Charles University in Prague
Faculty of Mathematics and Physics

Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen · Bolzano

Fakultät für Informatik | Facoltà di Scienze e Tecnologie informatiche | Faculty of Computer Science

MASTER THESIS

Giang Binh Tran

Combining of text-based semantics and vision-based semantics

Faculty of Mathematics and Physics, Charles University in Prague
& Faculty of Computer Science, Free University of Bozen Bolzano

Supervisors of the master thesis: Dr. Martin Holub, Dr. Marco Baroni
& Dr. Raffealla Bernardi

Study programme: *Erasmus Mundus Master in Language and Communication Technology*
Specialization: *Computational Linguistics*
Prague 2011

to me & my family & my love,

and to all who helped me make this thesis work

Acknowledgments

Foremost, I would like to give the deepest thank to my supervisors, Dr. Marco Baroni, Dr. Raffaella Bernardi and Dr. Martin Holub, for their guidance. They always appear when I need help, not only in thesis work but also in many aspects of life. I would admit that without their support, the thesis would not have been possible.

I am very grateful to the Erasmus Mundus program in Language and Communication Technology (EM-LCT) program which provides me financial supports for my studies.

I would express my gratitude to following people:

- Jan Hajic, Vladislav Kubon and Marketa Lopatkova and others from UFAL, Charles University in Prague.
- Valeria Fionda and the International Relations staffs of Free University of Bozen Bolzano.
- Elia Bruni from CLIC, CiMec Research Centre, University of Trento.

I sincerely acknowledge the BIT Program which provides me a chance to go to University of Trento and CiMec Research Centre to complete one interesting semester of my studies.

Perhaps I would haven't finished my master study without supports of the Institute of Applied Linguistics, Charles University in Prague; Faculty of Computer

Science, Free University of Bozen Bolzano and Language - Interaction and Computation Laboratory, University of Trento. I would like to give my honest appreciation for their kindly help.

I would like to thank Jasper Uijlings from DISI, University of Trento and my colleagues in the EM-LCT program for their interesting discussions and advices in image processing and NLP.

Rovereto, June, 2011.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague - July 10, 2011 **Giang Binh Tran**

Title: *Combining text-based semantics and vision-based semantics*

Author: **Giang Binh Tran**

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague

Supervisors of the master thesis:

Dr. Martin Holub, *Faculty of Mathematics and Physics, Charles University in Prague*

Dr. Marco Baroni, *CiMec Research Centre, University of Trento*

Dr. Raffaella Bernardi, *Faculty of Computer Science, Free University of Bozen Bolzano*

Abstract: Learning and representing semantics is one of the most important tasks that significantly contribute to some growing areas, as successful stories in the recent survey of Turney and Pantel (2010). In this thesis, we present an innovative (and first) framework for creating a multimodal distributional semantic model from state of the art text-and image-based semantic models. We evaluate this multimodal semantic model on simulating similarity judgements, concept clustering and the newly introduced BLESS benchmark. We also propose an effective algorithm, namely Parameter Estimation, to integrate text- and image-based features in order to have a robust multimodal system. By experiments, we show that our technique is very promising. Across all experiments, our best multimodal model claims the first position. By relatively comparing with other text-based models, we are justified to affirm that our model can stay in the top line with other state of the art models.

We explore various types of visual features including SIFT and other color SIFT channels in order to have preliminary insights about how computer-vision techniques should be applied in the natural language processing domain.

Importantly, in this thesis, we show evidences that adding visual features (as the perceptual information coming from images) is comparable (and possibly better) than adding further text features to the advanced text-based model; and more interestingly, the visual features can capture the semantic characteristics of (especially concrete) concepts and they are complementary with respect to the characteristics captured by textual features..

Keywords: semantics, vision-based, combination, image processing

Contents

Summary

Distributional semantic models use large text corpora to derive estimates of semantic similarities between words. The basis of these procedures lies in the hypothesis that semantically similar words tend to appear in similar contexts (Miller and Charles, 1991; Wittgenstein, 1953). For example, the meaning of spinach (primarily) becomes the result of statistical computations based on the association between spinach and words like plant, green, iron, Popeye, muscles. Alongside their applications in NLP areas such as information retrieval or word sense disambiguation (Turney and Pantel, 2010), a strong debate has arisen on whether distributional semantic models are also reflecting human cognitive processes (Griffiths et al., 2007; Baroni et al., 2010). Many cognitive scientists have however observed that these techniques relegate the process of meaning extraction solely to linguistic regularities, forgetting that humans can also rely on non-verbal experience, and comprehension also involves the activation of non-linguistic representations (Barsalou et al., 2008; Glenberg, 1997; Zwaan, 2004). They argue that, without grounding words to bodily actions and perceptions in the environment, we can never get past defining a symbol by simply pointing to covariation of amodal symbolic patterns (Harnad, 1990). Going back to our example, the meaning of spinach should come (at least partially) from our experience with spinach, its colors, smell and the occasions in which we tend to encounter it. We can thus distinguish two different views of how meaning emerges, one stating that it emerges from association between linguistic units reflected by statistical computations on large bodies of text, the other stating that meaning is still the result of an association process, but one that concerns the association between words and

perceptual information.

In this thesis, we try to make these two apparently mutually exclusive accounts communicate, to construct a richer and more human-like notion of meaning.

We present an innovative (and first) framework for creating a multimodal distributional semantic model from state of the art text-and image-based semantic models. We evaluate this multimodal semantic model on simulating similarity judgements, concept clustering and the newly introduced BLESS benchmark. We also propose an effective algorithm, namely Parameter Estimation, to integrate text- and image-based features in order to have a robust multimodal system. By experiments, we show that our technique is very promising. Across all experiments, our best multimodal model claims the first position. By relatively comparing with other text-based models, we are justified to affirm that our model can stay in the top line with other state of the art models.

We explore various types of visual features including SIFT and other color SIFT channels in order to have preliminary insights about how computer-vision techniques should be applied in the natural language processing domain.

Importantly, in this thesis, we show evidences that adding visual features (as the perceptual information coming from images) is comparable (and possibly better) than adding further text features to the advanced text-based model; and more interestingly, the visual features can capture the semantic characteristics of (especially concrete) concepts and they are complementary with respect to the characteristics captured by textual features.

Published papers:

Partial results of this thesis are published as:

††Giang Binh Tran and Elia Bruni and Marco Baroni. 2011. Convergence of text-based and vision-based semantics, Social Media Retrieval Summer School, Poster section, Antalya - Turkey, June.

††Elia Bruni and Giang Binh Tran and Marco Baroni. 2011. Distributional semantics from text and images. EMNLP - GEMS Workshop Edinburgh - UK, July.

Chapter 1

Introduction

Recently, computer scientists and engineers have obtained good enhancements in computer speed as well as storage power. However, the problem of building a computer that can independently think and give decision is still a big challenge. One of the insightful ideas mentioned in RSA 2008 (Jeff Hawkins, 2008) [40] is to apply neuron science into computer architecture to emulate people's brain. In spite of the fact that it is very complicated, it is not impossible. With current technology, we may think of this type of machine with languages learning ability that is comparable to human learners. Since language is one of the most powerful communicative means of human being, it should be worth exploring this topic.

What is Semantics? The problem of how to teach computer to understand natural language is often referred as Semantics studies, especially how to represent natural language in computer (i.e. semantic representation).

"Semantics is the study of the meaning of linguistic expressions. The language can be a natural language, such as English or Navajo, or an artificial language, like a computer programming language. Meaning in natural languages is mainly studied by linguists. In fact, semantics is one of the main branches of contemporary linguistics. Theoretical computer scientists and logicians think about artificial languages. In some areas of computer science, these divisions are crossed. In machine translation, for instance, computer scientists may want to relate natural language texts to abstract representations of their meanings; to do this, they have to design artificial languages for representing meanings." [71]

Children's language acquisition Linguistics and psychologists try to understand how children learn the meanings of words. Presenting one of the most important studies about that problem, Bloom (2000) answered many questions to make clear the language acquisition process in children.

Learning a word involves mapping a form, such as the sound dog, onto a meaning or concept, such as the concepts of dogs [10]

Among his analysis, some conclusions are:

- Children can grasp aspects of the meaning of a new word on the basis of a few incidental exposures, without any explicit training or feedback - in fact, even without any explicit act of naming
- Learning the meanings of words is not qualitatively different from learning facts about the world
- Children make the connection between words and what they refer to through their understanding of the referential intentions of others: they use their theory of mind to learn the meanings of words
- No word meaning can be learnt entirely through syntactic cues: syntax is certainly an important informational source as to the meanings of words, but it must be integrated with information obtained from other inferential mechanisms

Obviously, children are not able to access the Internet or dictionary to find new word definitions and concepts. They learn the language from the environment, which is where they can figure out the meaning of such words. They may require some support from adults/their parents, such as showing some hints or repeating words in different situations.

More specifically speaking, with some supervision they can learn new words from related visual cues, motions and old words they already know. These things bring us an idea: children, with no knowledge from the beginning, can acquire new language skills from the environment. This is the method computers should use. More generally speaking, we can teach a computer a new language by supervising it, giving it some basic language knowledge and a very rich perceptual environment.

Some of the latest cognition studies show that the information of meaning in a human brain, which is obtained by the cognitive system, is affected by experience and language (Andrews and Vigliocco; Andrews et al, 2009 & 2010) [1, 2, 3, 4]. In other words, information formed from perception and interaction with the environment (like emotional information) plays an important role in learning and representing word meaning, or semantic representation.

Language itself also gives compelling information for learning semantic representation across domains. In addition, linguistics studies show that the description of objects may loses its accuracy if it is only extracted from the text without consideration to the other perceptual information [6]. For example, *red light* may refer to the *red* color but does not keep the same meanings in *red light street*. In another example, "green banana" suggests a banana that is not ripe but it also suggests a banana that is visually green; "black man" refers to a man having dark skin as well as a man in black clothes. That shows us color words also convey non-linguistic information [62]. Vice-versa, non-color words, for example "ocean" in "ocean eyes", "sunflower" or "rose", is also able to transmit visual information. On another hand, description context effects owe boundlessness in natural language understanding. To take an illustration, the word "behind" expresses different meaning aspects in "clean behind the couch" and "hide behind the couch" [62]; more examples can be found in Roy and Reiter (2005). Thus, it would be interesting and useful if we use all of perceptual information such as information from colors rather than just text descriptions, as the combination for language learning to come over that disadvantage.

