual word occurrences across images has to be defined. As far as we know, the sum function has been the only function utilized so far. An example for the aggregation step is sketched in figure 9.3. VSEM offers an implementation of the sum function.

Transformations Once the concept-representing visual vectors are built, two types of transformation can be performed over them to refine their raw visual word counts: *association scores* and *dimensionality reduction*. So far, the vectors that we have obtained represent co-occurrence counts of visual words with concepts. The goal of association scores is to distinguish interesting co-occurrences from those that are due to chance. In order to do this, VSEM implements two versions of mutual information (pointwise and local), see Evert [2005].

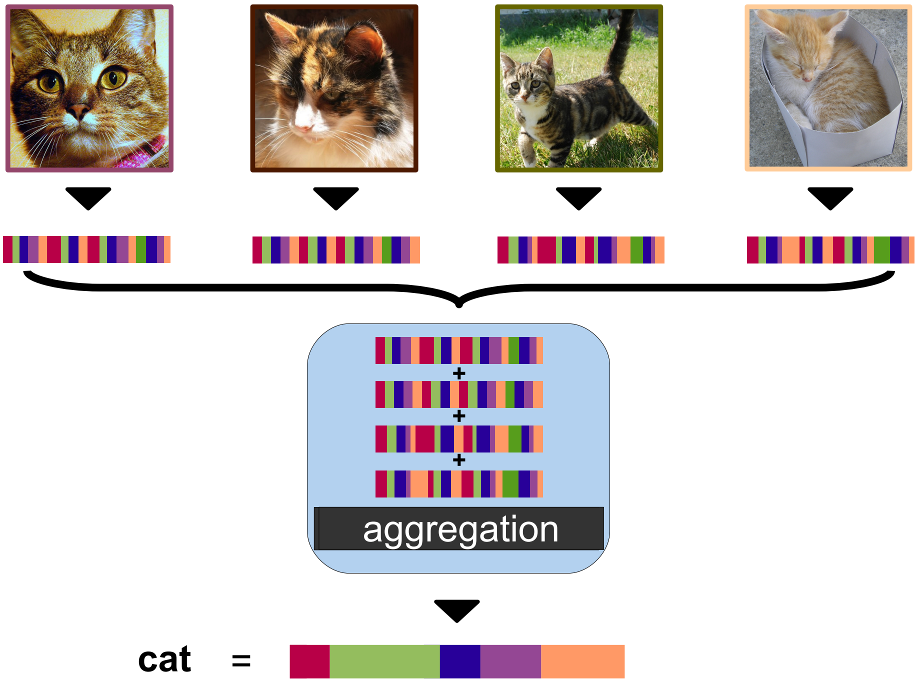


Figure 9.3: An example of a concept representation pipeline for **cat**. First, several images depicting a cat are represented as vectors of visual word counts and, second, the vectors are aggregated into one single concept vector.

On the other hand, dimensionality reduction leads to matrices that are smaller and easier to work with. Moreover, some techniques are able to smooth the matrices and uncover latent dimensions. Common dimensionality reduction methods are singular value decomposition [Manning et al., 2008], non-negative matrix factorization [Lee and Seung, 2001] and neural networks [Hinton and Salakhutdinov, 2006]. VSEM implements the singular value decomposition method.

9.1.2 Framework design

VSEM offers a friendly implementation of the pipeline described in Section 9.1.1. The framework is organized into five parts, which correspond to an equal number of MATLAB packages and it is written in object-oriented programming to encourage reusability. A description of the packages follows.

- `datasets` This package contains the code that manages the image data sets. As an example of how to write a wrapper for a given data set, the Pascal VOC wrapper is implemented (`PascalDataset`). To use a new image data set two solutions are possible: either write a new class which extends `GenericDataset` or use directly `PascalDataset` after having rearranged the new data as described in **`help`** `PascalDataset`.

- `vision` This package contains the code for extracting the bag-of-visual-words representation of images. In the majority of cases, it can be used as a "black box" by the user. Nevertheless, if the user wants to add new functionalities such as new features or encodings, this is possible by simply extending the corresponding generic classes.

