



European Master in
Language & Communication Technologies (LCT)

Multimodal Fusion for Combining Textual and Visual Information in a Semantic Model

Nam Khanh Tran

Supervisors: Dr. Marco Baroni
Prof. Dr. Manfred Pinkal

July, 2012

*to me & my family & my love,
and to all who helped me make this thesis work*

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisors, Dr. Marco Baroni, and Prof. Dr. Manfred Pinkal for their support and guidance during the completion of this thesis. With their encouragement and kind assistance, I found myself interested in broadening my knowledge in the area of computational semantics, examining different methods, solving problems and making improvements. The work in this thesis could not have been accomplished without their support.

My special thanks go to researchers in the Language, Interaction and Computation Laboratory (CLIC-CIMEC), especially Elia Bruni, who discussed and gave me a lot of invaluable advice. The general architecture and part of the experiments we present in this thesis are the product of joint work with Elia Bruni and Dr. Marco Baroni.

I would like to thank the committee of The European Masters Program in Language and Communication Technologies (LCT) for giving me such a fantastic opportunity and financial support to pursue my master degree. Many thanks to my coordinators PD Dr. Valia Kordoni, Prof. Nicu Sebe, and Mrs. Bobbye Pernice, Mrs. Marta Cazzanelli, Ms. Maria Jacob for their kind assistance with my living situations and residency in Europe.

Finally, I owe my loving thanks to my family and my love for their constant love and support. I always find my inspiration and encouragement from all of you, my dear family. A big hug to all of my friends in Vietnam, Saarbruecken, Trento for great times together.

Eidesstattliche Erklaerung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Trento, 25 June, 2012

Nam Khanh Tran

Abstract

In recent times, multimodal fusion that combines information from different data streams such as text and images, has attracted the attention of many researchers due to the benefit it provides for various analytical tasks. In the computational semantics community, recent work has extended distributional semantics to integrate features extracted from images as well as text. In this thesis, we propose a flexible and general framework for the integration of text- and image-based features that includes both a latent mixing phase in which text and image features are mapped together onto a lower-dimensional space, and a multimodal similarity estimation phase, where the two channels are linearly combined either at the feature or at the scoring level. The framework is of sufficient generality that, with specific parameter choices, it encompasses combination techniques that have been previously proposed in the literature. In addition, it can automatically pick the best model for a task given development data.

We evaluate our framework through extensive experiments on simulating similarity judgements, and concept categorization. A further experiment to examine the effect of different image-based features combined with a textual model into a multimodal architecture is also presented. It is shown that our system significantly outperforms a state-of-the-art text-based distributional semantic model on the MEN semantic relatedness evaluation dataset and, using the parameter values tuned on MEN, is also the top performer on the independent WordSim evaluation. In a similar manner, we obtain improvements on AP, a widely used concept classification benchmark, with the use of the parameters tuned on another dataset, BATTIG. The good performance of the multimodal models using a different test set from the one used for parameter tuning demonstrates the robustness of our approach. Moreover, we also show evidence that adding better visual features improves the performance of the system.

Keywords: Multimodal fusion, distributional semantics, latent semantic mixing, multimodal similarity estimation.

Contents

Acknowledgments	iii
Abstract	v
1 Introduction	1
1.1 Motivation	1
1.2 Contributions	2
1.3 Structure of the thesis	3
2 Literature Review	4
2.1 Distributional semantic models	4
2.2 Multimodal distributional semantics	7
2.3 Multimodal fusion for multimedia analysis	8
3 A General Framework for Multimodal Fusion	10
3.1 A general framework for multimodal fusion	10
3.2 Text- and image-based models construction	11
3.2.1 Text-based models	11
3.2.2 Image-based model	12
3.3 Latent multimodal mixing	14
3.4 Multimodal similarity estimation	16
3.4.1 Feature-level estimation	16
3.4.2 Scoring-level estimation	16
3.5 General form and special cases	17
4 Evaluation	18
4.1 Task definitions	18
4.2 Experimental setup	19
4.2.1 Evaluation datasets	19
4.2.2 Similarity and correlation measures	22
4.2.3 MixLDA	23
4.3 Model selection	23
4.3.1 Semantic relatedness	24

4.3.2	Concept categorization	26
4.4	Experimental results	26
4.4.1	Semantic relatedness	26
4.4.2	Concept categorization	28
4.4.3	Image-based model selection	29
4.4.4	Result summary	30
5	Conclusion and Future Work	32
5.1	Thesis achievements	32
5.2	Future work	33
A	Usage of the Modules Developed in this Thesis	35
A.1	Download link	35
A.2	Description of each module	35
A.2.1	Latent semantic mixing	35
A.2.2	Multimodal similarity estimation	36
A.2.3	MEN dataset	36
A.2.4	Evaluation	37
	Bibliography	38

List of Figures

2.1	Example of words' vectors in a semantic space	5
3.1	An overview of the general framework for multimodal fusion combining textual and visual information in distributional semantic models	11
3.2	Image examples with all their labels from ESP-Game dataset	13
3.3	The procedure to build a vector of visual words.	13
3.4	Feature-level similarity estimation	16
3.5	Scoring-level similarity estimation	16
4.1	Parameter search space for the FL strategy (k in terms of \log_2)	23
4.2	Performance of the models with Window20 as text-based resource by varying α and β parameters and fixing the latent dimension $k = 2^8$	24
4.3	Performance of the models with Window20 as text-based resource by varying the number of latent dimensions	25

List of Tables

4.1	Examples of word pairs and their semantic ratings in the MEN dataset	20
4.2	Examples of word pairs and their semantic ratings in the WS dataset .	20
4.3	Word pairs' types in MEN dataset	21
4.4	Examples of concepts and categories in BATTIG	21
4.5	Examples of concepts and categories in AP	22
4.6	The best values of α and β for each latent dimension (in terms of \log_2)	25
4.7	The best values of α , β and latent dimension (in terms of \log_2)	26
4.8	The best values of α , β and latent dimension (in terms of \log_2)	26
4.9	Spearman correlation of the models on MEN and WordSim (all coefficients significant with $p < 0.001$).	27
4.10	Top 10 pairs whose similarity is better captured by Text vs. FL.	28
4.11	Percentage purities of the models on AP dataset	29
4.12	The performance of different visual models	30
4.13	Spearman correlations of the models on MEN and WS (all coefficients significant with $p < 0.001$).	30

1.1 Motivation

In recent years, computer scientists and engineers have had the strong ambition that produces computers that can independently think and make decision like humans. However, one of the biggest obstacles to reach this goal is that computers currently understand very little of the meaning of human language. This problem is often referred as the semantic problem [52].

In linguistics, Semantics is the study of the meaning of linguistic expressions or how the components of language (words and phrases) correspond to concepts in the communicator’s mind. In the field of natural language processing, computational semantics concerns the usage aspect of language, where it studies automated methods for acquiring and using knowledge about linguistic meaning [48].

As one of the most popular and interesting methods in computational semantics, the distributional approach has recently been pursued by many scientists. Distributional semantic models (DSMs), also known as “word space” or “distributional similarity” models, approximate word meanings with vectors that keep track of patterns of co-occurrence of words in large collections of text (Landauer and Dumais (1997) [31]; Sahlgren (2006) [47]; Turney and Pantel (2010) [53]). In general, DSMs are based on some version of the *distributional hypothesis* proposed by Firth (1957) [20], Harris (1970) [24] and Miller & Charles (1991) [38]. The hypothesis states that the meaning of a word can be inferred from its usage and words that occur in similar contexts are semantically similar. Extended empirical evidence has confirmed that distributional semantics is very effective at harvesting meaning representations on a large scale, because words that have similar meanings tend to occur in similar textual contexts, and thus co-occur with the same context words. For example, the word *moon* may be observed in the same contexts as the word *sun*. As a result, it can be inferred that these words are semantically related. In contrast, *moon* is less similar to *cat* as cued by the fact that they rarely appear in the same contexts. Distributional semantic vectors can be used in any application that requires a measurement of semantic similarity or relatedness between words, ranging from document retrieval and classification to question answering, automated thesaurus construction and machine translation [53].

Despite its success, distributional semantics still severely suffers of the limitations that it represents the meaning of a word entirely in terms of connections to other words. Besides the obvious difference with the way in which humans acquire meaning (not only from language, but also from perception), this limitation leads to empirical weaknesses. For example, many studies [2, 3, 4, 45] have underlined how text-derived distributional models capture encyclopedic, functional and discourse-related properties of word meanings, but 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 feel the need to write down obvious statements such as “bananas are yellow”).

Very recently, concurrent advances in computer vision [22] and the availability of large amounts of Web 2.0 documents that combine words and images have made it possible to build distributional models of meaning that break out of the textual barrier and exploit co-occurrence patterns of words with image-extracted features, to achieve a richer and more cognitively grounded representation of word meaning [9, 18, 33].

An important issue arising in the context of these novel *multimodal* distributional semantic models is how to combine the features extracted from the textual and visual information channels, a process we refer to as *multimodal fusion*. Feng and Lapata (2010) [18] map text- and image-based features onto the same latent semantic dimensions, and represent word meanings as distributions over these dimensions. Bruni et al. (2011) [9] simply concatenate the textual and visual features. Leong and Michalcea (2011) [33], finally, compute separate similarity measures in textual and visual space and then combine them.

In this thesis, we present a general and flexible architecture for multimodal fusion, that encompasses all previously proposed strategies in a single parametrized pipeline. We show that, by fixing the parameters on training data, we derive multimodal models that, for the first time, outperform state-of-the-art purely text-based models in a statistically significant manner, as well as surpassing the models of Feng and Lapata, Bruni et al. and Leong and Michalcea, when the latter are trained on the same sources. The good performance of the multimodal models generalizes, moreover, to a different test set from the one used for parameter tuning, demonstrating the robustness of the method. Moreover, the framework is as an open prototype where several techniques can be examined in each phase without changing the general structure.