In computer science, learning and representing semantics is one of the most important tasks that significantly contribute to some growing areas, as successful stories in the recent survey of Turney and Pantel (2010)[72]. A good semantic representation can help us tackle various problems from attest words/concept similarity measurement to information retrieval or e-learning. Some evidences from recent works show that it is very possible to combine natural language information and other perceptual information to get learning systems that are much better than systems coming from individual information types. These systems can learn language from ambiguous knowledge (Mooney et al., 2008 & 2010) [14, 15], generate language as a sport caster, label images as an image annotator (Mooney et al.,2008 & 2010, Lapata et al. 2008 & 2010) [14, 15, 22, 23, 24] and detect vision activities (Gupta and Mooney., 2010, Roy et al., 2007) [29, 52], etc. The achievements from those systems are applied in various types of applications

to bridge the linguistic interaction between human and machine.

1.1 Aim and Objective

Practically, the overall aim of this thesis is to explore specialized topics in multimodal learning. Our goal is to make use of visual information and textual information and create a multimodal system that can better represent semantics than individual systems. We put an emphasis on combining state-of-the-art text-based semantic and vision-based (or image-based, since we employ information from a large image data set) semantics as well as assessing the role/quality of vision-based information in semantic learning/representation.

Theoretically, we try to make perceptual information from different source communicate, to construct a richer and more human-like notion of meaning. In particular, we concentrate on perceptual information coming from images, and we create a multimodal distributional semantic model extracted from texts and images, putting side by side techniques from NLP and computer vision.

Indeed, there has not been previous work that combined (physically) disconnected visual features to the state-of-the-art text-based semantic model in the literature (-perhaps works of Lapata and Feng (2010) is the first but with the possible exception of their technique that is not applied for combining separated state of the art models together), so in this thesis, we set a target in providing a first (and innovative) framework, as a prototype, to create a multimodal distributional semantic model from independent sources, namely vision-based model and text-based model.

Additionally, in this study we examine different approaches to the combination of semantic models and propose an effective algorithm to accomplish this task. Our algorithm has proven successful by a variety of our experiments.

Our objectives are:

- Extract and model concepts by visual features from (large) image data to create vision-based semantic space; explore which type of feature descriptions are suitable for the semantic representing task.
- Combine vision-based semantic space with state-of-the-art text-based semantic space to create a multimodal semantic model; design an effective algorithm to handle this task.
- Evaluate the quality of the model in well-known existing tasks including: se-

mantic similarity measurement, concepts categorization and the newly introduced BLESS data[1], by EMNLP conference organizers, which is a good benchmark for estimating the semantic representing capacity of models.

1.2 Contributions

The contributions of this thesis are pertain to the following points:
- Propose the framework to create a multimodal learning system (in the form of distributional semantic model) from two different sources: vision and text.
- Propose a good strategy to represent concepts meaning from images including the pipeline of extracting visual features as well as the most suitable type of visual features
- Propose good algorithms to combine vision-based semantic model and text-based semantic model to generate a united one that outperforms tasks.
- Prove that perceptual information from images can capture semantic relation among words/concepts and adding them to the state of the art text-based model is better than adding further text features. We also provide preliminary evidence for an integrated view of semantics where the more concrete aspects of meaning derive from perceptual experience, whereas verbal associations mostly account for abstraction.

1.3 Thesis Structure

This thesis contains 6 chapters. We will review the related works in the second chapter. Since our work is not only involved in Natural language processing techniques but also computer vision techniques, we will provide the background of image processing and the advanced approach in describing visual features from images in the chapter 3. The next chapter will describe the high-level architecture of our proposed framework. In the chapter 5, we go into the evaluation tasks. The chapter 6 concludes our results and analysis.

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

Chapter 2

Literature Review

This chapter is to overview the background of the thesis. It is devoted to the semantics representation and its applications. The basic idea is to describe semantics of words or concepts by their related distributional information and then represent them in the vector format. This technique is referred as vector space models of semantics. However, up to our best knowledge, current works in Computational linguistics (CL) fields are mainly based on text distributional information.

In fact, there are not many works on using vision-based information to improve CL systems but there are more studies in exploiting text-based information to improve vision-based systems such as multimedia retrievals or robotics. In this chapter, we will review works in semantic representation and other studies that explore combination of text-based information and vision-based information, although not all of them focus on semantics representation.

2.1 Introduction

Computers understand very little human language. That leads to hardness in giving friendly instructions to computers. Aiming to break this limitation, computer scientists try to construct modelling systems that can capture meanings of human language, or semantics. Their systems concentrate on representing semantics to enhance computer's learning ability. Traditional approach is to use text information since it is considered as the easiest media to communicate. Another branch of this approach is to use human voice, or speech information. Last decade has

seen a rising approach in using further information rather than text of computer scientists and cognitive scientists. Most of these works, we call situated systems or grounded systems, focus on combining textual information with perceptual information such as visual and emotional information.

2.2 Text-based Semantic Representation

Text-based semantics model is considered as the most effective approach for semantic representing. Some studies prove that using combination of several semantic types can help improve their systems. However, most of those works are strongly based on text-based semantic models, and other semantic information is added to help these systems more robust. Indeed, the text-based semantic model still wins admirations because of its convenience in building and its potency in problem solving. In this part, we will present the traditional approaches in constructing and applying text-based models: semantic network and distributional semantic model (DSM). The later part of this section will focus more on DSM since it is one of the most productive representation up to our best of knowledge.

2.2.1 Semantic Network

The principle of semantic network was born in pretty long time ago as a network of associatively linked concepts. It is conceived as a "presentational format that would permit the meanings of words to be stored, so that humanlike use of these meanings is possible" [18]. Originally, it is mainly designed to presents properties of thing rather than emotional meanings. Collin & Quillian (1969) proposed the earliest work in semantic network in which concepts are stored within a hierarchical structure. In their work, the network of concepts is displayed as a taxonomic tree where the levels for representing concepts ranging from the most abstract down to the most concrete. The concept will be linked to a list of its properties and then the meaning of a concept can be guessed from the concept it is linked to [18]. Inheriting from these directions, more semantic networks are constructed. Their primary characteristics are: each concept is a node in the semantic networks (graph); edge between 2 nodes is labelled by the relation between them, for example, "is" will be labelled on the edge linking "Dog" and "mammal" to represent *"Dog-is-mammal"*; and relatedness between 2 nodes is expressed by the distance between them. However, this model has severe limitations as a gen-

eral model of semantic structure. Its hierarchical structure clearly accepts only some certain taxonomically organized concepts, such as classes of animals, trees or humankind. Thus, semantic networks are mostly appropriate for small scale human-coded collections of concepts [67]. Starting from this limitation, Steyvers and Tenenbaum (2005) made an attempt to create a large-structure of semantic network from word association norms. However, comparing to the vocabulary of an adult speaker, the number of norms their network can represent is still much smaller [48]. Breaking this disadvantage, Harrington and Clark recently built the ASKNet semantic network by using output of a natural language parser. The ASKNet presents 1.5 million nodes with 3.5 million links extracted from 2 million sentences [32]. They claim that ASKNet has a strong ability to tackle the task of semantic measurement by testing on some evaluation test sets and showing good performance [31]. Another semantic network delivered from same approach as the ASKNet was developed by Wojtinnek and her colleagues (2010) [79]. They tackled the task of similarity measurement by exploring the relatedness of surrounding local networks and demonstrated an ability to overcome some problems of semantic analysis and representation. Even their reported results on WordSim353 are lower then most state-of-the-art models, but they showed a promising power of the automated semantic networks in resolving semantics problems. Figure 2.1 illustrates a semantic network [1]

2.2.2 Distributional Semantic Models

The most famous and effective approach for large-scale semantic representation perhaps is using distributional semantics to describe concept's meanings, called Distributional semantic models. In general, the Distributional semantic models (DSM) are the models relying on some version of distributional hypothesis proposed by Harris (1954) and Millar & Charles (1991). They claimed that the degree of semantic similarity among words can be modelled as a function capturing overlap among their linguistic contexts [8]. Its inference is the meaning of a word can be identified by its usage.

DSM is often in the form of high dimensional vector space so it is known as "word space", "vector space" or "semantic space". It proves a promise to solve a lot of problems related to semantic knowledge as well as lexical acquisition,

[1]Figure illustrates a subgraph of the semantic network constructured by Wojtinnek and Pulman (2010)

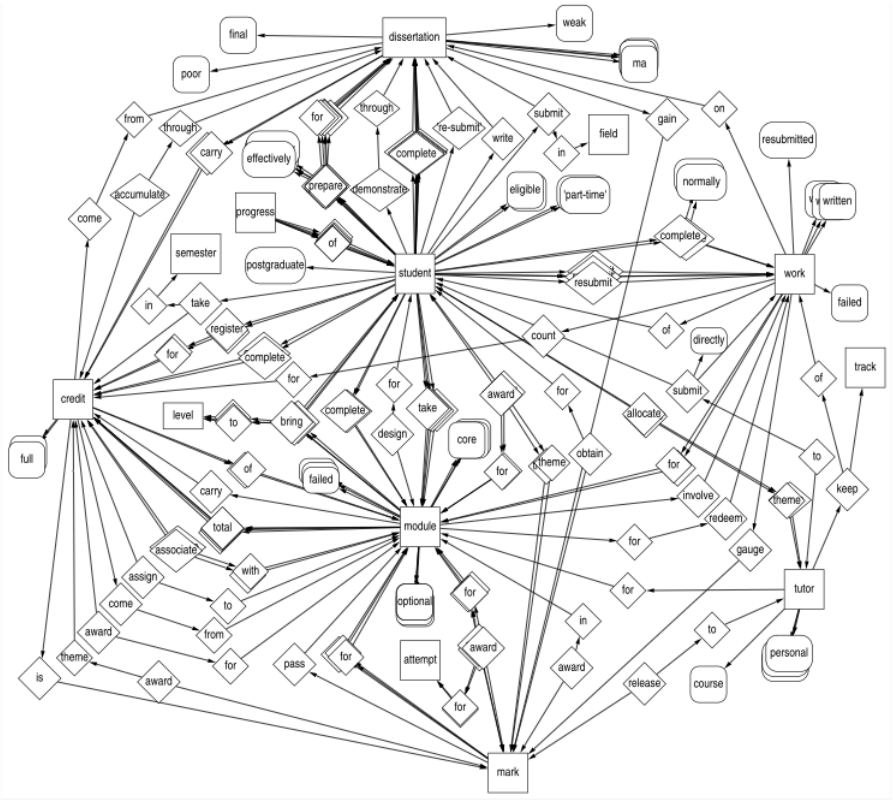

Figure 2.1: Illustration of *Semantic Network*

for instance, semantic relatedness measurement, word sense disambiguation, etc. Furthermore, DSM contributes a lot to the success of information retrieval where similar documents can be detected and queries can be reformed efficiently. One of the earliest IR system exploiting DSM is SMART of Salton (1971). His concept of indexing and representing documents as elements in vector space is still warm and is applied in the most successful search engines such as Google, Yahoo! or Bing. That idea works well until some recent decades when the modern search engines more and more capture DSM's advantages to attain better answers [72].