- `concepts` This is the package that deals with the construction of the image-based representation of concepts. `concepts` it the most important package of VSEM. It applies the image analysis methods to obtain the BoVW representation of the image data and then aggregates visual word counts concept-wise. The transformations utilities are contained in its `helpers` sub-package.

- `benchmarks` This package contains the code for benchmarking. Currently there are two benchmarks implemented: a semantic similarity benchmark and a semantic categorization benchmark . The two benchmarking classes are CategorizationBenchmark and SimilarityBenchmark. They both extend the abstract class GenericBenchmark.

- helpers This package contains supporting functions and classes. There is a general helpers with functionalities shared across packages and several package specific helpers.

9.1.3 Getting started

Installation VSEM can be easily installed by running the file vsemSetup.m. Moreover, pascalDatasetSetup.m can be run to download and place the popular dataset, integrating it in the current pipeline.

Documentation All the MATLAB commands of VSEM are self documented (e.g. help vsem) and an HTML version of the MATLAB command documentation is available from the VSEM website.

The Pascal VOC demo The Pascal VOC demo provides a comprehensive example of the workings of VSEM. From the demo file pascalVQDemo.mmultiple configurations are accessible. Additional settings are available and documented for each function, class or package in the toolbox (see Documentation).

Running the demo file executes the following lines of code and returns as output ConceptSpace, which contains the visual concept representations for the Pascal data set.

```
% Create a matlab structure with the
% whole set of images in the Pascal
% dataset along with their annotation
dataset = datasets.VsemDataset(configuration.imagesPath,'
   annotationFolder',configuration.annotationPath);


% Initiate the class that handles
% the extraction of visual features.
featureExtractor = vision.features.PhowFeatureExtractor();

% Create the visual vocabulary
vocabulary = KmeansVocabulary.trainVocabulary(dataset,
   featureExtractor);

% Calculate semantic vectors
conceptSpace = conceptExtractor.extractConcepts(dataset,
   histogramExtractor);

% Compute pointwise mutual
% information
conceptSpace = conceptSpace.reweight();

% Conclude the demo, computing
% the similarity of correlation
% measures of the 190 possible
% pair of concepts from the Pascal
% dataset against a gold standard
[correlationScore, p-value] = similarityBenchmark.
   computeBenchmark(conceptSpace,similarityExtractor);
```

References

Hervé Abdi and Lynne Williams. Newman-Keuls and Tukey test. In Neil Salkind, Bruce Frey, and Dondald Dougherty, editors, *Encyclopedia of Research Design*, pages 897–904. Sage, Thousand Oaks, CA, 2010. 51

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasça, and Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of HLT-NAACL*, pages 19–27, Boulder, CO, 2009. 57

Abdulrahman Almuhareb. *Attributes in Lexical Acquisition*. Phd thesis, University of Essex, 2006. 10

Abdulrahman Almuhareb and Massimo Poesio. Concept learning and categorization from the web. In *Proceedings of CogSci*, pages 103–108, Stresa, Italy, 2005. 11, 63

Mark Andrews, Gabriella Vigliocco, and David Vinson. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498, 2009. 3, 4, 14, 53, 118

Harald Baayen. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge, UK, 2008. 60

Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David Blei, and Michael Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003. 16, 118

Marco Baroni and Alessandro Lenci. Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88, 2008. 3, 53, 78, 80, 118

Marco Baroni and Alessandro Lenci. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010. 11, 13, 76

Marco Baroni and Alessandro Lenci. How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP GEMS Workshop*, pages 1–10, Edinburgh, UK, 2011. 50, 51

Marco Baroni, Eduard Barbu, Brian Murphy, and Massimo Poesio. Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, 34(2):222–254, 2010. 3, 11, 53, 63, 118

Lawrence Barsalou. Grounded cognition. *Annual Review of Psychology*, 59:617–645, 2008. 3

PR Beaudet. Rotationally invariant image operators. In *Proceedings of the International Joint Conference on Pattern Recognition*, pages 579–583, 1978. 20

Tamara Berg, Alexander Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy Web data. In *ECCV*, pages 663–676, Crete, Greece, 2010. 16, 118

Shane Bergsma and Randy Goebel. Using visual information to predict lexical preference. In *Proceedings of RANLP*, pages 399–405, Hissar, Bulgaria, 2011. 14, 119