1.2 Contributions

The most significant contributions of this thesis are the followings:

- We propose a general framework for integrating visual information into distributional semantics that not only encompasses previous models, but can automatically pick the best model for a task given development data.
- A hybrid fusion approach is presented where the Singular Value Decomposition method is first used to blend visual and textual information; then the two chan-

nels are combined linearly using weighted functions at the feature and scoring levels.

- We provide clear statistical evidence thorough experiments that multimodal models outperform purely textual models, that the hybrid fusion obtains significant improvements over simpler previous approaches and that the better the visual model integrated into textual models is, the better results we obtain.
- We introduce a new evaluation dataset (MEN) with the aim of evaluating multimodal semantic models. It contains thousands of items and several types of words that might be used to test the models precisely.

1.3 Structure of the thesis

The rest of this thesis is organized as follows:

Chapter 2 brings an overview of multimodal distributional semantics and discusses related work. It starts by introducing distributional semantics and its drawbacks that motivate the idea of combining information from multi-sources for improving performance on semantic tasks. After that, some approaches and proposed models are discussed.

Chapter 3 presents the main work of the thesis. It describes in detail our general framework to tackle the problem of multimodal fusion for combining textual and visual information in a semantic model. The chapter first sketches out the text- and image-based models construction and then specifies the two steps of the framework, that is, latent multimodal mixing and multimodal semantic estimation.

Chapter 4 presents a thorough experimental evaluation and discussion. It covers two different experiments, namely semantic relatedness and concept categorization. In both tasks, model selection is first described and then results are discussed.

Chapter 5 sums up the achievements throughout the previous chapters. Some future research topics are also mentioned in this chapter.

This chapter gives an overview of the problem of combining textual and visual information to extract semantic properties of words and related work. We first describe distributional semantic models and their drawbacks. These inspire the issue of integrating visual information into semantic models that is defined in Section 2.1.

The second part of this chapter provides short descriptions of some recent works in multimodal distributional semantics and points out their disadvantages which motivate us to propose our general framework for tackling the problem (see Section 2.2).

In the latter part, we review some multimodal fusion approaches for multimedia analysis. These methods focus on using information extracted from text to tackle vision-related tasks. They are related to the ones we propose in this thesis by mapping textual and visual data to the same latent space to leverage the overlapping of both data modalities (see Section 2.3).

2.1 Distributional semantic models

In recent years, there have been important studies on distributional models of semantic representation. Distributional models build semantic representations from the statistical patterns of co-occurrences of words in large-scale linguistic corpora. These models are based on the distributional hypothesis proposed by Harris (1970) [24] and Miller & Charles (1991) [38]: *The more similar the contexts in which two words appear, the more similar their meanings*. The context surrounding a given word provides information about its meaning. The context can be words around a target word or the documents, paragraphs or sentences that the target word occurs. Consequently, the meaning of a word can be identified by its usage.

Distributional semantic models typically represent word meanings as context vectors in a high-dimensional semantic space where each vector reflects the contextual history of a target word across the text corpus; and it is called as “vector space” or “semantic space”. Words’ vectors are described as points in the semantic space and words that are more semantically related tend to be closer in that space. For example, the word *dog* may be observed in same contexts as *cat*. As a result, these words are expected to be semantically related. On the other hand, *dog* and *computer* rarely occurs in similar contexts; therefore they are not related.

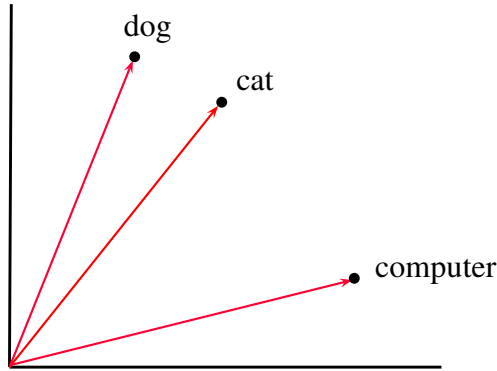


Figure 2.1: Example of words' vectors in a semantic space

Some typical examples of distributional semantic models include Hyperspace Analogue to Language (HAL; Lund and Burgess (1996) [35]), Latent Semantic Analysis (LSA; Landauer and Dumais (1997) [31]), and the Topic model (Griffiths et al. (2007) [51]). The HAL model considers words to be the context and acquires word meanings as a function of keeping track of how words are used in context by taking a moving window of N words in consideration and weighting co-occurrence distance to identify the word co-occurrences in a large linguistic corpus. The width N of the window can be varied depending on tasks [44] and in this thesis, we experiment with the width of 2 and 20 to examine the effects of differing the window size on model performance.

Latent Semantic Analysis is one of the most well-known semantic space model for extracting and representing the contextual-usage meanings of words by statistical computation applied to a large collection of texts [31]. It treats documents or sections of text instead of words as the context. The meaning of a word can be represented as a kind of average of the meaning of all the documents in which it appears, and the meaning of a document as a kind of average of the meaning of all the words it contains. LSA relied on the underlying idea that the aggregate of all the word contexts (e.g. documents or passages) in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and set of words to each other [50]. The key component in LSA is computing the singular value decomposition on the co-occurrence matrix [37]. It projects the high-dimensional semantic space onto a lower-dimensional one (*dimensionality reduction*) where it combines surface information into a deeper abstraction that captures the mutual implications of words and passages; so that words that may not have occurred in the same contexts but in similar contexts will come to have more similar vector representations in this reduced space and will thus be clustered closer together. LSA has been shown to be accurately reflect human-like word similarity judgements and other semantic memory phenomena [31].

The Topic model is considered as a different method from the other distributional models although it is also based on word co-occurrences within documents. The idea of topic modelling is that a document has been sampled from a mixture of latent topics. Each section in the document is associated with a probability distribution over a set of topics, and each topic specifies a probability distribution over words [51].

Each word in the model can be represented as a distribution over the latent topics. Using these representations, the relationships between words can also be inferred. Griffiths et al. (2007) show that the Topic model outperforms other models in the task of predicting word association and some other semantic phenomena.

In addition, Bullianria and Levy (2007) [11] presented an extension work for distributional semantic models where they explore several factors such as lexical association function. They found that transforming raw co-occurrence counts into positive point-wise mutual information scores (only positive scores are retained) helps to improve performance of the model.

Even though distributional semantic models have obtained significant enhancements in many applications, which require a measurement of semantic relatedness between words, such as document retrieval and classification, question answering, word sense disambiguation, machine translation, they have several limitations [2, 3, 4, 45]. Critics have argued that distributional models are inadequate psychological models of semantic learning and representation. The meaning of words cannot be result from word relationships alone; rather it needs to be grounded in perception and action. Andrews et al. (2009) [2] observed that distributional models are disconnected from the physical world. These models describe the relationship of words only to one another but not to the physical world or anything else beyond language itself; thus they alone are not plausible models of human semantic representation. To overcome this limitation, Andrews et al. describe how to combine experimental data, which can characterize the meaning of words in terms of objects and events in the world, and distributional models to learn semantic representations. In this way, knowledge acquired from language can be related to knowledge acquired from physical world. In this vein, Riordan and Jones (2011) [45] return to classic feature-based theories of semantic representation and presented comparisons in semantic representations of feature-based and distributional models using a semantic clustering task. They reported that while information relevant for learning semantic categories is redundantly coded in perceptual and linguistic experience, each source emphasizes different information that is useful for semantic clustering. Distributional models seem to add more weight to information about actions, functions, and situation, but less to information about direct perception related to objects. Therefore, they argued that the two sources could be integrated to obtain better understandings of word meanings. However, these works suffer of some limitations from practical and theoretical aspects. They use surrogates of true perceptual features defined by human subjects. Consequently, this limits the number of words can be described and lacks of the sort of implicit perceptual features of a given concept.

In this thesis, we focus on how to integrate perceptual information, specifically the information extracted from images, into distributional semantic models. Image-based features have been shown as complementary source of text-based features in harvesting the meaning of words [18]. As an intuitive example, from the text, we can harvest that *bananas* are *tropical* and *eatable* but not that they are *yellow*, that can be learned from *bananas* images. Section 2.2 describes current approaches for solving this problem, then our approach is presented afterwards.

2.2 Multimodal distributional semantics

In recent years, there are several works in improving a general purpose representation of word meaning by taking advantages of the combination of textual information and one more type of perceptual information, called multimodal semantics (it is also referred to “grounded language learning”) [18, 33, 9].

Feng and Lapata (2010) [18] propose the first multimodal distributional semantic model where they showed that it is possible to combine visual representation of word meanings into a joint bimodal representation constructed by using probabilistic generative latent topic models. Their generative probabilistic setting requires the extraction of text- and image-based features from the same mixed-media corpus. Specifically, they assume that the images and their surrounding text have been generated by a shared set of topics. However, there is the fact that images and documents represent distinct modalities: images are commonly described by a continuous feature space (e.g., color, shape, texture) whereas words in documents are discrete. In order to make both modalities possible to combine, they convert the visual features from a continuous onto a discrete space, thereby rendering image features more like word units. Since both images and documents, which are assumed expressing the same content, are now described as bags-of-words, a document and its associated image are represented as a mixture of verbal and visual words. A topic model is then learned on this concatenated representation. In this framework, word meaning is now represented by a word distribution over the set of latent multimodal dimensions or topics (Griffiths et al., 2007 [51]) derived from the surface textual and visual features and the similarity between two words can be estimated by measuring the topics they have in common. They experiment with a collection of documents downloaded from the BBC News website as corpus and the WordSim-353 evaluation test. They obtain better results when visual information is taken into account, but the overall performance is very modest. The main drawbacks of this approach are that the textual and visual data must be extracted from the same corpus, and that the generative probabilistic approach, while elegant, does not allow much flexibility in how the two information channels are combined. In addition, while they provided a proof-of-concept that using the visual modality leads to an improvement over their purely text-based model, no attempt has been made to evaluate the image-based models independently [33]. As shown below in Section 4.2.3, we reimplemented their approach (MixLDA) using as mixed-media training corpus the ESP-Game dataset, that is, the same source of labelled images we adopt for our model. This is possible because the dataset contains both images and textual labels describing them. For comparability, we used the same image processing pipeline as for the other models in this thesis. 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 3.3).