The success of IR brings about a trend in computational linguists to improve the quality of DSM model. In this line, there are quite many interesting works on enhancing DSM and examining its quality on some related linguistic tests. For example, Rapp used the DSM approach in form of vector-based representation to tackle the TOEFL solving problem. His system attains 92.5% accuracy on multiple-choice question which is much higher than average human-rating score (64.5%) [61]. Similarly, Turney constructed a DSM and tested it on the SAT

college entrance test. His model scores 56% on the multiple choice questions and thus can be comparable to the average results of human (57%) [73]. While many of works concentrating on constructing a particular DSM that can be applied to a specific task at hand, Baroni and Lenci provided a DSM namely Distributional Memory **(DM)** to overcome many tasks at once [8]. They support the concept "one distributional model, multiple tasks" and argue against "one semantic task, one distributional model" which owns a great limit of the current state of the art. In their work, various linguistic relations of verbs, objects, subjects and others are employed to form a highly dimensional vector model which later scores highly in similarity judgements, noun and verb categorization, and others. All in all, those successful stories just to emphasize that DSM is earning highly impacted influence in the currently active fields.

2.3 Distributional Semantic Modelling

There is often understood that modelling distributional semantics is to model co-occurrence information between words (or concepts) in large corpora. DSMs are often classified into 2 categories: *unstructured DSM* and *structured DSM*. We will go into details of the modelling approach for those types of DSMs.

2.3.1 Unstructured Models

Unstructured DSMs are models that don't use the linguistic structures to compute co-occurrences. In contrast, they rely on some degrees of the lexical distance between target elements and context elements to identify co-occurrences. In other words, a co-occurrence happens when the target elements appear "close enough" to the context element. People use set a window size for co-occurrence recognizing. Formally, this type of DSMs is referred as the 2-way structured matrix $M_{|B|x|T|}$ (i.e, formal definition of semantic space of Pado & Lapata, 2007) where B is the set of basis elements of context and T is target elements that can be compared to each others by information captured in B [8]. This approach is simple but capable of simulating the semantic representation.

To make it clear, we may look at the text *"the boy plays football"* and *"this book belongs to Elia"*: "boy" and "the" are shared features of "play" and "football"; "Elia" and "book" also share the feature "this", "to". We don't take any linguistic relation here.

One of the earliest unstructured DSMs was provided by Schutze (1992,1993,1997) [36, 37, 38] in which they showed simply co-occurrence statistics computed from text resource can demonstrate a substantial amount of semantic information. In his works, he computed context vectors from a fixed window size (i.e, 1000 characters (1992), 1001 4-grams (1993)). Next, the comparison in meaning of words can be analysed from their corresponding vectors. Interestingly, he showed that using a method namely Singular Value Decomposition (SVD) to compress/select the best subset dimensions from the statistics will improve a lot the quality of this model. In generality, SVD is a method related to the standard Principle Component Analysis (PCA) to reduce the dimensional size of non-square matrix in a least squares sense in order to select important columns which are optimal for a target function. It is rather sophisticated so we can not go into the details. Its descriptions can be found in Landauer and Dumais (1997), Manning and Schutze (1999) or other studies [61, 12].

Similarly to this line, Lund and Burgess (1996) constructed a framework namely Hyperspace Analogue to Language model (HAL) and showed how simple patterns from available corpus can effectively handle related-semantics problems. Instead of using n-grams in character level, they take windows of 10 words in consideration to identify the lexical co-occurrences in a large corpus of 160M words coming from the USENET newsgroup discussions. HAL is based on a corpus where the lexical co-occurrence is used for producing a high-dimensional semantic space. HAL weights lexical co-occurrences and then creates a **nxn** co-occurrence matrix. Each word will be represented as a *2n* dimensional space from its corresponding row and column. Lund and Burgess used the Euclidean Distance computation to indicate how the words are similar to each others [43]. Additionally, the semantic similarity measurement can be applied by other methods such as Cosine Similarity or Manhattan Distance.

The size of window in computing co-occurrence can be changed variously depending on tasks. For example, Rapp (2003) used a window size of 2 on a sparse corpus to count word co-occurrences. His framework proves that word sense can be induced well from that simple statistics. However, he argued that SVD is not so convincing in preserving the information for sense induction and emphasized that the larger corpora the better quality in representing the word sense.

Presenting one of the most noticable works of unstructured DSM, Landauer and Dumais adopted a slightly different approach of information retrieval namely

Latent Semantic Analysis (LSA, it is similar to Latent Semantic Index in Information Retrieval), placing emphasis on reducing dimensionality of word vectors [44]. It is often referred as the term-document approach where context elements co-occur with target elements if they are in a same documents. Like most of the work in unstructured DSMs, they also used a fixed window size to construct matrix of co-occurrence, nevertheless, they was able to demonstrate how word co-occurrence data was adequate for children's learning vocabulary simulation. That makes clear the significant role of co-occurrences in meaning representation from the psychological point of view.

The most recent model of unstructured DSMs, Bulliaria and Levy extracted semantic representation from the statistics of word co-occurrence. They argued that some aspects for word meaning can be emitted from patterns of the word co-occurrence and validated their idea on some testing sets of TOEFL, semantic categorization and syntactic categorization. In experiment, they counted the number of times *n(c,t)* each context element c occurs with target element t in a certain window size, transformed to the probabilistic score and then formed a corresponding vector of t that is used for later semantic similarity computation. They also recommended that using *Positive PMI* to represent probabilistic scores (i.e. components of target's vector) and *Cosine Similarity* in measurement semantic similarity should bring about better results than other methods [12]

2.3.2 Structured Models

Similar to unstructured DSMs, the structured DSMs can capture the information of co-occurrences from text resource. Nevertheless, it distinguishes itself at extracting linguistic relations between words. Co-occurrence statistics is computed between a pair of words by pattern that links them. Generally, a linguistic relation is represented by 3 elements (triple): 2 words and a syntactic or lexical-syntactic link. In almost cases, the link reflects the lexical-semantic properties related to those words and can be determined from sentence structures. It bases on the hypothesis that surface structured relations cue their semantic relation [5, 16, 19, 74, 60, 64]. Let's look back at the previous examples *"the boy plays football"*: "play" is considered as a relevant context connecting "boy" and "football" but it should not be the context property of "the". Take more illustrations, structured DSMs would not regard "very well" as a linguistic context for "football" but would rather to regard it as a context for "play" in the sentence *"the*

boy plays football very well"; and it also would not consider "eat" as a legitimate property for "red" in the sentence "the teacher eats a red apple" [8].

Regarding comparison of structured models and unstructured models, computational linguists claim the structured ones are at least not worse than the unstructured ones [60, 64, 8]. It is because structured models shapes its semantic representation by taking syntactic structures into account. Furthermore, it requires some preliminary text processing procedures such as parsing, pattern extractions which take pretty much information about the unstructured statistical information in the corpus. Therefore, the structured DSMs may capture some information of unstructured co-occurrences as well as deeper information for that unstructured models may not reach. In addition, the preprocessing procedures practically work like filterers and only interest in linguistics-reflected patterns of sentences, so that the structured DSMs tend to more sparse and more selective. Experiments also suggest that structured models perform slightly better than their unstructured sisters [60, 64, 8]

In presentation, structured DSMs seems to be more flexible than unstructured DSMs. They can be represented in 2-way matrix forms of ether by dropping one elements or joining 2 elements from each triple (relation) [74, 60, 8]. They also can represented as a list of 3-ways tensors that can support many tasks at once as what Baroni and Lenci (2010) constructed or can be converted into large semantic networks, or graphs, which can simulate the ASKNet [32]. Nonetheless, there is a two-edged blade in their representation. On one hand, the different combinations of 3 elements provide different insights into corpus and then can perform well in different tasks. On another hand, in 2-way matrix form, they lose high demonstrative competence of the large set of triples (relations) acquired from large corpus [8].

2.4 Semantics in Multimodal Systems

There are some certain works in improving learning ability of computers that already succeeded in some fields such as robotics, computer science and cognitive science. These works are often referred to "grounded language learning" , "situated learning system" (in this thesis, we adapt a term "multimodal system" as the general description about that approach). Most of these works study the combination of linguistic information and one more type of perceptual information.

Some works are designed for general purpose of meaning representation while some others pay attention to language learning, generating and understanding.

Proposing one of the first studies in this direction, Siskind demonstrated a translation of video input into structured representation and developed a temporal representation that captures relationship between objects inferred from visual observations. The system can recognize the occurrence of events, which are described by simple spatial-motion verbs (Siskind and Morris, 1996).

Taking advantages of the integration of language descriptions and visual input, some systems are able to generate natural language to describe sport events. For example, the VITRA system can handle language generation task for soccer matches and traffice scenes (Herzog and Wazinski, 1994), and more recently, Sportcasters can procedure various descriptions for soccer games simulated by the Robocup simulator (Mooney et al., 2008). VITRA bridges the gap between vision system and natural language (NL) by encoding spatial relations and interesting motion events into verbal descriptions. In the beginning, visual data is transformed into geometrical representation, and later, explicit links between sensory data and NL expression are formed for NL generating purpose. Mooney and his colleagues developed a series of the Sportcasters systems in which they use semantic parsers to transform the description of visual information into the logical semantic representation and then generate NL basing on that (Mooney et al, 2008, Chen and Mooney, 2010). Additionally, they address the problem of learning language from ambiguous data into their systems, nevertheless, they don't process visual information, instead they use prefab of abstract symbolic activity descriptions, and their systems are designed for working on video annotation or vision recognition.

Roy and his colleagues present a number of systems that connect natural language to perceptual environment (Roy and colleagues., 2002& 2004& 2005). Some of them learn word's meanings by analyzing speech in related object's vision while some others model and acquire language's grammar from the scene descriptions. They also developed a framework for grounding the meanings of words by connecting them to a network of sensory motors (Reiter and Roy, 2005). More presently, Fleischman and Roy (2007) used both captions and motion descriptions of baseball video to retrieve relevant clips given a textual query. Their direction is quite close to the work of Gupta and Mooney (2010) where video, as modelled as "bag of visual words", and closed captions are used for motion detection. Their work showed that language can improve the quality of vision scene identification.