Brent Berlin and Paul Key. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA, 1969. 80

Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortexl*, 12:2767–2796, 2009. 98, 99, 110

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 10, 118

Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. In *Proceedings of ICCV*, pages 1–8, Rio de Janeiro, Brazil, 2007. 25

Anna Bosch, Andrew Zisserman, and Xavier Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4), 2008. 32

Vicki Bruce, Patrick R Green, and Georgeson Mark A. *Visual perception: Physiology, psychology, and ecology*. Psychology Pr, 2003. 98

Elia Bruni, Giang Binh Tran, and Marco Baroni. Distributional semantics from text and images. In *Proceedings of the EMNLP GEMS Workshop*, pages 22–32, Edinburgh, UK, 2011. 13, 58, 119

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. Distributional semantics in Technicolor. In *Proceedings of ACL*, pages 136–145, Jeju Island, Korea, 2012a. 14, 55, 119

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of ACM Multimedia*, pages 1219–1228, Nara, Japan, 2012b. 35, 119

Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006. 11, 53

John Bullinaria and Joseph Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526, 2007. 36, 44

John Bullinaria and Joseph Levy. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, 44:890–907, 2012. 36

Curt Burgess. Theory and operational definitions in computational memory models: A response to Glenberg and Robertson. *Journal of Memory and Language*, 43(3):402–408, 2000. 2

Juan Caicedo, Jaafar Ben-Abdallah, Fabi González, and Olfa Nasraoui. Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. *Neurocomputing*, 76(1):50–60, 2012. 38

John Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell*, 36(4):679–698, 1986. 77

Kai-min Chang, Tom Mitchell, and Marcel Just. Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation. *NeuroImage*, 56:716–727, 2011. 95

Linda L Chao, James V Haxby, and Alex Martin. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature neuroscience*, 2(10):913–919, 1999. 98

Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of BMVC*, Dundee, UK, 2011. xiv, 29, 33, 121, 122

David L. Chen and Raymond J. Mooney. Learning to interpret natural language navigation instructions from observations. In *Proceedings of AAAI*, pages 859–865, San Francisco, CA, 2011. 16

Kenneth Church and Peter Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. 45, 101

Stephen Clark. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics, 2nd edition*. Blackwell, Malden, MA, 2013. In press. 1, 10, 94

Max Coltheart. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33, 1981. 46

Andrew Connolly, Lila Gleitman, and Sharon Thompson-Schill. Effect of congenital blindness on the semantic representation of some everyday concepts. *Proceedings of the National Academy of Sciences*, 104(20):8241–8246, 2007. 3

Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, Prague, Czech Republic, 2004. 25, 87

James Curran and Marc Moens. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, pages 59–66, Philadelphia, PA, 2002. 10

Manuel de Vega, Arthur Glenberg, and Arthur Graesser, editors. *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford University Press, Oxford, UK, 2008. 2

Jia Deng, Wei Dong, Richard Socher, Lia-Ji Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255, Miami Beach, FL, 2009. 13, 100

Russell A Epstein. Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in cognitive sciences*, 12(10):388–396, 2008. 98

Katrin Erk. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653, 2012. 10, 94

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 87, 88, 117

Stefan Evert. *The Statistics of Word Cooccurrences.* Dissertation, Stuttgart University, 2005. 9, 36, 88, 123

Mark D. Fairchild. Status of cie color appearance models, 2005. 77

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of CVPR*, pages 1778–1785, Miami Beach, FL, 2009. 15, 16, 118

Ali Farhadi, Mohsen Hejrati, Mohammad A. Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of ECCV*, Crete, Greece, 2010. 16, 118

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998. 100

Pedro Felzenszwalb, Ross Girshick, David McAllester, and Deva Deva Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010. 86, 88

Yansong Feng and Mirella Lapata. Visual information in semantic representation. In *Proceedings of HLT-NAACL*, pages 91–99, Los Angeles, CA, 2010. xv, 12, 13, 58, 59, 119

Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. *Advances in Neural Information Processing Systems*, 2007. 15

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002. 11

John R. Firth. *Papers in Linguistics, 1934-1951*. Oxford University Press, Oxford, UK, 1957. 1