In order to overcome some drawbacks of Feng and Lapata’s approach, Bruni et al. (2011) [9] propose instead to directly concatenate the text- and image-based vectors that are independently constructed from different sources to produce a single multimodal vector to represent words, as in what we call Feature-Level fusion. In this approach, the text-based distributional vector representing a word taken from

a state-of-the-art distributional model (DM [5]) is concatenated with a vector representing the same word with visual features, extracted from all the images in the ESP Game collection. That means that the image-based features are not restricted to be extracted from the same corpus texts are taken from. Therefore, the impact of the visual and textual components can be assessed quantitatively and qualitatively separately, and the image-based model can be added to a state-of-the-art text-based distributional model. They evaluate the performance of their approach in different tasks, specifically, semantic relatedness, concept categorization and capturing semantic neighbours of different classes. For all tasks, they obtain promising performance, although appreciably lower than the results we report here. That is because the method is too simple, it does not make use of overlapping or complementing between two data modalities. This inspires us to propose our framework where we add a further phase, latent multimodal mixing, to leverage this overlapping. As we discuss in Section 3.5, their approach is a special case of ours, and we again report its results in Section 4.4.

Leong and Mihalcea (2011) [33] propose a strategy, akin to what we call Scoring-Level fusion below, that comes up with separate text- and image-based similarity estimates, and combine them to obtain the multimodal score. Leong and Mihalcea use two combination methods, that is, summing the scores and computing their harmonic mean. They make use of ImageNet, a large-scale ontology of images (Deng et al., 2009)[14], as their image source. Since ImageNet leverages on reliable semantic annotation in WordNet, the propagation of errors caused by unreliable annotations can be effectively circumvented. They experiment with small subsets of WordSim-353 and obtain significant correlations, even though not at the same level we report. Regardless of the performance of the individual image-based model, they also show that the multimodal approach improves over the standalone text-based model for both knowledge-based and corpus-based methods. As we discuss in Section 3.5, their summing strategy is a special case of our general framework, and we report its performance on our evaluation datasets in Section 4.4. Moreover, we perform a further step including defining a parametrized combination function where summing method is a specific case, and tuning instead of fixing the parameter on development data to obtain the best performance.

2.3 Multimodal fusion for multimedia analysis

In recent years, the multimedia and computer vision communities have shown much interest in using information extracted from textual data associated to images to tackle vision-related tasks such as automated image annotation, textual descriptions of images generation, region naming or exploiting textual information to induce visual attributes [29, 6, 16, 17, 30]. While Barnard et al. (2003) [29] 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, both Farhadi et al. (2010) [17] and Kulkarni et al. (2011) [30] aim to create more natural descriptions for images than just tags.

There are also studies addressing the topic of multimodal media retrieval that use multimodal fusion techniques related to the ones we propose in this thesis [42, 27, 25, 46]. Pham et al. (2007) [42] study the effect of Latent Semantic Analysis (LSA) on multimedia document retrieval and automatic image annotation. Each modality of a multimedia document including text and image is processed separately to obtain the text- and image-based matrices. The two matrices are fused by row-concatenating and then projecting onto a latent space to gain the reduced matrix. The process is similar to what we perform in latent multimodal mixing step. After that, given a query with both text and image, the document retrieval is performed in the latent space. Caicedo et al. (2012) [27] present an approach that is similar to what we propose in this thesis. However, they make use of non-negative matrix factorization to generate multimodal representations that integrate visual features and text information. They reported that the approach outperforms the response of the system in both image indexing and search tasks, when compared to multimodal latent semantic spaces generated by singular value decomposition. It motivates us to perform more analysis in further work in order to compare both approaches in text processing tasks.

The two above approaches are categorized as early fusion strategies for multimedia image analysis. Another category called late fusion (concatenating outputs of single-modality systems) is also worth of attention. Escalante et al. (2008) [25] present several methods or late fusion of heterogeneous sources for multimedia image retrieval. This approach consists of building several retrieval systems including uni-modal and multi-modal search methods using subsets of the same collection of documents. The outputs of these different independent retrieval methods are then combined to obtain the final list of ranked documents where they consider a simple score based on a weighted linear combination of the document rank through the different lists. The idea of this approach is analogous to the scoring level in our multimodal similarity estimate phase (see Section 3.4).

In all aforementioned efforts, the goal of these works was to use information extracted from text to improve image-related tasks, and typically they attempt to automatically annotate or retrieve specific images or to associate specific textual passages to specific images. In contrast, we propose an approach that exploits visual information extracted from image collections to enhance a general purpose representation of word meaning. Our approach is categorized as a hybrid fusion strategy where we first perform latent multimodal mixing using singular value decomposition (early fusion) and then estimate multimodal similarity combining information from two channels at feature level (early fusion) and scoring level (late fusion).

Chapter 3

A General Framework for Multimodal Fusion

The chapter introduces our general framework for multimodal fusion integrating visual information into distributional semantic models. We start by describing the data pre-processing step that uses text and image analysing techniques to obtain vectors in the textual and visual modalities. Then, the two phases of the framework are presented including a *latent multimodal mixing* phase in which the two data modalities are mapped together onto a lower-dimensional space, and *multimodal similarity estimation* where the two channels are combined either at the feature or at the scoring level.

3.1 A general framework for multimodal fusion

Figure 3.1 describes our general framework. It takes the text and image matrices from a data pre-processing phase called text- and image-based models construction as its inputs. Basically, the preparation step processes a text corpus and labelled image data by using text/image processing techniques to represent target words in terms of separate text- and image-based feature vectors, that are stored in target-by-feature matrices. The general framework consists of two steps:

- **Latent multimodal mixing:** This step generates a latent semantic space using both data modalities. The goal is to find a set of latent factors that explain the relationships between multimodal features. In order to do that, the text and image matrices are row-normalized and concatenated, obtaining a single matrix whose row vectors are projected onto a lower dimensionality space by Singular Value Decomposition, and then projected back to obtain a concatenated matrix with the original features but lower rank.
- **Multimodal similarity estimation:** Information in the text- and vision-based (sub)matrices is combined in two ways to obtain similarity estimates for pairs of target words: at the feature level and at the scoring level. Under the first strategy, the concatenated text- and image-based feature vectors of words are used as their single representation, and similarity of pairs is estimated by

comparing these multimodal representations. In the second approach, independent similarity estimates are obtained in text and image space, and the two scores are combined to provide the final verdict on the degree of similarity of a pair.

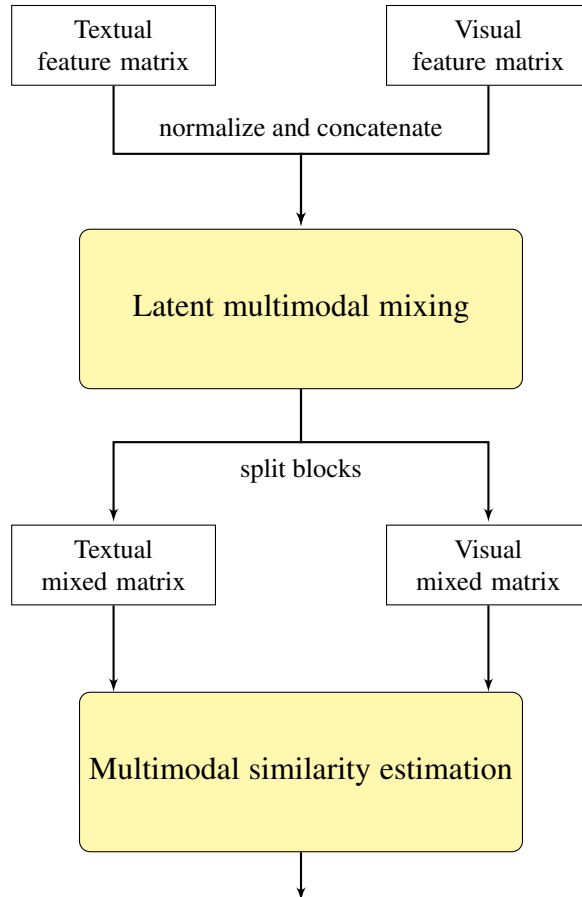


Figure 3.1: An overview of the general framework for multimodal fusion combining textual and visual information in distributional semantic models

In the following subsections, we first describe briefly the data preparation step in Section 3.2 and then specifically the two steps in Section 3.3 and 3.4

3.2 Text- and image-based models construction

3.2.1 Text-based models

Text-based distributional semantic models approximate the meaning of words with vectors that record their distributional history in a corpus [53]. We experiment here with two models based on counting co-occurrences within a window of fixed width, in the tradition of HAL [35]. Window-based models have been shown to be at the

state of the art in various semantic tasks [44, 47], and the high correlations with similarity ratings we report below confirm their quality. Moreover, Bruni et al. (2012) showed that the window-based models outperform other text-based models including a document-based model and a syntax-based model (DM) in the same semantic similarity tasks we consider here. In this thesis, we examine two variants, **Window2** and **Window20**. Window2 records sentence-internal co-occurrence with the nearest 2 content words to the left and right of each target word (expected to capture tighter, taxonomic relations). Window20 considers a larger window of 20 words to the left and right of the target (expected to capture broader, topical relations). All the parameters to be described next are common to both models and were picked without tuning.

Co-occurrence statistics are extracted from the freely available ukWaC and Wackypedia corpora combined (size: 1.9B and 820M tokens, respectively).¹ Raw co-occurrence counts are transformed into non-negative Local Mutual Information (LMI) scores, a close approximation to the popular Log-Likelihood Ratio weighting [15]. It is computed as following formula:

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

where t and w stand for the target and context words, respectively; $P(x)$ is the probability of x calculated by maximum likelihood using relative frequency.