More into using visual information to enhance solving techniques in basic NLP, Barnard and Johnson exploited vision-based information to tackle word sense disambiguation (Barnard and Johnson, 2005), Kelleher and his colleagues built a system that is able to generate and interpret referring object expression (Kelleher et al., 2005). Most recently, Lapata and her colleagues have constructed a computational model based on texts and images. In their work, images are translated to visual features and used with co-occurred words for topic inferring. Their model, taking the advantages of Latent Dirichlet Allocation (LDA) and probabilistic estimation, can manage the task of image annotation and text illustration well even it could be better if it uses more information of available image captions for generating model (Lapata and Feng, 2010). Like this vein, Feng and Lapata (2010) also presented a multimodal semantic space that is earned directly from documents and their associated images. This semantic space is extracted from a large corpus, employs LDA analysis, and infers the meanings of words. The model gets a significant improvement in handling the task of semantic similarity measurement. Although Feng and Lapata's work is very heartening, it has several drawbacks [13]. Firstly, the model requires the extraction of information from mixed-media data so that it constricts its methods of textual and visual features extraction. That means it is not easy to add visual features to the state-of-the-art text-based models like the model extracted from Wikipedia or other larger corpus. Virtually, the quality of the model is limited, and it is hard to judge how effective the visual feature contribute to the already good text-based model. Secondly, their joint model is trained in a such way of a mix between textual features and visual features, thus, it is very hard to separately evaluate effects of vision and text on the overall model.

Regarding how the combinational semantic model of perceptual information is represented, there are some insightful methods that overcome challenges of mixed semantics learning. For instance, Andrew and his colleagues proposed the HMM model (2009) and LDA model (2010), and Piantadosi presented the Bayesian model [57, 3, 2]. These models can deal with two well-known substantial challenges to learners: referential uncertainty and subset problem. Although they use the semantic representation for different purposes, they clearly showed the success and importance of multimodal semantics model. Nonetheless, their systems, such as Andrew's, rely on speaker-generated features so they require a lot of manually encoded. That explains why the main disadvantage of those models is lying on the scale of the models. They only handle a limited number

of concepts that is not comparable to what human use daily.

Chapter 3

Image Processing: Foundation and Approach

The aim of this thesis is to discovery the convergence of vision-based semantics and text-based semantics. Normally, vision-based semantics is often extracted from labelled images or video and deals with image processing and video processing. Between them, processing video is more complicated than processing images since it requires split video scene into series of images or frames and then applies image processing techniques to each of them. In this thesis, we don't concentrate on how to split video scene to images sequence and analyse them so we are not going to propose new methods to process images or video scene. On the contrast, we are more interested in exploring how vision-based semantics can contribute to the text-based semantics models. Therefore, employing state of the art techniques in image processing makes more sense for us in this topic.

In this chapter, we will present the current cutting-edge methods in Image processing, or Computer Vision (CV) in more generally speaking, as the CV background of the thesis.

3.1 Image Formation

Image in computer is a 2-dimensional matrix of pixels whose values are proportional to the brightness of the relative point in the scene [56]. Normally, the image is represented as NxN matrix pixels of m-bit values, that means it has the point size of N and its brightness is in a range of m-bit values (2^m values). Thus, the

larger m is, the more colors and higher contrast that image can expresses. For example, with m equals to 8, the image is often in black and white (gray) color with brightness ranges between 0 and 255.

Color image is stored a bit sophisticatedly than gray one but it still follows that strategy (i.e, matrix of pixels of m-values). However, instead of using one image plane of pixels, it uses several pixel components. For example, the common RGB image has three pixel components corresponding to red, green and blue; CMYK image has four pixel components corresponding to cyan, magenta, yellow and black, etc. Each component contains information specifying pixels' intensities. The combination of those component therefore provides a lot of colors [56]

3.2 Feature Extraction

3.2.1 Overview

Feature extraction and matching are important for many computer vision applications and vision-related fields. Currently active features are felt into 2 main categories: keypoin features (aka. interest points), and edge features. The keypoint features focus on salient and specific location of images such as mountain peaks, shape of eyes, leaves' corners while edge feature focus on group of curves, local appearance or boundaries of some objects [70]. However, keypoint features are often considered as stronger features and outperforms the edge features in image matching tasks, ie. finding image locations or a sparse set of corresponding locations in different images. The reason is they are very distinctively and selectively extracted from the large database of features with high probability, which is useful for accurate location matching in 2D images [46, 70]. An practical example can be found in [76, 45]. On the contrast, edge features, with its object fenceline, are robust in 2D object or 3D occlusion events recognition (both are delineated by visible contours) [70]. An alternative version of edge features are line features that are applied for rectangle detection.

To detect keypoint features, the key idea is to determine image locations which can correspondingly appear on other images[70]. Perhaps the most famous approach of feature extraction in computer vision is Scale Invariant Feature Transform (SIFT) which transforms features of the images into scale-invariant coordinates. It main advantage is that it procedures a large numbers of features densely covering the images. Therefore, it seriously takes the quality of the image

in to account to decide the quality of features [46]. SIFT approach is described by a sequence of following procedures:

(i) Scale-space extrema detection: it searches over all scales and image locations. It is often implemented effectively with the Difference-of-Gaussian (DoG) function do determine keypoints that are invariant to scale and orientation.

(ii) Keypoint localization: when a candidate is determined, it is passed to the model that measure their stability to be considered as the keypoint.

(iii) Orientation assignment: Each keypoint are assigned with one or more orientations by exploring local image gradient directions. Further operations on images will be performed relatively to the assigned orientation, scale and location of each feature.

(iv) Keypoint descriptor: construct representation of local image gradients that can reflect the local shape distortion and change in illumination.

The more details about SIFT approach could be found in [46, 70]

3.2.2 Pyramid Matching

In image processing, we may need to change the resolution of an image as a preliminary step before making further analysis. Examples include that it may need some changes in the resolution to match the output setting of some devices such as printer or computer screen; or to match the objects when we don't know exactly size and scale. Image pyramid is a technique to fulfil these requirements. Sometime it is more known for multi-resolution image representation. The main direction is to look for objects or patterns at different scales and perform multi-resolution changes and to speed up the coarse-to-fine search algorithm [70]. Traditional pyramid techniques like Laplacian pyramid, Gaussian pyramid work in the same idea: forming different levels of scale; at each level, a daughter of its parent-level image is generated by sampling by a factor of 2. Figure 3.1 illustrates the traditional image pyramid multi-resolution changes through levels [70].

Starting from Image pyramid idea, Grauman and Darrell (2005) proposed *pyramid matching* kernel to locate an approximate correspondence between two sets of *highly dimensional* vectors of features. It places a sequence of coarser grids over the features space and considers 2 points are matched with each other if they are in the same cell of any grid of resolutions. Its weights of the matches found vary differently by levels: weights are higher in finer resolutions and vice versa) [27, 45] Thus, it can perform accurate matching in high dimensional space.

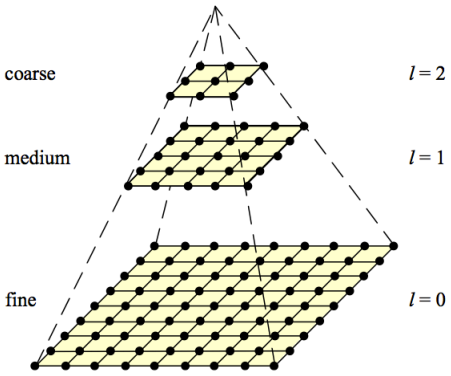

Figure 3.1: Illustration of *traditional image pyramid* multi-resolution changes through levels (Szeliski, 2010)

Specifically speaking, assuming that we have L levels of grids. The grid at resolution level l has sizes of 2^l cells for each of its dimensions, for a total of $D = 2^{dl}$ cells. The number of matches at level l is determined as the histogram intersection of two corresponding histograms of sets

$$I^l = \sum_{i=1}^{D} min(H_X^l(i), H_Y^l(i))$$

where $H_X^l(i)$ and $H_Y^l(i)$ are the number of points from 2 sets of vectors that fall into the i^th cell of the grid. Since it contains the number of matches found in the finer level *l+1*, so the actual matches found in the level l is:

$$I^l - I^{l+1}$$

Grauman and Darrell (2005) used the weights for each level l with:

$$\frac{1}{2^{L-l}}$$

The Graumann and Darrell's *pyramid matching kernel* is defined:

$$Kp^L(X,Y) = I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(I^l - I^{l+1})$$

3.2.3 Spatial Matching

Pyramid matching kernel has a withdraw: it works with an orderless image representation and discards all spatial information, therefore, it is not able to capture shapes or segment objects from image's background. Overcoming this, Lazebnik & Schmid & Ponce (2006) proposed *Spatial Matching Scheme* which performs the pyramid matching in the 2D image space and uses traditional clustering algorithms in feature space. They demonstrated that their approach significantly enhances the bag of visual words (boVW) model which already showed impressive levels of performance [45] (we will discuss BoVW in the following section). The general concept of Spatial matching is to quantize/group all feature vectors into distinguished types and only consider 2 feature vectors matched to each other if they are same type. After that, Pyramid matching kernel is used to find matches in each type and summed up to form the *Spatial matching kernel* finally. The formal formula is:

$$K^L(X, Y) = \sum_{m=1}^{M} Kp^L(X_m, Y_m)$$

3.3 Bag of Visual Words

Bag of Visual Words (BoVW) is an powerful approach of computer vision that is employed from NLP. In NLP it is "bag of words" but we adopt BoVW for its analogous model in computer vision. "bag of words" (BoW) is a dictionary-based method used in NLP and Information Retrieval. Its main idea is to represent a text, or document, as an unordered collection of words without grammar consideration. Similarly, in BoVW, each image are treated as a document and then be described as a bag of visual words which are formed from local interest points (keypoints - defined as salient image patches). This method is widely applied in computer vision to tackle various type of tasks, especially image categorization [66, 17, 55, 11, 59]. The following will describe BoVW procedures:

Keypoint extraction: keypoints are extracted from each image over image data set. They are in the form of large dimensional vectors. For example, SIFT keypoint are often in a vector of 128 dimensions.

Vector quantization: The keypoint vectors are projected in the same space and clustered into groups. Each group is considered as a visual word. One of the most common vector quantization techniques is K-means.

Image descriptor: vector quantization allows each original keypoint mapped into one visual word. Let number of visual words is n, then each image will be represented as a *n-dimensional* vector where each dimension' value reflects the occurrence number of corresponding visual word. By that way, the image is in a form of fixed dimensionality spare vector instead of a very large and variant set of keypoint vectors [13].

Figure 3.2 illustrates above strategy.