Arthur Glenberg and David Robertson. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 3(43):379–401, 2000. 2

Melvyn A. Goodale and David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15:20–25, 1992. 98

Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Proceedings of ICCV*, pages 1458–1465, Beijing, China, 2005. 33

Kristen Grauman and Bastian Leibe. *Visual Object Recognition*. Morgan & Claypool, San Francisco, 2011. 20, 21, 26, 33, 120

Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer, Boston, MA, 1994. 10

Tom Griffiths, Mark Steyvers, and Josh Tenenbaum. Topics in semantic representation. *Psychological Review*, 114:211–244, 2007. 4, 10, 12

Peter Hagoort. On Broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9):416–423, 2005. 99

Thorsten Hansen, Maria Olkkonen, Sebastian Walter, and Karl Gegenfurtner. Memory modulates color appearance. *Nature Neuroscience*, 9:1367–1368, 2006. 2

David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664, 2004. 14, 15

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990. 2

Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50, 1988. 20

Zellig Harris. Distributional structure. *Word*, 10(2-3):1456–1162, 1954. 1

Hedi Harzallah, Frédéric Jurie, and Cordelia Schmid. Combining efficient object localization and image classification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 237–244. IEEE, 2009. 86, 88

James Haxby, Ida Gobbini, Maura Furey, Alumit Ishai, Jennifer Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425–2430, 2001. 94

Geoffrey Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, 2006. 115, 124

Alexander Huth, Shinji Nishimoto, An Vu, and Jack Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012. 94

Michael Jamieson, Afsaneh Fazly, Suzanne Stevenson, Sven Dickinson, and Sven Wachsmuth. Using language to learn structured appearance models for image annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):148–164, 2010. 16

Brendan Johns and Michael Jones. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4(1):103–120, 2012. 4, 14

Nancy Kanwisher and Galit Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476):2109–2128, 2006. 98

George Karypis. CLUTO: A clustering toolkit. Technical Report 02-017, University of Minnesota Department of Computer Science, 2003. 64

Michael Kaschak, Carol Madden, David Therriault, Richard Yaxley, Mark Aveyard, Adrienne Blanchard, and Rolf Zwaan. Perception of motion affects language processing. *Cognition*, 94:B79–B89, 2005. 3

Brent Kievit-Kylar and Michael Jones. The Semantic Pictionary project. In *Proceedings of CogSci*, pages 2229–2234, Austin, TX, 2011. 4

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008. 102

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of CVPR*, Colorado Springs, MSA, 2011. 16, 118

C H Lampert, M B Blaschko, and T Hofmann. Efficient Subwindow Search: A Branch and Bound Framework for Object Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:2129–2142, 2009. 16

Thomas Landauer and Susan Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997. 4, 10, 76, 118

Ivan Laptev. Improving object detection with boosted histograms. *Image and Vision Computing*, 27(5):535–544, 2009. 88

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of CVPR*, pages 2169–2178, Washington, DC, 2006. 33, 88, 100, 121

Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2001. 124

Chee Wee Leong and Rada Mihalcea. Going beyond text: A hybrid image-text approach for measuring word relatedness. In *Proceedings of IJCNLP*, pages 1403–1407, 2011. 13, 43, 58, 119

Tony Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998. 22

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982. 33

Max Louwerse. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3:273–302, 2011. 3, 71

Max Louwerse and Louise Connell. A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, 35: 381–398, 2011. 71

David Lowe. Object recognition from local scale-invariant features. In *Proceedings of ICCV*, pages 1150–1157, 1999. 23, 31

David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 2004. 23, 31, 87, 120

Will Lowe. Towards a theory of semantic space. In *Proceedings of CogSci*, pages 576–581, Edinburgh, UK, 2001. 9

Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28:203–208, 1996. 4, 10, 35

Chris Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999. 10

Chris Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008. 41, 124

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of ICML*, Edinburgh, UK, 2012. 17

Bruce D McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, 7(7):293–299, 2003. 98

Robert D McIntosh and Thomas Schenk. Two visual streams for perception and action: Current trends. *Neuropsychologia*, 47(6):1391–1396, 2009. 98