From the corpora, we harvest vector representations for lemmas, and in particular the top 20K most frequent nouns, 5K most frequent adjectives and 5K most frequent verbs in the combined corpora. The same 30K nouns, verbs and adjectives are also employed as contextual elements.

3.2.2 Image-based model

We use image data where each image is associated with one or more words or tags. We use the ESP-Game dataset,² containing 100K images labeled through a game with a purpose in which two people partnered online must independently and rapidly agree on an appropriate word to label the selected images [56]. The images are a random sample from the Web, and they are not chosen to be representative of the words they are labeled with. There are 20,515 distinct words in the dataset, with an average of 4 words per image. We build vectors with visual features for each word in the ESP dataset.

The visual features are extracted with the use of a standard bag-of-visual-words (**BOVW**) representation of images, inspired by NLP [7, 13, 41, 49, 54, 58]. The approach relies on the notion of a common vocabulary of “visual words” that can serve as discrete representations of all images. The process to extract visual words and use them to create image-based feature vectors to represent target textual words is schematically illustrated in Figure 3.3, for a hypothetical example in which there is

¹<http://wacky.sslmit.unibo.it/>

²<http://www.espgame.org>

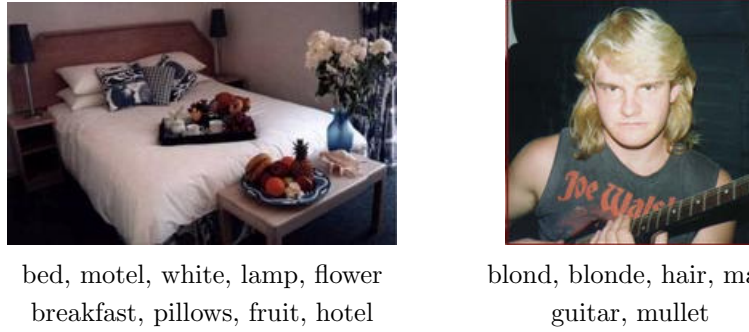


Figure 3.2: Image examples with all their labels from ESP-Game dataset

only one image in the collection labelled with the word *horse*. Following the standard procedure in vision feature extraction, from every image in the dataset, keypoints are automatically detected with the use of the Difference of Gaussian (DoG) detector and represented as low-level feature vector by using the Scale-Invariant Feature Transform (SIFT) [34]. These vectors are then projected onto a common space and grouped into a number of clusters. Each cluster is treated as a discrete visual word (vector quantization). To extract information from a particular image, each keypoint is mapped onto a visual word, such that the image can be represented as a BoVW feature vector, by counting the instances of each visual word in the image. We also use spatial histograms to introduce weak *geometry* in BoVW [21, 32]. The main idea is to divide the image in several (spatial) regions, perform the entire BoVW for each region, and then concatenate the vectors.³

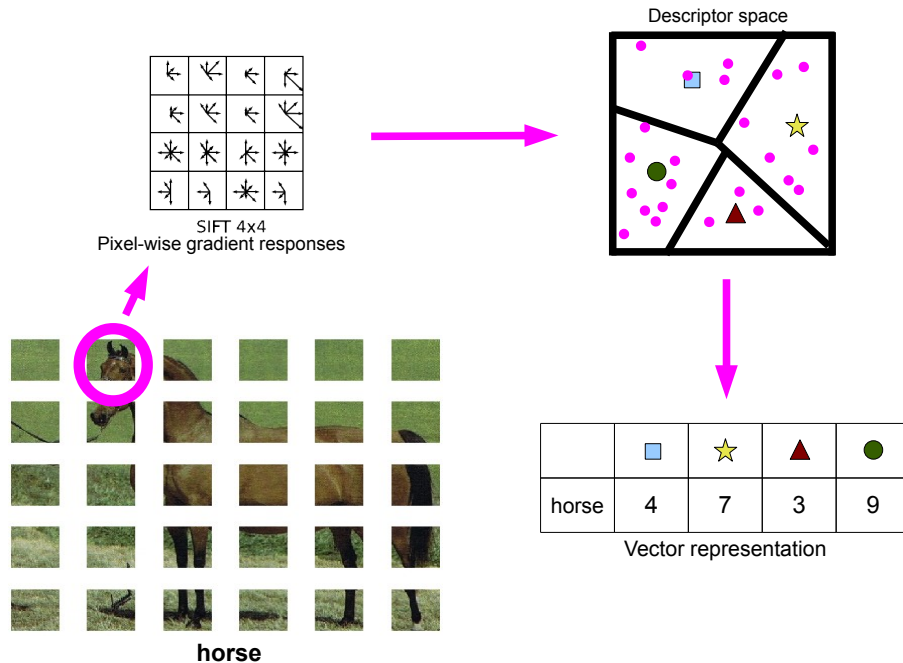


Figure 3.3: The procedure to build a vector of visual words.

³We use VLFeat <http://www.vlfeat.org> for visual word extraction [55]

In this thesis, we extract color (HSV-based) SIFT descriptors on a regular grid with five pixels spacing, at four multiple scales (10, 15, 20, 25 pixel radii) and zeroing the low contrast descriptors. To map SIFT descriptors to visual words, we cluster them in the 128-dimensional space using the k -means clustering algorithm. We set k , the number of visual words, to 5,000. The spatial regions are obtained by dividing the image in 4×4 , for a total of 16 regions. Crossing the values for k with the spatial region, we increase the feature dimensions 16 times, for a total feature number of 80K dimensions. The parameters of the visual word extraction pipeline are standard choices in the computer vision literatur and were adopted without tuning.

Once the BoVW representation are built, each target textual word is associated to the list of images which are labelled with it; the visual word occurrences across the list of images is summed to obtain the co-occurrence counts associated with the target word. The raw counts are then transformed into LMI scores as the formula in Section 3.2.1

3.3 Latent multimodal mixing

This step generates a latent semantic space using both data modalities. The aim is to find a set of latent factors that explain the relationships between multimodal features. It takes the two data modalities, herein denoted by $M_t \in R^{w \times n}$ and $M_v \in R^{w \times m}$ as its inputs.⁴ The former is a matrix whose rows are indexed by w target words and whose columns correspond to the n text terms. The latter has m columns to represent visual features and w rows for target words as well. The w rows of these matrices represent words at intersection of text- and image-based target words. Each of the two matrices provide information about the occurrence of different features for each target word in which values account for the number of times a particular element appears with the target word. In our case, these values in both matrices are LMI scores transformed from raw counts. The purpose of a latent factor model is to try to explain these occurrences by characterizing both target words and features using a set of factors inferred from occurrence patterns.

There are several techniques that can be used to perform this task such as Non-negative Matrix Factorization, Singular Value Decomposition (**SVD**). In this thesis, we adopt SVD to compute the latent factors.⁵ This strategy consists of estimating a rank-reduced factorization of the feature matrix in terms of its eigenvectors and eigenvalues. The latent factors obtained by using SVD are orthogonal factors that preserve the maximum amount of variance from the original data representation [36, Ch. 18]. Moreover, by performing SVD on the concatenated textual and visual matrix (blended matrix), we project the two types of information into the same space, where they are described as linear combinations of principal components. 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 overlaps. As a toy example, consider the concepts *moon* and *ball*. These are visually

⁴Feature vectors for each target word are normalized to have L2 norm $l_2 = 1$, in both textual and visual matrices

⁵Computed with SVDLIBC: <http://tedlab.mit.edu/~dr/SVDLIBC/>

highly similar, but the corresponding words are not likely to occur in similar contexts in text. Therefore, the original text features *moon* and *ball* might not be highly correlated. However, after performing SVD on the blended matrix or multimodal space, they might both be associated with (have high weights on) the same principal 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).

Following the description in Pham et al. (2007), the SVD of a matrix M of rank r is a factorization of the form

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

where

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

Using SVD, any matrix M can be factored into an orthonormal matrix U , a diagonal matrix Σ , and an orthonormal matrix V^t . The singular values in Σ can be ordered from largest to smallest, where the larger values correspond to the vectors in U and V that are more significant components of the initial M matrix. The principal components of the data are represented by the largest singular values in Σ and their corresponding rows of U and columns of V^t . In our context, the multimodal matrix M is constructed by normalizing two feature matrices separately and then concatenating $M = \gamma M_t \oplus (1 - \gamma) M_v$ with $\gamma \in [0, 1]$, a weighting parameter that controls the relative importance of the two data modalities, and \oplus the vector concatenate operator. In our experiment, we set $\gamma = 0.5$ to give the same importance to visual and text data.

When making use of the SVD results, we usually keep the first k components, that is, the principal components of M (k largest values from Σ matrix and corresponding rows of U and columns of V^t), resulting in the smaller matrices U_k, Σ_k, V_k^t . In this way, the reduced matrix M_k is represented by

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

While the reduced matrix M_k keeps the same number of columns/dimensions as M , its rank is now k ($k < r$). The size k is a free parameter that we tune on the MEN development data (see Chapter 4). This mixed multimodal representation aims to find relationships between visual features and text terms, since both of them are aligned in the same feature matrix M . Several similar approaches using SVD were proposed by [26, 42], in which visual and text features called terms are aligned to generate a multimodal latent semantic representation for tackling the task of image retrieval.

After factorization, we split the reduced matrix M_k back into the original textual and visual blocks obtaining two matrices `Mixed_Text` and `Mixed_Image` that contain multimodally mixed text- and image-based representations of the target words.

3.4 Multimodal similarity estimation

This step is to estimate the similarity based on the multimodally mixed representations of the target words, that is, `Mixed_Text` and `Mixed_Image` matrices. In this thesis, we use linear weighted combination, one of the most widely used methods [43], in which the information obtained from different modalities is combined in a linear fashion.

Following previous researches [12, 23, 57], we estimate similarities for pairs of target words in two ways, at the feature level (**FL**) and at the scoring level (**SL**). Under the first strategy, the feature vectors of words in `Mixed_Text` and `Mixed_Image` matrices are normalized and concatenated, then used as a single representation, and similarity of pairs is estimated by comparing these multimodal representations. The latter approach estimates similarity of a pair independently in text and image space, then combines these two scores based on the degree of this similarity.