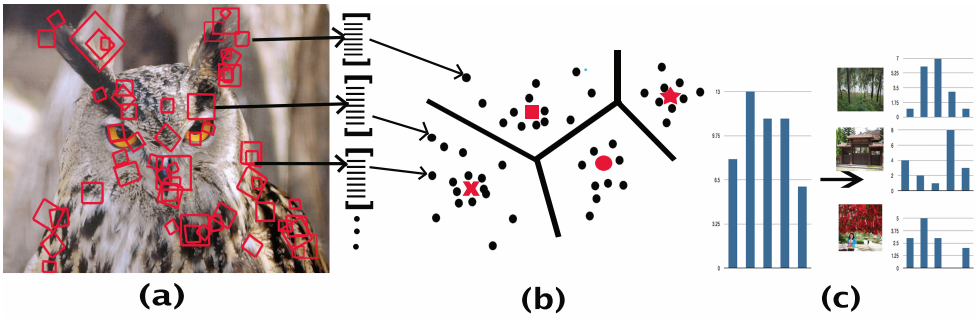

Figure 3.2: Illustration of *bag of visual words* procedure: (a) detect and represent local interest points as descriptor vectors (b) quantize vectors (c) histogram computation to form BoVW vector for the image

In fact, the information that image BoVW vector contains is often referred as the low-level appearance of patches from 2D images. Those patches bring about rich local information of the images and often point out the area around corners and edges in the image objects [59]. However, the content that BoVW captures varies a lot depending on the how we extract keypoints, regarding types of keypoint, quantization algorithm as well as number of visual words we selected.

3.4 Color Descriptions in Image

This section is dedicated to various types color features in image processing which achieve a good performance in computer vision. Actually, the intensity-based features have been widely used but the color features just have been proposed recently to improve distinguishable ability of the system.

3.4.1 Histograms

RGB Histogram: The RGB histogram is a combination of three 1-D based histograms (R(ed), G(reen) and B(lue)). This histogram doesn't reflect invariant characteristics [77, 76].

Opponent Histogram: Similar to the RGB histogram, Opponent histogram is also a combination of three 1-D histograms that are based on different channels o the opponent color space:

$$(O1, O2, O3) = \left(\frac{R-G}{\sqrt{2}}, \frac{R+G-2B}{\sqrt{6}}, \frac{R+G+B}{\sqrt{3}}\right)$$

the channels O1 and O2 reflect the color information of the image. Channel O3 reflects intensity information [77, 76]

Hue Histogram: In the HSV color space, hue is often considered being unstable near the grey axis. The certainty of the Hue is inverely proportional to the saturation [58]. That is the reason Hue histogram is often more vigorous if each sample of the Hue is weighted by its respective saturation. Concerning light intensity, it reflects scale-invariant and shift-invariant [77, 76]

rghistogram: *rg*histogram is histogram representing color information of images. *rg*histogram deliveries from RGB color model when it is normalized. In that case, $r + g + b = 1$ so b is redundant if we only take color information (r and g) into account.

$$(r, g, b) = \left(\frac{R}{R+G+B}, \frac{G}{R+G+B}, \frac{B}{R+G+B}\right)$$

It is scale-invariant and invariant to light intensity changes, shadows and shading [76]

3.4.2 SIFT Descriptors

SIFT the SIFT descriptor is the type of keypoint descriptor that is extracted by SIFT approach (see section 3.2 above for more details, again). It is proposed by Lowe (2004) to describe the local shape of a region using edge orientation

histograms. Its properties can be summed up in 2 points: (i) Gradient of an image is sift-invariant

(ii) Not invariant to light color changes

it is often in the form of 128-dimensional vector.

HSV-SIFT Bosch and his colleagues (2007) computed SIFT descriptors over all 3 channels of the HSV color model. Each descriptor consists of 3x128 dimensions where each channel occupies 128 dimensions. It does not posses any invariance properties although the H color model is scale-invariant and shift-invariant. The reason is it is computed as the combination of the 3 channels H,S, and V [77]. The drawback of this descriptor is that the periodicity of the hue channel and the instability of the hue for low staturation is not addressed.

HueSIFT Weijer and his colleagues (2006) concatenated hue histogram with SIFT descriptor to form HueSIFT. It address the instability of the hue near the grey axis as well as the periodicity of the hue channel. In addition, it reflects scale-invariant and shift-invariant properties as the hue histogram [76].

OpponentSIFT OpponentSIFT uses SIFT features to describe all channels of the opponent color space. It remains the properties of the opponent color space: O3 indicates the intensity information and O1 and O2 indicate color information which is not invariant to changes of the light intensity [77].

rgSIFT As synthesised description of van de Sande and his colleagues (2008), rgSIFT are added for the r and g chromaticity components of the normalized RGB color model which is scale-invariant. It also owns shift-invariant properties, nevertheless, the color part of the feature is not invariant to the changes of the illumination colors [76].

RGB-SIFT RGB-SIFT are SIFT descriptors delivered from every RGB channel independently. It is equal to the transformed color SIFT features that are computed by normalizing every channel of RGB and computing SIFT descriptor for each. It is scale-invariant and shift-invariant as well as invariant to light color changes [77].

Chapter 4
Semantic Model in Multimodal Space

This chapter will discuss our approach in combining image-based semantics and text-based semantics. We propose a framework that employs state-of-the-art techniques in both computer vision and natural language processing to create a multimodal semantic space. The key idea is to construct text-based and image-based co-occurrence models separately and then combine them together. However, we are not going to propose a new model for the text-based semantics for our purpose. The reason is that there exists a numbers of promising text-based semantic models that are staying on the state-of-the-art lines. Instead, we keep the framework open to all of DSMs that are publicly available off-the-shelf and has been shown successful. We also provide a method to boost up the model in order to have deep insights into how effectively our multimodal semantic system can do. Beside all of that, we pay our attention on forming a good vision-based model and various ways to concatenate it with the text-based model.

4.1 System Architecture

The Figure 4.1 presents a diagram with the architecture of our framework[1]. Its working flow is involved in text-based and vision-based DSM modelling in form of high dimensional vectors, visual feature extraction based on top-tier techniques (keypoints extraction), bag of visual words modelling, tag modelling and vector concatenation. To make it easy to follow, we first describe our procedure to build

[1]The source code of the system is published at: https://github.com/s2m

both text-based and vision-based DSMs. However, we stress the latter since it is the more novel part of the procedure. Then, we describe our combination techniques to integrate both models. We underline keeping our system architecture at an abstract level to enlarge scopes of applications.

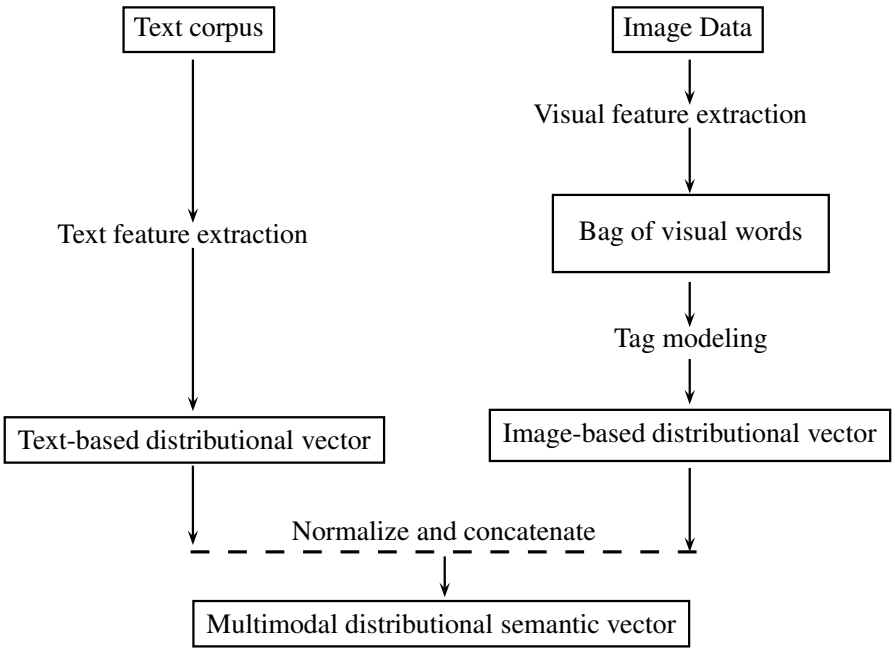

Figure 4.1: Overview of our system architecture

4.2 Text-based semantic model

Looking at the pros and cons of text-based semantic family including semantic networks and DSMs. We decided choosing DSMs for its generality and power in modelling semantics of words/concepts instead of semantic networks (*see 2.2.2*). As mentioned previously, we don't consider proposing a new DSM for our framework and also don't concretely attach our framework to any specific DSMs. On the contrast, we keep the architecture of our multimodal system open to various DSMs. We have a couple of reasons to do that:

Firstly, if choose/propose a specific published model fixed for the framework, we vividly restrict ourself by that model together with its advantages as well as disadvantages. Up to our best knowledge, each DSM has its own strength and weakness. They are often designed to some specific purposes so that some can be

good for some tasks but not good for other tasks (see *ACL wiki of state-of-the-arts* [2] or [8])

Secondly, the open framework makes itself easy to be shared and competed with other types of models. It is also straightforward for computational linguists or computer vision scientists who want to explore the convergence of text and vision by employing their proposed text-based models or vision-based models regarding their new visual features.

The only enquired requirement from the text-based model in our framework is that they need to be transformed in the form of high dimensional vectors (as is the most general form of DSMs). Thus, the whole DSM is encoded in a matrix in which each target word is represented by a row vector or weights representing its association with collocates in a corpus. We will go into details about the text-based model in the section 5.1.2

4.3 Image-based semantic model

4.3.1 Image Data

We use the image data set that contains a large number of images that are labelled (manually rather than automatically to ensure the quality of the Image data, although a fully automated ones would be cool). We assume that the word labels of the images, that we call **tags**, somehow related to the context that the image express. We base on a hypothesis that tags and image should share some salient semantic information. Figure 4.2 illustrates an image and its tags.