Ken McRae, George Cree, Mark Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, 2005. 3, 15

Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, 2005. 20, 23, 117

Earl K Miller, David J Freedman, and Jonathan D Wallis. The prefrontal cortex: categories, concepts and cognition. *Philosophical Transactions of the Royal*

Society of London. Series B: Biological Sciences, 357(1424):1123–1136, 2002. 98

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

Tom Mitchell, Svetlana Shinkareva, Andrew Carlson, Kai-Min Chang, Vincente Malave, Robert Mason, and Marcel Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191–1195, 2008. 95, 96, 99, 101, 109

Saif Mohammad. Colourful language: Measuring word-colour associations. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 97–106, Portland, Oregon, 2011. 84

David Moore and George McCabe. *Introduction to the Practice of Statistics*. Freeman, New York, 5 edition, 2005. 59

Frank Moosmann, Bill Triggs, Frederic Jurie, et al. Fast discriminative visual codebooks using randomized clustering forests. *Advances in Neural Information Processing Systems 19*, pages 985–992, 2006. 87, 88

Brian Murphy, Partha Talukdar, and Tom Mitchell. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of *SEM*, pages 114–123, Montreal, Canada, 2012. 95, 99, 101

Gregory Murphy. *The Big Book of Concepts*. MIT Press, Cambridge, MA, 2002. 11

Douglas Nelson, Cathy McEvoy, and Thomas Schreiber. The University of South Florida word association, rhyme, and word fragment norms. `http://www.usf.edu/FreeAssociation/`, 1998. 15

David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2161–2168, 2006. ISBN 0-7695-2597-0. 25

Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *Proceedings of ECCV*, pages 490–503, Graz, Austria, 2006. 32

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001. 77

Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007. 9, 10, 44

Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom Mitchell. Zero-shot learning with semantic output codes. In *Proceedings of NIPS*, pages 1410–1418, Vancouver, Canada, 2009. 95

Diane Pecher, René Zeelenberg, and Jeroen Raaijmakers. Does pizza prime coin? Perceptual priming in lexical decision and pronunciation. *Journal of Memory and Language*, 38:401–418, 1998. 41

Marius V Peelen and Paul E Downing. Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, 93(1):603–608, 2005. 98

Francisco Pereira, Greg Detre, and Matthew Botvinick. Generating text from functional brain images. *Frontiers in Human Neuroscience*, 5(72), 2011. Published online: `http://www.frontiersin.org/human_neuroscience/10.3389/fnhum.2011.00072/abstract`. 95

Florent Perronnin, Jorge Sanchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of ECCV*, pages 143–156, Berlin, Heidelberg, 2010. 29, 121

Trong-Ton Pham, Nicolas Maillot, Joo-Hwee Lim, and Jean-Pierre Chevallet. Latent semantic fusion model for image retrieval and annotation. In *Proceedings of CIKM*, pages 439–443, Lisboa, Portugal, 2007. 41

Massimo Poesio and Abdulrahman Almuhareb. Identifying concept attributes using a classifier. In *Proceedings of the ACL Workshop on Deep Lexical Semantics*, pages 18–27, Ann Arbor, MI, 2005. 11

Friedemann Pulvermueller. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6:576–582, 2005. 2

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of WWW*, pages 337–346, Hyderabad, India, 2011. 11

Reinhard Rapp. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the 9th MT Summit*, pages 315–322, New Orleans, LA, 2003. 35

Joseph Reisinger and Raymond J. Mooney. Multi-prototype vector-space models of word meaning. In *Proceedings of NAACL*, pages 109–117, Los Angeles, CA, 2010. 11, 34

Brian Riordan and Michael Jones. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):1–43, 2011. 3, 53, 118

Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bern Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *Proceedings of CVPR*, San Francisco, CA, 2010. 16

Klaus Rothenhäusler and Hinrich Schütze. Unsupervised classification with dependency based word spaces. In *Proceedings of the EACL GEMS Workshop*, pages 17–24, Athens, Greece, 2009. 11

Herbert Rubenstein and John Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965. 11

Alexander T Sack. Parietal cortex and spatial cognition. *Behavioural brain research*, 202(2):153–161, 2009. 98