3.4.1 Feature-level estimation

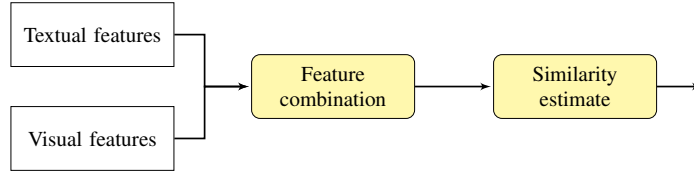


Figure 3.4: Feature-level similarity estimation

Figure 3.4 illustrates the FL strategy, where first we use the linear weighted combination method to build a single multimodal representation of words from textual and visual features, and then we estimate the similarity of pairs. The combination function is defined as

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

where F_t and F_v stand for text- and image-based feature vectors of target words and \oplus is vector concatenate operator. The parameter α with $\alpha \in [0, 1]$ is tuned on the MEN development data for the optimal accomplishment.

3.4.2 Scoring-level estimation

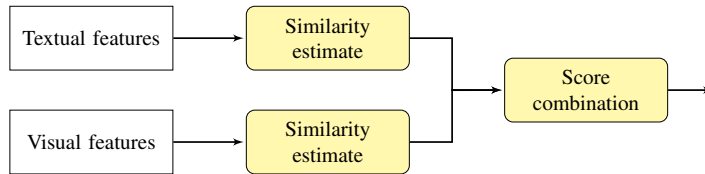


Figure 3.5: Scoring-level similarity estimation

Figure 3.5 describes the SL method, where text- and vision based (sub)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

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

where S_t and S_v stand for scores calculated based on textual and visual features, respectively. The parameter β with $\beta \in [0,1]$ is again calculated on the MEN development data.

3.5 General form and special cases

Given fixed and normalized text- and image-based matrices, our multimodal approach is parametrized by the size k of latent semantic space, α (the weight of text component in FL similarity estimate) and β (the weight of text component of SL). Moreover, when $k = r$, with r the rank of the original combined matrix, SVD returns the original combined matrix (no 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 Chapter 4). The simple approach of Bruni et al. (2011), in which the two matrices are concatenated without mixing, is the parametrization $k = r$, FL , $\alpha = 0.5$ (called **NaiveFL** model). The summing approach of Leong and Mihalcea (2011) corresponds to $k = r$, SL , $\beta = 0.5$ (called NaiveSL model). 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 $\beta = 0$ (these are the **Mixed_Text** and **Mixed_Image** models).

Reducing all these and other models to the same parametrized approach means that, giving development data for a specific task that requires similarity measurements, we can use these development examples to optimize the parameter vector. We can thus discover in a data-driven way which of the various models is the 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).

In this chapter, we present two different experiments, that is, semantic relatedness and concept categorization, in evaluating our proposed approach for multimodal fusion combining textual and visual information in distributional semantic models. In addition, we also provide a comparison between different image-based models on the integration into a textual model.

The first part of the chapter defines briefly the two evaluation tasks and then we describes our settings for those experiments specifying the evaluation datasets and computation measurements. For the semantic relatedness task, we evaluate our models on not only a widely used benchmark (WordSim353) but also a new dataset (MEN) wherein a subset of MEN is used to tune parameters. For the task of concept categorization, we use the Battig dataset for obtaining the most suitable parameters, then test our models on AP dataset (see Section 4.2).

In the latter part of this chapter, we report model parameter tuning and experimental results, then present our analyses. Both in the two experiments, we show that, by fixing the parameters on training data, we derive multimodal models that, for the first time, outperform state-of-the-art purely text-based models in a statistically significant manner, as well as surpassing some of previous approaches when trained on the same sources. The good performance of the multimodal models generalizes, moreover, to a different test set from the one used for parameter tuning on both tasks, demonstrating the robustness of our approach. The last experiment suggests that we might use better visual features in order to improve performance of the system (see Section 4.4).

4.1 Task definitions

To evaluate performance of our approach, we conduct experiments on two different tasks, that is, **semantic relatedness** and **concept categorization**. The former is the task of quantifying the degree in which some words or concepts are related, considering not only similarity but also semantic relationship among them. The latter is the task of classifying words or concepts into classes or categories.

Many applications, in Natural Language Processing (NLP) and other fields, benefit from calculating measures to determine numerically how semantically related two

words are. In the literature related to this topic, three main kinds of measures are considered by different authors or sometimes interchangeably by the same authors: semantic similarity, semantic relatedness and semantic distance. We adopt here the interpretation given in [10].

- Semantic similarity: It is defined by considering lexical relations of *synonym* (e.g. “car,automobile”) and *hypernymy* (e.g. “car,vehicle”)
- Semantic relatedness: It covers any kind of lexical or functional association such as meronymy (e.g. “finger,hand”), antonymy (e.g. “hot,cold”) or any kind of functional relationship or frequent association (e.g. “penguin,Antarctica”). So, it is considered as a more general concept than *semantic similarity*.
- Semantic distance: It is the inverse of *semantic relatedness*. The more two terms are semantically related, the more semantically close they are.

In this thesis, we use only the term of semantic relatedness and consider similarity as a special case of relatedness.

Categorization is the task of grouping words into classes or categories depending on their meaning similarities. It plays a prominent role in cognitive research on concepts and meanings as well as the ability to hierarchically arrange concepts into taxonomies [40]. Recently, research in corpus-based semantics focuses on investigating whether distributional similarity could be used to group words into semantically coherent categories [5]. In this thesis, we explore a wider aspect, that is, examining if visual information can help distributional semantics to improve on this task.

4.2 Experimental setup

4.2.1 Evaluation datasets

MEN It is a new evaluation benchmark¹, introduced by the author and his collaborators [8], with the express purpose of evaluating multimodal semantic models. It consists of 3,000 pairs of words, randomly selected from words that occur at least 700 times in the freely available ukWaC and Wackypedia corpora combined with the size of 1.9B and 820M tokens, respectively and at least 50 times as tags in the ESP image collection, with [0,1]-normalized semantic relatedness ratings. In order to avoid picking unrelated pairs only, we sampled the pairs so that they represent a balanced range of relatedness levels according to a text-based semantic score. The dataset was created by crowdsourcing, using the Amazon Mechanical Turk via the CrowdFlower interface (CF).² We presented CF workers with a list of two candidate pairs of words wherein each pair was randomly matched with 50 comparison pairs. Then, the CF workers were asked to judge which pair is more semantically correlated than the other. By this way, each pair was rated against 50 comparison pairs, thus obtaining a final score on a 50-point scale (then normalized between 0 and 1). Table 4.1 presents some examples of word pairs and their semantic ratings in MEN dataset.

¹<http://clic.cimec.unitn.it/~elia.bruni/MEN.html>

²<http://crowdflower.com>

word-1	word-2	score
car-n	automobile-n	1.00
canine-j	puppy-n	0.88
white-j	yellow-j	0.70
happy-j	sleep-v	0.44
feel-v	fun-n	0.32
chess-n	cigarette-n	0.06
bakery-n	zebra-n	0.00

Table 4.1: Examples of word pairs and their semantic ratings in the MEN dataset

We did not instruct the subjects about the difference between similarity and relatedness. We used only the second term and gave examples involving both similarity, that is a special case of relatedness (“car-automobile”) and relatedness (“wheels-car”). 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 technique based on control “gold standard” pairs ensures that only the data of speakers with a good command of English were retained.

WordSim353 It is a widely used benchmark proposed by Finkelstein and his colleagues (2002)[19].³ WordSim353 dataset (**WS**) was constructed by asking 16 subjects to rate directly a set of 353 word pairs on a 10-points similarity scale and averaging those ratings. We use the dataset in the format that contains POS information for each target word. Some examples of pairs of words and their ratings are shown in Table 4.2.

word-1	word-2	score
computer-n	keyboard-n	7.62
planet-n	galaxy-n	8.10
drink-n	car-n	3.04
country-n	citizen-n	7.31
tiger-n	animal-n	7.00
football-n	basketball-n	6.81

Table 4.2: Examples of word pairs and their semantic ratings in the WS dataset

We create MEN dataset aimed at enabling a larger-scale and more appropriate evaluation for multimodal semantic models than WS. MEN not only contains more pairs of words but also covers more word pairs’ types than WS that consists of mostly *noun-noun* word pairs except 3 *noun-adjective* and one *adjective-adjective* pairs (see Table 4.3). Moreover, we present a different way to obtain human judgements for these pairs of words. Unlike WS which asked subjects to give word pairs specific scores on a 10-points scale, we asked CF workers to give binary decision on which word pairs are more related. Psychologically, people find it simpler and more accurate

³<http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/>

to make binary decisions (which one of two word pairs is more related) than giving directly specific ratings on a scale.

Type	Percentage
noun-noun	67
verb-noun	10
noun-adj	17
verb-verb	1
adj-verb	2
adj-adj	3

Table 4.3: Word pairs’ types in MEN dataset

We used 2,000 items in the MEN dataset for model tuning and 1,000 items for evaluation. Our models cover 252 WS pairs, therefore, the correlations reported in the thesis are not directly comparable to those reported in other studied that used WS.

For the task of concept categorization, we use two well-known datasets: BATTIG and Almuhareb-Poesio (**AP**), in which the former is used for model tuning and the models are tested on the latter with tuned parameters. The task is to classify words (concept) into categories which is crucially important in both computer science and cognitive science. The aim of this experiment is to show potentially how the models perform on applying distributional information to other problems related to semantics [5].

BATTIG It is a small dataset, introduced by Baroni and his colleagues, 2010 [3] which classifies 82 concepts into 10 categories but based on battig’s norms. It contains only concrete basic-level concepts belonging to categories as shown in Table 4.4.