4.3.2 BoVW Construction

The key approach to form the image-based vector space is to use the BoVW method (see 3.3 for a more detailed description of the BoVW) which already became a powerful approach in representing image by features. Following what has been increasingly recognized as standard procedures in vision feature extraction, we use the Difference of Gaussian (DoG) detector to automatically detect keypoints from images. We use the Scale-Invariant Feature Transform (SIFT) to represent those keypoints in term of 128-dimensional real-valued descriptor

[2] http://aclweb.org/aclwiki/index.php?title=State_Of_The_Art

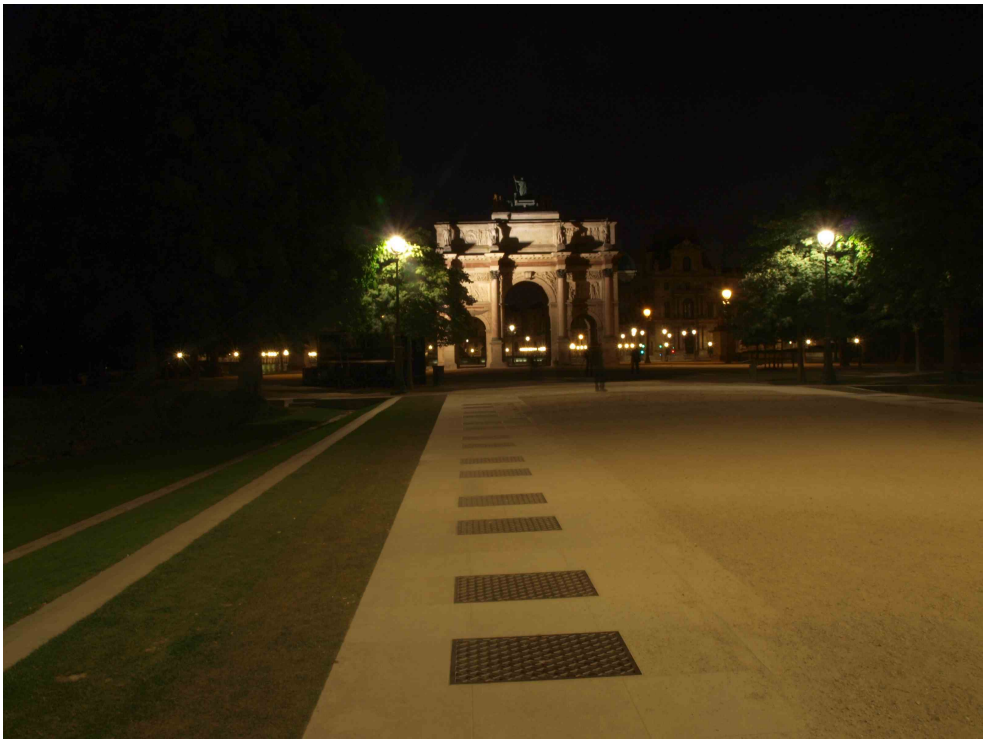

Figure 4.2: Example of image and its tags: *night, tree, gate, street*

vectors. We chose SIFTs for their characters at invariance to image scale, orientation, noise and other properties (see 3.4.2). We apply the pipeline of spatial pyramid matching scheme proposed by Lazebnik and his colleagues (2004) [45] to form spatial histogram of keypoints, called BoVW. For brief reminding, we group descriptors into clusters by K-means algorithm and consider each cluster as a visual word. This algorithm also allows us to map keypoints to a visual word they belong to, so that we can create histograms for each image. Spatial pyramid approach is applied to help us enrich the histogram models by dividing the image into small pieces and integrating histograms of pieces together (see chapter 3.2 for more detailed description about feature extraction techniques). Figure 4.3 demonstrates an illustration of spatial histograms obtain in the end.

4.3.3 Tag Modelling

The previous procedure provides us an useful representation in term of occurrences of visual words of each image in our image data set. However, we just have information of image separated to textual words information. To depict the

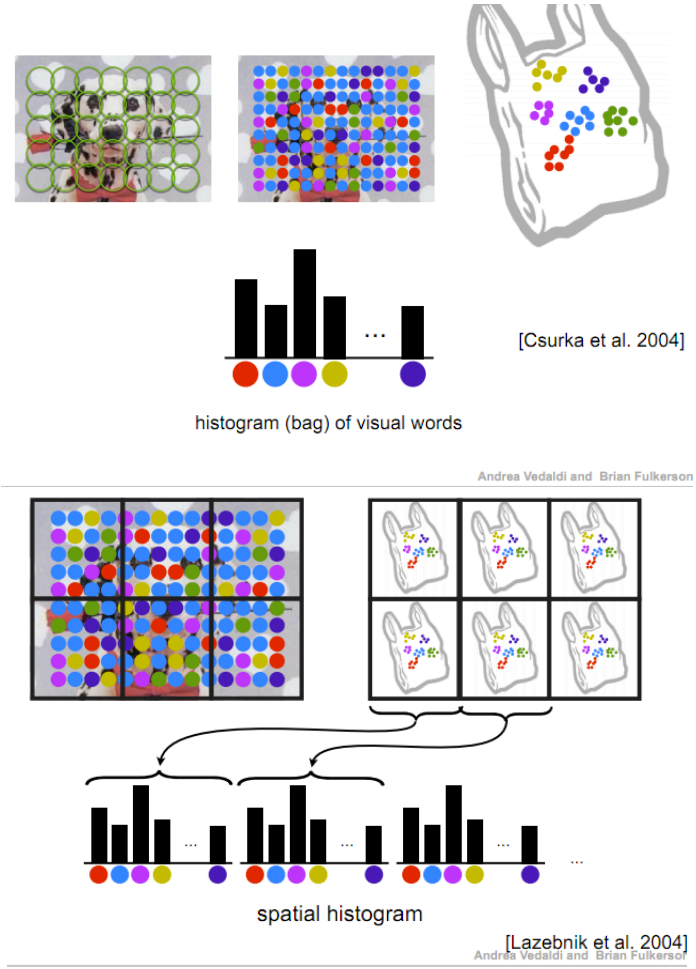

Figure 4.3: illustration of spatial histograms of the image, Vedaldi and Fulkerson (2010)

connection between textual context and visual context of the image, we associate the tag to BoVW of all images it is labelled to and sum visual word occurrences across those list of images. This step is called *tag modelling*. That results a raw frequency histogram of visual words for each tag. We transformed them into Local Mutual Information **(LMI)** scores computed between each tag and visual word. LMI is an association measure that closely approximates the commonly used Log-Likelihood Ratio while being simpler to computer [20]. It is simply computed as following formula:

$$LMI(T, w) = p(T, w) \log(\frac{p(T,w)}{p(T)p(w)})$$

where T and w is correspondingly *tag* and *visual word* we are looking at; *p(x)* is probability of x which here we estimate by maximum likelihood using relative frequency.

Finally, we obtain an image-based distributional semantic model which is a matrix consisting of rows, or tag vector, for summarizing the distributional history of the tag in the image collections.

4.4 Integrating Distributional Models

We integrate the two distributional vectors to construct our multimodal semantic space. Notice that a word is represented both in text-based model where it plays a role of a target word and vision-based model where it plays a role of a tag. They are all in a form of high dimensional vector. Our procedure of integration two distributional models simply combines two corresponding vectors. However, the challenge of this step is lying in the spaces of vectors. In fact, we are staying in two different spaces with different dimensions and we don't see the relation between two space yet. Let's say the text-based vector is:

$$u = (u_1, u_2, u_3, ..., u_n)$$

and the vision-based vector is:

$$v = (v_1, v_2, v_3, ..., v_m)$$

where *m and n* is the dimensional size of 2 spaces.
First of all, we need to bring them to the same scale (aka. vector normalization) because they posses different scale sizes at the moment. It disadvantages us if there exists a situation that one vector dominates the values of the other one. In that such case, the combination work is meaningless. We use the common *L2 normalization* technique to transform those vectors into 1-unit length vectors:

$$X = \frac{X}{||X||}$$

From now, we uniformly refer to the normalized vectors when mentioning U and V. Next, our aim is to figure out the function $f(u, v)$ that results the integrated

representation V_f of multimodal distributional semantic space. We propose following methods:

Linear combination We concatenate the two vectors disregarding any weight scheme between them.

$$V_f = u||v = (u_1, u_2, ..., u_n, v_1, v_2, ..., v_m)$$

Linear weighted combination Linear weighted combination works pretty similar to the pure linear combination method above. The only difference is that each vector is multiple with a score before concatenating step. Sum of all scores should be 1.

$$V_f = \alpha * u||\beta * v = (\alpha u_1, \alpha u_2, ..., \alpha u_n, \beta v_1, \beta v_2, ..., \beta v_m)$$

where $\alpha + \beta = 1$ Actually, the Linear combination can be considered as a special case of Linear weighted combination when $\alpha = \beta = 0.5$.

Parameters Estimation for combination (PE) Most of works in DSMs share a characteristic: they treat all features, or relations among words, equally. However, treating all features equally may bring about flaws because some features are more important than others. In our proposal, we provide an approach to overcome that drawback by estimating parameter for weighting features of 2 vectors. This work can be done by machine learning method. Unfortunately, machine learning approach should be costly in computation since we are working with hundreds of thousands features (as is high dimensional vector space) meanwhile the target function for machine learning algorithms is difficult to set up because we aim to some co-efficient benchmarks for semantic measurement - that means the target function is pretty unclear for now.

Staying above all that problems, we provide the algorithm based on an idea: estimating the parameters for groups of features instead of individual feature. To do so, we estimate the importance of features and sort them following bellow procedures:

(1) Do the *topn* selections for text−based features and the vision−based features;
For example: we can select the top 4K features, group them into the 1st *block*; then select the next top 4K features and group them into the 2nd *block*; so on and so forth. The top dimensions are picked based on their cumulative Local Mutual Information mass.

In case the distributional model is represented in a space of huge dimensions (e.g, 700M dimensions), we make the impact of features by selection only limited top n dimensions. We show in the experiments that trimming the distributional models in this way does not have a negative impact on its performance, so that we are justified in claiming that we are still working with state-of-the-art models. Table 4.1 demonstrates the topn algorithm

```
Begin
    Initialize the Cumulative scores of all features as zeros

    For Word in total words do
    Begin
         Update Cumulative scores of all features of Word
    End For

    Sort the Cumulative scores

End
Output N features having highest cumulative scores
```

Table 4.1: Topn feature extraction algorithm

(2) Weight every group of n features and finally combine them together.

$$V_f = \alpha_1 * block_1 || \alpha_2 * block_2 ... || \alpha_k * block_k$$

where $\alpha_1 + \alpha_2 + ... + \alpha_k = 1$; and k is the total number of block *topn* we extracted.

Nevertheless, there raises two issues: (a) how to evaluate our parameters? and (b) how to estimate parameters (α) to get a good model?

To evaluate our parameters: we propose a simple (and very common) method which is where we pick up the parameters' values corresponding to the best model based on our evaluation benchmarks. We will discuss more about the evaluation benchmarks in the next chapter.