Magnus Sahlgren. An introduction to random indexing. `http://www.sics.se/~mange/papers/RI_intro.pdf`, 2005. 9, 10

Magnus Sahlgren. *The Word-Space Model*. Dissertation, Stockholm University, 2006. 10, 36, 53, 62

Magnus Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–53, 2008. 35

Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors. *International Journal of computer vision*, 37(2):151–172, 2000. 21

Hinrich Schütze. *Ambiguity Resolution in Natural Language Learning.* CSLI, Stanford, CA, 1997. 10, 101

Carina Silberer and Mirella Lapata. Grounded models of semantic representation. In *Proceedings of EMNLP-CoNLL*, pages 1423–1433, Jeju, Korea, 2012. 4, 14

Carina Silberer and Mirella Lapata. Models of semantic representation with visual attributes. In *Proceedings of ACL*, Sofia, Bulgaria, 2013. 15

Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, pages 1470–1477, Nice, France, 2003. 25, 87, 117

Mark Steyvers. Combining feature norms and text data with topic models. *Acta Psychologica*, 133(3):234–243, 2010. 4

Richard Szeliski. *Computer Vision : Algorithms and Applications.* Springer, Berlin, 2010. 77, 117

David Therriault, Richard Yaxley, and Rolf Zwaan. The role of color diagnosticity in object recognition and representation. *Cognitive Processing*, 10(4):335–342, 2009. 3

Richard Tillman, Vivek Datla, Sterling Hutchinson, and Max Louwerse. From head to toe: Embodiment through statistical linguistic frequencies. In *Proceedings of CogSci*, pages 2434–2439, Austin, TX, 2012. 71

Peter Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010. 1, 9, 10, 12, 95, 118

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, pages 680–690, Edinburgh, UK., 2011. 45

N Tzourio-Mazoyer, B Landeau, D Papathanassiou, F Crivello, O Etard, N Delcroix, B Mazoyer, and M Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002. 99

Jasper RR Uijlings, Arnold WM Smeulders, and Remko JH Scha. Real-time visual concept classification. *Multimedia, IEEE Transactions on*, 12(7):665–681, 2010. 87

Jasper RR Uijlings, Arnold WM Smeulders, and Remko JH Scha. The visual extent of an object. *International journal of computer vision*, 96(1):46–63, 2011. 94

J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *IJCV*, 2013. 122

Koen van de Sande, Theo Gevers, and Cees Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010. 32, 87

Koen van de Sande, Jasper Uijlings, Theo Gevers, and Arnold Smeulders. Segmentation as selective search for object recognition. In *Proceedings of ICCV*, pages 1879–1886, Barcelona, Spain, 2011. 86, 87, 88

James Van Overschelde, Katherine Rawson, and John Dunlosky. Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, 50:289–335, 2004. 63

Andrea Vedaldi and Brian Fulkerson. Vlfeat – an open and portable library of computer vision algorithms. In *Proceedings of ACM Multimedia*, pages 1469–1472, Firenze, Italy, 2010. 32, 119

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001. 88

Luis Von Ahn. Games with a purpose. *Computer*, 29(6):92–94, 2006. 27

Gang Wang and David A. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *Proceedings of ICCV*, pages 537–544, Kyoto, Japan, 2009. 16

Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Proceedings of CVPR*, pages 3360–3367, San Francisco, CA, 2010. 29

Julie Weeds. *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, Department of Informatics, University of Sussex, 2003. 44

Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, UK, 1953. Translated by G.E.M. Anscombe. 1

Jun Yang, Yu-Gang Jiang, Alexander Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In James Ze Wang, Nozha Boujemaa, Alberto Del Bimbo, and Jia Li, editors, *Multimedia Information Retrieval*, pages 197–206. ACM, 2007. 25

Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, and Zengfu Wang. Visual query suggestion. In *Proceedings of ACM Multimedia*, pages 15–24, Beijing, China, 2009. 16

Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis. Technical Report 01-40, University of Minnesota Department of Computer Science, 2003. 64

Song Chun Zhu, Cheng en Guo, Ying Nian Wu, and Yizhou Wang. What are textons? In *Proceedings of ECCV*, pages 793–807, Copenhagen, Denmark, 2002. 77