Concept	Category
aeroplane-n	vehicle
apple-n	fruit
bean-n	vegetable
bicycle-n	vehicle
birch-n	tree
blouse-n	clothes
bowl-n	kitchenware

Table 4.4: Examples of concepts and categories in BATTIG

AP It is a widely used concept categorization benchmark created by Almuhrdb and Poesio, 2006 [1] consisting of 402 concepts from WordNet balanced in terms of frequency and ambiguity. These concepts are categorized into 21 distinguished classes based on their similarity relatedness. Unlike BATTIG, AP contains not only concrete but abstract concepts as shown in Table 4.5.

In our experiment, we cover 77 concepts from 10 classes in BATTIG, and 231

Concept	Category
love-n	feeling
life-n	motivation
joy-n	feeling
headache-n	pain
monkey-n	animal
mortgage-n	assets
motorcycle-n	vehicle

Table 4.5: Examples of concepts and categories in AP

concepts belonging to 21 groups in AP. Even though we do not deal with full coverages of two datasets, what we obtained is fairly distributed in all categories from the original datasets.

4.2.2 Similarity and correlation measures

For the semantic relatedness task, in order to measure the similarity between two words, we use the widely used Cosine computation that provides a fair comparison with other approaches. Given two vectors of attributes, X and Y wherein $X = \{x_1, x_2, \dots, x_n\}$ and $Y = \{y_1, y_2, \dots, y_n\}$, the cosine similarity θ is represented using a dot product and magnitude as

$$\theta = \cos(X, Y) = \frac{\sum_{i=1}^n x_i * y_i}{\sqrt{\sum_i x_i^2} \times \sqrt{\sum_i y_i^2}}$$

To compare the performances of our models to human ratings, we use the Spearman’s rank correlation coefficient. For a sample of size n , the n scores X_i, Y_i are converted to ranks x_i, y_i , then the Spearman coefficient ρ is computed as:

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

For the task of concept categorization, following the original proponents, we compute the *purity* score by using CLUTO toolkit [28]. The purity score is defined as follows: if n_r^i is the number of items from i -th true (gold standard) class that were assigned to the r -th cluster in which n is the number of items and k is the number of cluster, then:

$$Purity = \frac{1}{n} \sum_{r=1}^k \max_i(n_r^i)$$

The higher the purity score is the more quality clusters have. In the best case, clusters reach to the purity of 100% and as cluster quality deteriorates, purity approaches 0.

4.2.3 MixLDA

In order to compare to Feng and Lapata’s approach called MixLDA, we reimplemented it using description keywords for comparability. First, we create a joint matrix that mixes textual and visual information from the ESP dataset. For each image, we extract the image-based features (via BoVW as discussed above) and use the words labelling 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 labelling textual word or a visual word extracted from that image. We obtain a matrix of size $90\text{K} \times 100\text{K}$, with 10K textual words, 80K visual words and 100K documents (images). The Latent Dirichlet Allocation model is trained on this matrix and tuned on the MEN development data by varying the number of topics K_t .⁴ Under MixLDA, each target word in an evaluation set is represented by the vector giving its distribution over the K_t latent topics.

4.3 Model selection

The multimodal data representation has several parameters that must be instantiated. These include the size k of latent space and the values for α and β . We explored the parameter space on held-out data in which given development data for a specific task, the parameters are instantiated from these development examples. Specifically, we tune the parameters for the semantic relatedness and concept categorization models separately using 2,000 items in the MEN and BATTIG datasets, respectively.

We varied increasing values of k as powers of 2, starting from a very small one $2^4 = 16$ to a large one $2^{12} = 4096$. We also varied the weighting parameters α and β from 0 to 1 inclusively with step 0.1. As discussed above in Section 3.1, the framework consists of two steps, that is, latent multimodal mixing and multimodal similarity estimate, where the latter takes the two matrices resulting from the former as its input. Therefore, for each value of k in the first step, we tuned 11 values for each α or β depending on which estimate strategies (FL or SL) is used. Consequently, we created 11 multimodal models for each k and strategy, and totally 99 models for each strategy.

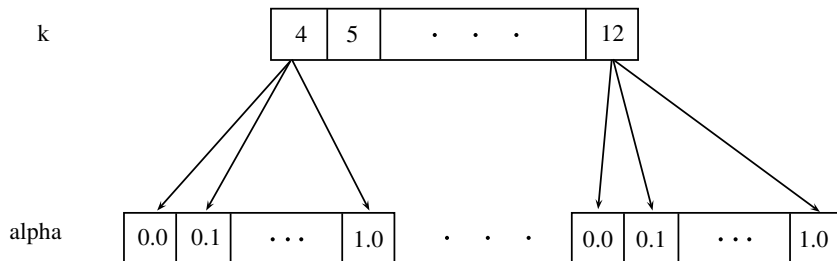


Figure 4.1: Parameter search space for the FL strategy (k in terms of \log_2)

In this thesis, we report the tuning phase in the following fashion. In the FL or SL strategy, for each k , we first tune α or β from 0 to 1 and store the best model in this

⁴LDA was computed with Gensim: <http://radimrehurek.com/gensim/>

manner (we called it as *local tuning*). We end up with 9 best models according to each k after the local tuning phase. In order to obtain globally the optimal parameters, we compare these models and retain the model gaining the best result. These optimal parameters are then used to evaluate on the testing data.

4.3.1 Semantic relatedness

For this task, we use 2,000 items in the MEN dataset for model tuning, then evaluate our model on 1,000 remaining word pairs in MEN and WS datasets with the tuned parameters. Following the fashion described above, we first vary the values of α and β in the FL and SL strategies, respectively for each k . Figure 4.2 presents an example with $k = 2^8$ and the use of Window20 as textual source.

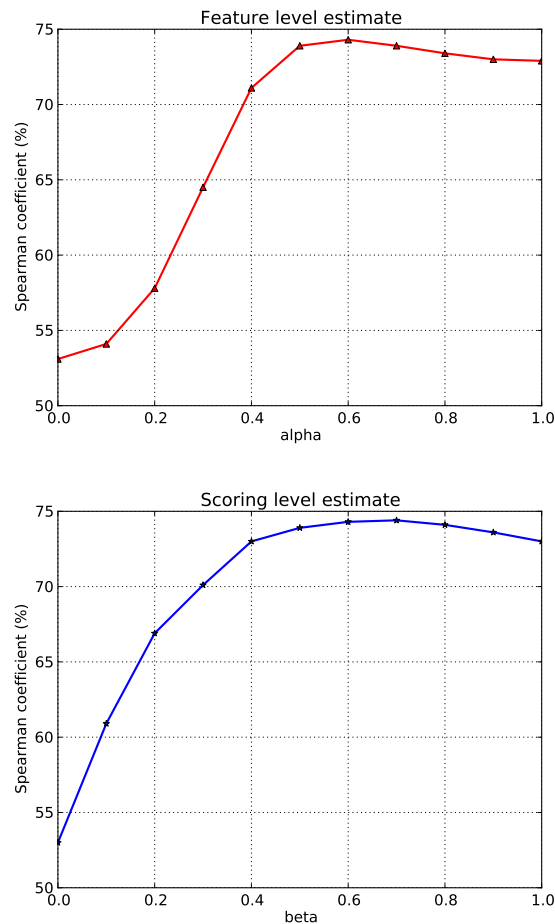


Figure 4.2: Performance of the models with Window20 as text-based resource by varying α and β parameters and fixing the latent dimension $k = 2^8$

In Figure 4.2, after the local tuning computation, we obtain the best models with $\alpha = 0.6$ with FL and $\beta = 0.7$ with SL for $k = 2^8$. We explore the parameters in the similar fashion for other values of k and end up with locally best models for each k as shown in Table 4.6

k	4	5	6	7	8	9	10	11	12
α	0.7	0.7	0.7	0.6	0.6	0.5	0.5	0.5	0.5
β	0.8	0.8	0.8	0.8	0.7	0.6	0.5	0.5	0.5

Table 4.6: The best values of α and β for each latent dimension (in terms of \log_2)

These local best models are compared in terms of k as illustrated in Figure 4.3 in order to achieve the global best model.

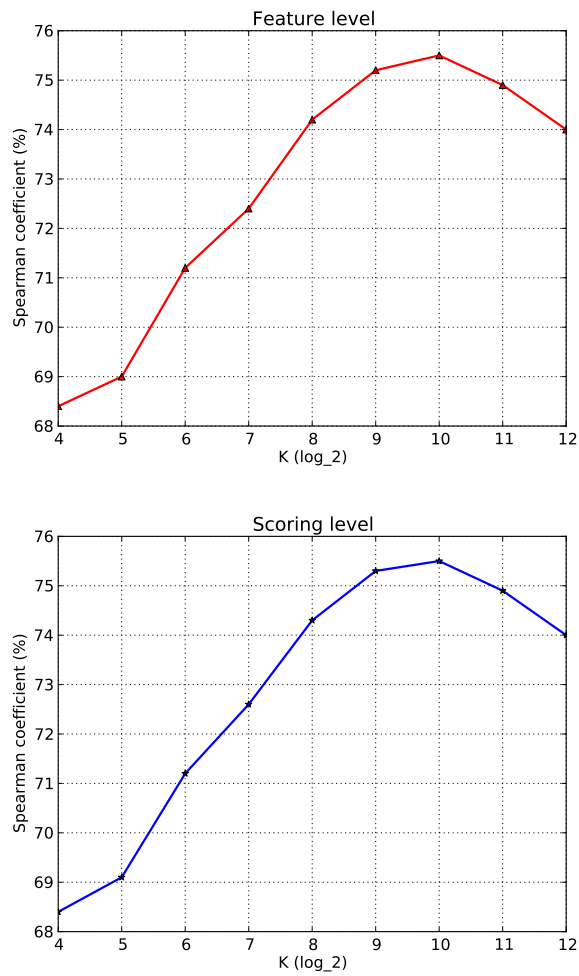


Figure 4.3: Performance of the models with Window20 as text-based resource by varying the number of latent dimensions

Table 4.7 shows the optimal parameters we obtain for the models with different text-based resources (Window2 and Window20). These values are then used to perform on the testing data.