To estimate the parameters: we propose a *divide and conquer* technique. In our case that is: the problem of combining **k blocks** of features can be reduced to the problem of combining **2 blocks** of features. We will do *k-1* times of combination. For instance, firstly we vary α_1 and α_2 to combine the *1st block* and the *2nd block*; then we pick up the values of αs that results the best score in our benchmark. We consider the combination of the 1st block and 2nd block as *a new block* and use that new block to combine with the *3rd block*,... so on and so forth. In each of total of k-1 combinations, the sum of αs parameters related to 2 blocks in our consideration should be 1. Table 4.2 presents our algorithm to combine blocks of features.

```
Begin
    Current_Block = block_1
    For i:= 2 to k do
    Begin
         Best_Block = Current_Block
         For \alpha_1 := 0 to 1.0 in step of 0.1
         Begin
                  \alpha_2 = 1.0 - \alpha_1
                   New_Block =  \alpha_1 * Current_Block || \alpha_2 * block_i

                   Evaluate (New_Block)

                   if New_Block is better than Best_Block then
                   Begin
                          Best_Block = New_Block
                          Store the \alpha values for later reference if needed
                   End If

                   Update: Current_Block = Best_Block
          End
      End For
End
Output Current_Block
```

Table 4.2: Parameters Estimation algorithm to combine k blocks of features

Chapter 5

Evaluation

This chapter is to describe our experiments and results in evaluating the multimodal semantic model.

In the first part of this chapter, we will speak about our settings for experiment in which we specify the corpus/data set we use for our experiment. We understand that suitable data are vital for a fair judgement about the proposed approach. Then, we consider testing our model on famous evaluation sets which can provide us insightful assessment about semantic representation and similarity measurement.

One of the interesting works done is visual feature extraction. We use various parameter settings with advanced techniques from the computer vision community. That is the reason for us to claim that we extracted good features from images. In addition, since our image data set for experiment is pretty huge, we borrow techniques from distributed system. All of these works are hidden behind our framework architecture but it is appropriate to mention them here.

In the later part of this chapter, we will report our results and our analysis. The results are promising enough to show that our model could stay in the top-tier line of state-of-the art level. Perhaps more fascinating than the testing scores we have is a proof of contribution of vision in semantic representation. The vision-based features are at least as good as text-based features. This suggests that visual words, although hard to be seen or revealed, can capture the semantic relation among contextual objects of images.

5.1 Experimental Setup

5.1.1 Evaluation benchmarks

WordSim353 We manage our most extensive evaluation on the *WordSim353* **(WS)** data set[1] proposed by Finkelstein and his colleagues (2002) [25]. It is a widely used benchmark constructed by asking 16 subjects to rate a set of word pairs on a scale of 10-point about similarity between two pair words and averaging those ratings. For instance, *coast/shore, 9.10*, *day/summer, 3.94*, *rooster/voyage, 0.62*, etc. There is a total of 353 pairs divided into 2 groups: similarity (we call WS-Sim) and relatedness (we call WS-Rel). WS-Sim is specified for measuring similarity while WS-Rel is designed for measuring relatedness. WS-Sim *semantically similar* (e.g., synonyms or coordinate terms) and WS-Rel - *semantically related* (e.g., meronyms or topically related concepts).

WS-Rel examples include:

```
computer-n keyboard-n 7.62
Jerusalem-n Israel-n 8.46
planet-n galaxy-n 8.11
canyon-n landscape-n 7.53
OPEC-n country-n 5.63
day-n summer-n 3.94
day-n dawn-n 7.53
country-n citizen-n 7.31
planet-n people-n 5.75
```

WS-Sim examples include:

```
tiger-n cat-n 7.35
tiger-n tiger-n 10.00
plane-n car-n 5.77
train-n car-n 6.31
television-n radio-n 6.77
media-n radio-n 7.42
bread-n butter-n 6.19
cucumber-n potato-n 5.92
```

[1] http://alfonseca.org/eng/research/wordsim353.html

We use the data set in a format that contains POS information for each target word.
Visit *http://alfonseca.org/eng/research/wordsim353.html* for more information about the data set.

We evaluate models in terms of the Spearman correlation of the cosines they produce for the WordSim pairs with the average human ratings for the same pairs. With the models based on Google ESP data set, we cover around 73% of whole set (i.e. 260 pairs), therefore, it does not make sense for us to compare the results to the state of the art in the literature. However, our results are still more competitive than that of the state of the art model running on our version of WS. We will describe the Google ESP image data in the next section.

Additionally, because WS-Sim and WS-Rel can report different views about semantic relation, our target is that we not only test our models on all WS but also on each of them separately. Nevertheless, we observe the similar improvement of our multimodal semantic model with respect to the traditional text-based semantic model in both WS-Sim and WS-rel (refer to the next section 5.2.1 for more detailed reports). That means our new model can work well with both the semantic similarity aspect and semantic relatedness aspect.

We use the common Cosine computation because it is widely used in other studies; that provides us a fair comparison with other approaches :

$$\text{cosine(x,y)}=\frac{\Sigma_{i=1}^{n} x_i * y_i}{||x|| * ||y||}$$

To verify if the conclusion reached on WS extend to different semantic tasks, we use further famous benchmarks for which we have a good coverage: Rubenstein and Goodenough Similarity Judgments (**RG**), and Noun categorization. Although our results are not presented on fully covered testing sets, it is still good to have a look into the DM and other state-of-the-art models following the ACL Wiki site [2] just to have a relative comparison. In fact, the state of the art model brings about pretty similar results on our version of testing set to that on the corresponding fully covered version reported on the Wiki site.

Rubenstein-Goodeneough (RG) Our second challenge comes from the classic data set of Rubenstein and Goodenough (1965). It is a human-rating set

[2] http://aclweb.org/aclwiki/index.php?title=State_Of_The_Art

consisting of 65 noun pairs rated by 51 subjects on the scale of 0-4. The average rating for each pair represents the perceived similarity between them. For example:

```
boy-n lad-n 3.820
boy-n sage-n 0.960
boy-n rooster-n 0.440
automobile-n car-n 3.920
automobile-n cushion-n 0.970
automobile-n wizard-n 0.110
```

Generally, RG is quite similar to WordSim but it contains a smaller number of pairs. Following the earlier literature, we use *Pearsons r* to examine how good the cosines in the mulimodal semantic space between the nouns in each pair correlate with the ratings are. The results (expressed in terms of percentage correlations) are presented in Section 5.2 below . In our version of RG test, we have 47 pairs covered by the models, covering about 73 % of the full RG test. That coverage is quite similar to what we have with WordSim.

Noun categorization - AP and Battig We estimate our model's effectiveness on this task because the task of classifying words into classes or categories is a very important task in both computer science and cognitive science. It provides clear insights into concepts and meaning as well as the ability to hierarchically arrange concepts into taxonomies(Murphy, 2002). Moreover, it can show the potentiality of applying distributional information to various problems related to semantics (for additional information, refer to Baroni and Lenci, 2010). In this task, we use 2 well-known data sets: Almuhareb-Poesio (**AP**) (Almuhareb, 2006), and the new **Battig** set [7]. In short, AP set owns 402 concepts from WordNet, balanced in terms of frequency and ambiguity. They should be categorized into 21 distinguished groups by their relatedness of meaning. For example, *basketball, bowling* should be put into *game* class. Similar to AP set in term of structure, the Battig set consists of 82 concepts from 10 categories. Even though we do not have full coverages of those data sets, what we got (i.e, 230 concepts in AP and 72 concepts of Battig) is fairly distributed in all categorizes of the original data

sets. Therefore, we believe that the test is fair enough for assessing our proposed models/framework.

Examples of the AP data and Battig data include:

```
==AP==
acacia-n        tree
acceptance-n    legal
ache-n  pain
acne-n  illness
aeon-n  time
agency-n        socialunit
aircraft-n      vehicle
airplane-n      vehicle
airstream-n     atmospheric
allocation-n    assets
allotment-n     assets
....
==Battig==
aeroplane-n     vehicle
apple-n fruit
bean-n  vegetable
bear-n  land_mammal
bicycle-n       vehicle
birch-n tree
blender-n       kitchenware
blouse-n        clothes
boat-n  vehicle
bowl-n  kitchenware
bra-n   clothes
broccoli-n      vegetable
```

Following what successfully done in literature, we compute the *purity* score of noun clustering based on our distributional models by using CLUTO toolkit (Karypis, 2003) then compare with the published results using the same toolkit and experimental settings. To make it more clear, we calculate the similarities of n nouns and create the similarities matrix of size nxn. After that, we take the

matrix as the input for CLUTO's scluster algorithm with the repeated bisections with global optimization (*rbr*) clustering method. Cluster quality is evaluated by percentage purity (Zhao and Karypis, 2003). If n_r^i is the number of items from the i-th true (gold standard) class that were assigned to the r-th cluster, n is the total number of items and k the number of clusters, then:

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

In the best case (perfect clusters), purity is 100% and as cluster quality deteriorates, purity approaches 0.

BLESS Recently, (computational) linguists have raised a debate about reliability of the Spearman and Pearson coefficient scores in measuring semantic similarities. Even though they are considered as of the most trustworthy tools for that task, some argue that those scores are fragile. Providing an alternative solution, Baroni and Lenci bring forth the BLESS data set. (**BLESS**) is the "Baroni-Lenci Evaluation of Semantic Similarity" data set made available by the GEMS 2011 organizers[3]. The data set contains 200 concrete nominal concepts, each paired with a set of words that instantiate the following 5 relations: hypernymy (spear/weapon), coordination (tiger/coyote), meronymy (castle/hall), typical attribute (an adjective: grapefruit/tart) and typical event (a verb: cat/hiss). Concepts are moreover matched with 3 sets of randomly picked unrelated words (nouns, adjectives and verbs). For each true and random relation, the data set contains at least one word per concept, typically more. The relata were selected by the authors using rich sources such as WordNet, ConceptNet, Wikiapedia and the semantic norm of McRae and colleagues. More detailed descriptions can be found in the available publicity of BLESS.

Following the GEMS guidelines, we apply a model to BLESS as follows. Given the similarity scores provided by the model for a concept with all associated words within a relation, we pick the term with the highest score. We then *z-standardize* the 8 scores we obtain for each concept (one per relation), and we produce a boxplot summarizing the distribution of z scores per relation across the concepts (i.e., each box of the plot summarizes the distribution of the scores picked for each relations, standardized as we just described). Boxplots are produced accepting the default boxplotting option of the R statistical package[4] (boxes extend from

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

[4]http://www.r-project.org

first to third quartile, median is horizontal line inside the box).