Models	Latent dimension	α	β
Window2	9	0.7	0.8
Window20	10	0.5	0.5

Table 4.7: The best values of α , β and latent dimension (in terms of \log_2)

4.3.2 Concept categorization

In this task, we use Battig dataset for model tuning then used tuned parameters to evaluate our models on AP dataset. For both sets, we cluster the words (concepts) based on their pairwise cosines in the semantic space defined by a model using the CLUTO toolkit. We first calculate cosine scores between every possible pair of words in the dataset, creating a similarity matrix which is then used as the input for CLUTO’s scluster algorithm. We use CLUTO’s built-in repeated bisections with global optimization method (*rbr*) accepting all of CLUTO’s default values in order to make our results comparable to previous approaches.

As the line of model tuning in semantic relatedness experiment, we combine every possible values of the size k of latent semantic space and weighting parameters α , β to create different numbers of semantic space that might be used to compute the similarity matrix. The tuned parameters obtaining the highest Spearman correlation on BATTIG are retained to evaluate on AP dataset. Because this procedure is similar to that in the first experiment, we do not describe it explicitly here. Table 4.8 reports the optimal parameters for the models using Window2 and Window20 as textual resources, respectively.

Models	Latent dimension	α	β
Window2	7	0.8	0.9
Window20	9	0.5	0.5

Table 4.8: The best values of α , β and latent dimension (in terms of \log_2)

4.4 Experimental results

4.4.1 Semantic relatedness

Table 4.9 reports the correlations on MEN (the 1,000 pairs not used for parameter tuning) and WS datasets with the use of Window2 and Window20 as textual models, respectively. The first two rows show results of the original text- and image-based models, before any mixing. Text shows comparable performances on both datasets. 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 results of the previous approaches we took into consideration [9, 18, 33]. While the NaiveFL 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

<i>Model</i>	Window 2		Window 20	
	<i>MEN</i>	<i>WS</i>	<i>MEN</i>	<i>WS</i>
Text	0.73	0.70	0.68	0.70
Image	0.43	0.36	0.43	0.36
NaiveFL	0.75	0.67	0.73	0.67
NaiveSL	0.76	0.69	0.74	0.64
MixLDA	0.30	0.23	0.30	0.23
Mixed_Text	0.77	0.73	0.74	0.75
Mixed_Image	0.55	0.52	0.57	0.51
FL	0.78	0.72	0.76	0.75
SL	0.78	0.71	0.77	0.72

Table 4.9: Spearman correlation of the models on MEN and WordSim (all coefficients significant with $p < 0.001$).

NaiveSL (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. This is probably due to the fact that Feng and Lapata’s approach is constrained to using the same source for the textual and the visual model and our image dataset is a poor source of textual data. Note in any case that the reported MixLDA coefficients are comparable to those obtained in the original Feng and Lapata experiments.

Analysing now the effects of our fusion strategies, we can first see a uniform enhancement on both MEN and WS for Mixed_Text and Mixed_Image. Mixed_Text 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 [39], with $p < 0.01$. Looking then at the similarity estimate combination strategies, both SL and FL equal or do better than any other model. Not only they outperform Text on MEN with high statistical significance ($p < 0.001$), but they are significantly better than Mixed_Text on MEN with Window20 ($p < 0.01$). Between the two approaches, combining the features before measuring similarity (FL) seems more fruitful than combining similarity estimates (SL), when we look at the overall result patterns.

It is worth remarking in any case that while Mixed_Text is a bit worse than the combined models, it still achieves high correlations with the human judgments and it has an extremely high correlation with the best model FL ($\rho = 0.98$). This suggests that most of the benefits of multimodality are already captured by latent mixing. Note that Mixed_Text is an attractive model because it has less parameters than the whole pipeline and it is more compact than FL, since it discards the visual features.

To acquire a more qualitative view of how multimodality is contributing to meaning representation, we first picked the top 200 most similar pairs from the combined human norms, so that we would be confident that they are pairs for which similarity should be high, and then we looked, within this subset, at those pairs with the

most pronounced difference in cosines between Text and FL, with Window20 as our textual source. That is, the first column of Table 4.10 gives pairs that are considered very similar by humans and where similarity was better captured by Text, the second column pairs where similarity was better captured by FL.

Text	FL
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 4.10: Top 10 pairs whose similarity is better captured by Text vs. FL.

We can notice that 7/10 of the relations better captured by FL 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 we can suppose to 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 similarities, such as synonymy and coordination.

4.4.2 Concept categorization

Table 4.11 reports percentage purities on the AP dataset for the models with the use of Window2 and Window20 as textual model, respectively. The first two rows show purities with original text- and image-based models before any smoothing. Even though Image performs significantly worse than Text, it does capture some aspects of meaning. This is in accordance with what was found in Bruni et al.’s study.

While the NaiveFL approach (analogous to Bruni et al.’s method) obtains slightly better result with Window2, it attains lower performance with Window20. Thus, generally speaking, adding visual features by simply concatenating without smoothing does not significantly improve performance. In the inverse manner, NaiveSL (equivalent to Leong and Mihalcea’s approach) loses several purity points with Window2 but

Model	Window2	Window20
Text	73	65
Image	26	26
NaiveFL	74	64
NaiveSL	65	66
Mixed_Text	74	67
FL	74	69
SL	75	69

Table 4.11: Percentage purities of the models on AP dataset

gains inappreciably improvements with Window20. Again, combing original textual and visual information at the decision level does not obtain considerably enhancement in this task.

The last three rows report the results of our fusion strategies. Mixed_Text shows improvements compared to Text for the models with the use of both textual models in which with Window20 we obtain slightly better purity difference (2%) than that with Window2 (1%). It is the first one showing enhancement with both text-based models. Analysing now the multimodal similarity estimate strategies, we can see that both FL and SL equal or perform better than any other model. They not only outperform Text (2% and 4% improvement with Window2 and Window20, respectively), but they gain considerably better results than NaiveFL and NaiveSL. Compared to Mixed_Text, they are also significantly better with both Window2 and Window20.

In short, in the task of concept categorization, mapping combined text- and image-based features onto a lower dimensional semantic space and then combining these information with different multimodal estimate strategies also performs better than purely textual models and some previous methods performing on original features.

4.4.3 Image-based model selection

This experiment aims to see the effect of different image-based features combined with a textual model into a multimodal architecture. We suppose that the better visual features we have the better multimodal models we obtain. To achieve this purpose, these models are examined on the task of semantic relatedness.

In order to compare different visual features, we follow our pipeline framework taking Window20 as text-based resource and image-based features using HSV- or RGB-based SIFT descriptors, respectively. Table 4.12 shows the performance of these two visual models on MEN and WS. From the table, Image₁ (HSV-based) obtains better results than Image₂ (RGB-based); thus we expect that Window20 performs better when combined with Image₁ than with Image₂.

Let Multi-modal₁ and Multi-modal₂ be multimodal models with the usage of Image₁ and Image₂ as image-based models, respectively. The parameters (k , α , β) in the fashion we described are first tuned on 2,000 items in the MEN dataset. The models with the tuned parameters obtaining the best correlations are used to perform on the remaining items in MEN and WS datasets.

Model	MEN	WS
Image ₁	0.44	0.36
Image ₂	0.39	0.33

Table 4.12: The performance of different visual models

Table 4.13 presents Spearman correlations of Multi-modal₁ and Multi-modal₂ on MEN and WS, respectively. As we expected, the Multi-modal₁ model obtains significant higher results than Multi-modal₂ on the WS dataset (5% and 2% improvement with FL and SL estimate strategies, respectively). Similarly, on the MEN dataset, Multi-modal₁ is also slightly better than Multi-modal₂.

<i>Model</i>	Image ₁		Image ₂	
	<i>MEN</i>	<i>WS</i>	<i>MEN</i>	<i>WS</i>
FL	0.76	0.75	0.76	0.70
SL	0.77	0.72	0.76	0.70

Table 4.13: Spearman correlations of the models on MEN and WS (all coefficients significant with $p < 0.001$).

The results inspire us to exploit more effective methods for extracting visual features from images. Moreover, multimodal models still obtains good performance across different image models, showing robustness of the approach.

4.4.4 Result summary

To sum up, we have evaluated our multimodal models on the tasks of semantic relatedness and concept categorization. For the former task, we conduct experiments on MEN and WS in which one part of MEN was used for model tuning. Our best models perform significantly better than purely text-based models wherein obtaining an improvement of 5% and 2% on MEN and WS with Window2, 9% and 5% on MEN and WS with Window20. Moreover, our proposed approach shows considerably improvements both on MEN and WS compared to previous approaches (Bruni et al., 2011; Fend and Lapata, 2010; Leong and Mihalcea, 2011). For the latter task, we first tune model parameters on Battig dataset then used these tuned parameters to evaluate our models on AP dataset. While previous methods can obtain enhancement on AP compared one purely textual model, they lose several purity points compared to another text-based model. Our fusion strategies combining textual and visual information at the feature level and scoring level after mixing, that is, FL and SL, outperform any other method. Compared to purely textual models, they gain 2% and 4% improvement with Window2 and Window20, respectively.

More interestingly, Mixed_Text achieves high results on both tasks even though it is a bit worse than the combined models. This suggests that most of the benefits of multimodality are captured by latent mixing.

Furthermore, we also show that a better quality of image-based features can im-

prove the performance of the system.

Overall, we show that by fixing the parameters on training data, the multimodal models we derived outperform state-of-the-art purely textual models, as well as surpassing some of previous methods when trained on the same sources. Moreover, by performing model tuning on the training sets of a rather different kind than the corresponding test sets, we demonstrate the robustness of our approach.

5.1 Thesis achievements

The integration of image-based features into textual models has recently attracted a lot of attention. In this thesis, we have given an overview of recent works in this field, then compared them and discussed some of their limitations. The main objective of this study is to tackle this problem in an adaptable and effective way.

We have proposed and implemented a general and flexible architecture, that not only encompasses all previously proposed strategies in a single parametrized pipeline, but can automatically choose the best model for a specific task given development data. The framework consists of two steps, namely latent semantic mixing and multimodal semantic estimation. The former aims to map both text- and image-based features onto a lower-dimensional semantic space in order to leverage the overlapping information between these two modalities. The two channels are then combined in a linear weighted fashion at the feature and scoring levels. The flexible aspect of the framework is illustrated by the way in which we can use different techniques in each step, for example Singular Value Decomposition or Non-negative Matrix Factorization might be used in the latent semantic mixing phase. In addition, the previous methods are shown as special cases of our framework by setting the model parameters appropriately. To the best of our knowledge, we propose what is the most current general framework in this field.

We demonstrated in various experiments that simple low-level visual features can enrich the semantic representation of word meaning with information that cannot be extracted from text alone. Additionally, this approach obtains a significant enhancement over all previous proposed methods on both the task of estimating degrees of semantic relatedness and clustering concepts into categories. Furthermore, a good quality of visual features might lead to improvement of multimodal models.

Last but not least, we have created the new MEN benchmark to evaluate the similarity of word pairs. The MEN dataset contains not only more items but also more varied typology of words than WordSim-353. We hope it will become a widely used benchmark in this field.

5.2 Future work

In future work, we will focus, first of all, on improving performance by examining different mixing techniques, such as Non-negative Matrix Factorization, that can be used in the latent semantic mixing step. In addition, another promising route will be that of assigning different weights to the textual and visual components on a word-by-word basis, instead of doing it globally, for example depending on an automatically determined imageability score to be assigned to each word (intuitively, visual information might be assigned a larger weight for concrete rather than abstract words).

Exploiting more effective techniques in computer vision community to extract better visual features from collections of images might be another direction for future work. Moreover, in the quality of creating a collection with high quality images might bring benefits to extracted features. In this thesis, we use the ESP image dataset where it contains good textual annotations but quite low-quality images. Therefore, we are creating a new image dataset (CleanedFlickr) with the purpose of enhancing quality of visual features.

Last but not least, we will evaluate multimodal models in more semantic tasks and applications, investigating in particular whether the measures of semantic relatedness they provide are helpful for the analysis of text.

Appendix A

Usage of the Modules Developed in this Thesis

A.1 Download link

The implementation of all modules in the FUSE project can be downloaded from <https://github.com/s2m/FUSE>. Modules developed in this thesis are described in detail in next subsections. The MEN dataset is available from <http://clic.cimec.unitn.it/~elia.bruni/MEN.html>.

A.2 Description of each module

In this thesis, we developed various modules for the implementation of the framework, namely latent semantic mixing and multimodal similarity estimation. Besides that, some modules in evaluation are also presented. All of these modules are written in Python.

A.2.1 Latent semantic mixing

This module contains two submodules and can be downloaded from <https://github.com/s2m/FUSE/tree/master/core/src/fusion>.

The first module is Decomposer:

- Description: Parse the blended matrix M into three matrices U , S , V_t using Singular Value Decomposition.
- Library required: sparsesvd in <http://pypi.python.org/pypi/sparsesvd>
- Input:
 - The blended matrix M in the following format:
target_word value1 value2 ... valueN
 - Four files to store list of target words, U-matrix, S-matrix and V_t -matrix
- Output: Output matrices in files.

The second module is Mixer:

- Description: Map the concatenated matrix onto a lower-dimensional space by keeping k largest singular values in S-matrix.
- Input:
 - Four files obtained from Decomposer module
 - The value of the size k
 - The file to store the output matrix
- Output: The matrix with the same columns as the blended one but with a lower rank k .

A.2.2 Multimodal similarity estimation

This module can be found from https://github.com/s2m/FUSE/tree/master/core/src/fusion/multimodal_similarity_estimation, containing two submodules.

The first one is Feature_Combiner:

- Description: Combine the features of two modalities using a linear weighted function.
- Input:
 - Text- and image-based feature matrices
 - α - weighting parameter
 - The file to store combined features
- Output: Combined matrix in the file

The second module is Score_Combiner:

- Description: Combine two modalities at the scoring level using a weighted combination function.
- Input:
 - Text- and image-based scores
 - β - weighting parameter
 - The file to store combined scores
- Output: Combined scores in the file

A.2.3 MEN dataset

This module can be downloaded from <https://github.com/s2m/FUSE/tree/master/eval/src/MEN-dataset>, consisting of various modules to create the MEN dataset. They includes the preparation for crowdsourcing job and the extraction of results.

A.2.4 Evaluation

The implementation of evaluating the approach can be downloaded from <https://github.com/s2m/FUSE/tree/master/eval/src>. Some utilities, such as computing cosine score between two words and computing Spearman coefficient, are also found from that link.

Bibliography

- [1] Abdulrahman Almuhareb. *Attributes in Lexical Acquisition*. PhD thesis, PhD thesis, University of Essex, 2006.
- [2] Mark Andrews, Gabriella Vigliocco, and David Vinson. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498, 2009.
- [3] 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.
- [4] Marco Baroni and Alessandro Lenci. Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88, 2008.
- [5] Marco Baroni and Alessandro Lenci. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.
- [6] Tamara Berg, Alexander Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy Web data. In *ECCV*, pages 663–676, 2010.
- [7] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image Classification using Random Forests and Ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007.
- [8] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the ACL 2012*, South Korea, 2012.
- [9] 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, 2011.
- [10] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32, 2006.

- [11] John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526, 2007.
- [12] Marcel Worring Cees G. M. Snoek and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.
- [13] 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, 2004.
- [14] Jia Deng, Wei Dong, Richard Socher, Lia-Ji Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [15] Stefan Evert. *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University, 2005.
- [16] 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.
- [17] 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*, 2010.
- [18] Yansong Feng and Mirella Lapata. Visual information in semantic representation. In *Proceedings of HLT-NAACL*, pages 91–99, Los Angeles, CA, 2010.
- [19] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppín. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.
- [20] J. R. Firth. A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*, pages 1–31, 1957.
- [21] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *In ICCV*, pages 1458–1465, 2005.
- [22] Kristen Grauman and Bastian Leibe. *Visual Object Recognition*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.
- [23] David L. Hall and James L. Llinas. An introduction to multisensor data fusion. In *Proceedings of the IEEE: Special Issues on Data Fusion*, pages 6–23, 1997.
- [24] Z. Harris. Distributional structure. In *Papers in structural and transformational linguistics*, pages 775–794, 1970.

- [25] Luis Enrique Sucar Hugo Jair Escalante, Carlos A. Hrnadez and Manuel Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 172–179, 2008.
- [26] Peter G.B. Enser Jonathon S. Hare, Paul H. Lewis and Christine J. Sandom. A linear-algebraic technique with an application in semantic image retrieval. In *Proceedings of the 5th international conference on Image and Video Retrieval*, pages 31–40, 2006.
- [27] Fabio A. Gonzlez Juan C. Caicedo, Jaafar BenAbdallah and Olfa Nasraoui. Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. *Neurocomputing*, 76:50–60, 2012.
- [28] Geogre Karypis. Cluto: A clustering toolkit. Technical report, Technical Report 02-017, University of Minnesota Department of Computer Science, 2003.
- [29] David Forsyth Nando de Freitas David M. Blei Kobus Barnard, Pinar Duygulu and Michael I. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003.
- [30] 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*, 2011.
- [31] Thomas K. Landauer and Susan T. Dumais. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.
- [32] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR 2006, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.
- [33] 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, Chiang Mai, Thailand, 2011.
- [34] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, 60:91–110, 2004.
- [35] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentations, and Computers*, 28:203–208, 1996.
- [36] Chris Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.

- [37] Dian I. Martin and Michael W. Berry. *Mathematical Foundations Behind Latent Semantic Analysis*. Handbook of latent semantic analysis, 2007.
- [38] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6:1–28, 1991.
- [39] David Moore and George McCabe. *Introduction to the Practice of Statistics*. Freeman, New York, 5 edition, 2005.
- [40] Gregory Murphy. *The Big Book of Concepts*. MIT Press, Cambridge, MA, 2002.
- [41] 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.
- [42] 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.
- [43] Abdulmotaleb El Saddik Pradeep K. Atrey, M. Anwar Hossain and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Springer Multimedia Systems Journal*, pages 345–379, 2010.
- [44] Reinhard Rapp. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the 9th MT Summit*, pages 315–322, New Orleans, LA, 2003.
- [45] 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.
- [46] Jia Li Ritendra Datta, Dhiraj Joshi and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 2008.
- [47] Magnus Sahlgren. *The Word-Space Model*. Dissertation, Stockholm University, 2006.
- [48] Diarmuid O Seaghdha. Distributional approaches to semantic analysis. *Lectures in HIT-MSRA Summer Workshop on Human Language*, 2011.
- [49] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.
- [50] Peter W. Foltz Thomas K. Landauer and Darrell Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.
- [51] Mark Steyvers Thomas L. Griffiths and Joshua B. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114:211–244, 2007.
- [52] Richmond H. Thomason. What is semantics? *Lectures at University of Michigan*, 1996.

- [53] Peter Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.
- [54] Jasper R. R. Uijlings, Arnold W. M. Smeulders, and Remko J. H. Scha. Real-time Visual Concept Classification. *IEEE Transactions on Multimedia*, 12, 2010.
- [55] Andrea Vedaldi and Brian Fulkerson. Vlfeat – an open and portable library of computer vision algorithms. In *Proceedings of the 18th annual ACM international conference on Multimedia*, 2010.
- [56] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326, Vienna, Austria, 2004.
- [57] C-YLin M. R. Naphade C.Neti H.J.Nock W. H. Adams, G.Iyengar and J.R.Smith. Semantic indexing of multimedia content using visual, audio and text cues. In *EURASIP J. Appl. Signal Process*, pages 170–185, 2003.
- [58] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Multimedia Information Retrieval*, pages 197–206, 2007.