5.1.2 Text-based semantic model

As discussed in the section 4.2, we keep the framework open to various DSMs. Thus, it could be best if we can do experiments on all available DSMs which have been showed successful in semantic representation. Since most of available DSMs are designed specifically to be good in some tasks (*one task, one distributional model*), doing experiments on all of them will make sense for a just evaluation on how effective the vision-based semantics contributes to text-based semantics. However, it obviously takes too much cost in computation and time. We would prefer leaving that for further research. Alternatively and fortunately, there exists a text-based semantic model namely Distributional Memory (**DM**) that is proven successful for many tasks at once [8]. It has been shown to be near or at the state of the art with a good variety of semantic tasks such as modelling similarity judgements, concept categorization, predicting selection preferences, relation classification and more. Considering DM's merit, we strongly believe that this model can be considered as a representative for cutting-edge DSMs in our evaluation tasks. We are justified in claiming that using DM not only saves our cost in computation and time but also satisfies our purpose to examine the strength of the multimodal semantic space as well as effectiveness of the vision-based semantic model. We lay stress on measuring effectiveness of vision-based semantic model because at the moment, there are not many studies in that area published, up to our knowledge.

The DM model (more precisely, the most successful TypeDM version) [5] is proposed by Baroni and Lenci (2010). It is trained on a large corpus of around 2.8B tokens coming from Web documents, the Wikipedia and the BNC. The DM is a structured DSMs in which links connecting collocates and target words are used to label the collocates themselves. It does not compute the score of the links by looking up only frequency of co-occurrence (or strength of association) but also variety of surface forms that express the links to avoid losing semantic information. The links are determined by a mixture of dependency parse information and lexico-syntactic patterns, resulting in distributional features such as in form of *as_adjective_as*, *subject_verb*, *attribute_noun*, etc. For example, for the word fat and the feature of animal, the raw score is 9 because fat co-occurs with

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

9 different forms of the feature *(a fat of the animal, the fat of the animal, fats of animal. . .)*. In total, there are 25.336 distinct links in DM and that can be seen as a tensor of size 30K x 25K x 30K with density 0.0005 % [8].

We transform the DM to its 2-way matrix form of semantic space with 30K rows (target words) represented in a space of more than 700M dimensions. Since our visual dimension extraction algorithms are maximally producing 32K dimensions *(see Section 5.1.3 below)*, we make the impact of text features on the combined model directly comparable to the one of visual features by selecting only the *topn* DM dimensions (with n varying as explained below). The top dimensions are picked based on their cumulative Local Mutual Information mass (see more details in algorithm 4.1). Again, we emphasize in the experiments below that trimming DM in this way does not have a negative impact on its performance, so that we are fair to argue we are adding visual information to a state-of-the-art text-based semantic space.

5.1.3 Vision-based semantic model

Image Data In our experiments, we use the ESP-Game data set[6] ESP game is a system that allows people to label images while playing and is proposed by von Ahn and Dabbish [78] and acquired by Google Inc. The game is played by two players not in communication. They are assigned number of images that both can see (one image at a time) and guess strings to label the image. Only a string appearing in common for both players is accepted. The ESP-Game data set is divided into 2 parts and published by the authors: the first one contains 50K images with tags (labels) that form a vocabulary of 11K distinct words. The image labels contain 6.686 tags on average (2.375 s.d.); the second one contains 100K images (mostly includes the first one) with similar tag's statistics. We did experiments on both of them instead of using the second one only. The reason is that there is a high number of parameters in visual information extraction algorithm, for example: number of clusters, step size for spatial pyramid and others which should be tested carefully to fit the framework. That will cost us a lot of time and computation, especially with a large image data set. Therefore, we explore the parameter settings on the first data set and evaluate with WS, then use the results as a criteria to extract features of later large data sets: second ESP-Data (100K images). Note that the first part contains 50K images, so we

[6]http://www.gwap.com/gwap/gamesPreview/espgame

consider it large enough to be the representative for parameters selection. We refer to them as **ESP50K**, **ESP100K** in our experimental reports.

From out point of view, ESP-Game is interesting since on one hand, it is rather large and we know that the tags are related to the images. On the other hand, ESP-Game is not the product of experts labelling representative images, but of a noisy annotation process of often poor-quality or uninteresting images (e.g, logos) randomly downloaded from the Web. Thus, our algorithms must be able to exploit large-scale statistical information while being robust to noise.

Visual Feature Extraction Generally, we use a standard pipeline in computer vision to extract visual features from images (Szeliski, 2010). Standard version **SIFT** descriptors are extracted on a regular grid with 5 pixels spacing, at four multiple scales (10, 15, 20, 25 pixel radii), zeroing the low contrast ones (our experimental results on ESP-Game data 50K shows those settings provides the best result in general). Descriptors are then quantized on a number of visual words that we varied between 250 and 2000 in steps of 250. We then computed a one-level 4x4 pyramid of spatial histograms (Grauman and Darrell, 2005), consequently increasing the features dimensions 16 times, for a number that varies between 4K and 32K, in steps of 4K. We used the VLFeat[7] implementation for the entire pipeline (Vedaldi and Fulkerson, 2008). From the point of view of our distributional semantic model construction, the important point to keep in mind is that standard parameter choices such as the ones we adopted lead to distributional vectors with 4K, 8K, . . . , 32K dimensions, where a higher number of features corresponds, roughly, to a more granular analysis of an image.

In addition, we use further types of SIFTs features enhancing color descriptions (we call **color SIFTs** to distinguish them from the standard **SIFT**) which have been showed successful in visual concept classification [75, 77] including **HSV-SIFT**, **OpponentSIFT**, **RGB-SIFT**, **rgSIFT**. The visual features are extracted by the implementation of Koen van de Sande [8] accepting best parameters from the authors. Briefly, the descriptors are extracted from every pixel in the image with Gaussan Derivative filter with a sigma of 0.667. Various colour channels are calculated separately and then concatenated together. While normal SIFT is described by 128 dimensional vectors, colour SIFTs are in the form of 384 dimensional vectors so we use PCA techniques to reduce the dimensionality

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

[8] http://koen.me/research/images/colordescriptors.png

by a factor of 3 in order to obtain 128 dimensional vectors.

All extracted features are then used to construct the vision-based semantic space as technically described in the section 4.3. In the experiments below, the SIFT vision-based model is referred as **image** and the color SIFT vision-based model is referred as **color-VM**.

5.1.4 Model Integration

Linear combination We observed that our visual information extraction procedure naturally results in vectors of dimensionality from 4K to 32K in steps of 4K. To balance text and image information, we use DM vectors coming from our topn algorithm, which also range from 4K to 32K in steps of 4K. However, we don't just combine a visual model with its related textual model of same dimensionality (e.g, 4K text with 4K vision, 8K text with 8K vision, etc.). We additionally combine each of visual feature vectors with each of textual feature vectors and obtain 64 combined models, in order to get an extensive outlook of the multimodal vector space model. We call those model **combined** models. Since in the experiments on WS (Section 5.1 below) we observe best performance with 32K text-based features, we report here later experiments with only (at least) 32K dimensions. Similar patterns to the ones we report are observed when adding image-based dimensions to text-based vectors of different dimensionalities. In the experimental result below, we refer to the text model of topn 32K features as **text**

Importantly, we remarked above our goal to compare impact of visual features to the textual features: adding visual features is as good as or better than adding further text-based features. We notice an improvement of adding visual features to the text-based model, and we should ask whether the same improvement could also be obtained by adding more text-based features. To control for this possibility, we also consider a set of purely text-based models that have the same number of dimensions of the combined models, that is, we compare the combined model of **d** dimensionality to the text-based **d** dimensionality. For example, the combined models of **64K** dimensionality (aka, model from 32K text-based model and 32K vision-based model) to the purely 64K text-based model, and so on and so forth. We refer to these models **text+** in our experimental reports. However, the comparison is only meaningful if we treat the features equally. That is why we use the **linear combination** for that target.

Linear weighted combination Leaving the consideration of a thoroughly fair comparison of text-based features and vision-based features aside, we are also interested in how good model our framework can generate and how effective vision-based features and text-based features compare to each other in our multimodal semantic space. We applied linear weighted combination for each of 8 text-based model (ranging from 4K to 32K in steps of 4K) with each of 8 vision-based model (also ranging from 4K to 32K in steps of 4K). The weighting parameter α varies from 0.0 to 1.0 in steps of 0.1 ($\beta = 1.0 - \alpha$). That is, $\alpha = 0.0$ means the combined model is actually the vision-based model and no information of text-based model takes part in the combination. $\alpha = 1.0$ brings about opposite meaning.

Parameter Estimation for combination We combine the 32K text-based model (text) and the 20K vision-based model by our Parameter Estimation algorithm (PE) (the reason why we choose those models will be explained in the section 5.2) . The reason is both of them own a large number of features and they show the best performance overall. We divide the set of features into groups of 4K features and then apply the PE algorithm for find the possibly optimal weighting schemes for these groups. We tune the PE on WordSim data set and with the ESP50K vision-based models, then take them over ESP100K vision-based models and other experimental benchmarks. The result indicates that our algorithm works pretty well. We will present further discussion on the following section.

5.2 Results and Discussion

5.2.1 WordSim

Linear combination The results on WS of the combined models with SIFT based on ESP50K are reported in the Table 5.1.

Among text-based topn models, the DM32K obtains the best result overall in term of Spearman coefficient. However, all *topn* models get quite similar results to each others, with very tiny differences. That additionally points up that dimensional size of trimmed DM models does not affect a lot the performance of semantic similarity measurement. In other words, trimming DM model by the *topn* algorithm can still preserve the quality of the text-based *state of the art* model. That judgement becomes more concrete with the result stated in the Table 5.2 where original DM, trimmed DM (text) and *text+* own quite a similar

+Visual features	*4K*	*8K*	*12K*	*16K*	*20K*	*24K*	*28K*	*32K*
DM4K	**50**	**48**	**49**	**49**	**49**	**49**	**49**	**49**
DM8K	**50**	**49**	**49**	**49**	**50**	**50**	**50**	**49**
DM12K	**50**	**49**	**50**	**49**	**50**	**50**	**50**	**50**
DM16K	**51**	**49**	**49**	**50**	**50**	**50**	**50**	**50**
DM20K	**51**	**49**	**50**	**50**	**50**	**50**	**50**	**50**
DM24K	**51**	**49**	**50**	**50**	**50**	**50**	**50**	**50**
DM28K	**51**	**50**	**50**	**50**	**50**	**50**	**51**	**50**
DM32K	**51**	**50**	**50**	**50**	**50**	**50**	**51**	**50**

Table 5.1: Performance of combined models on WordSim in Spearman coefficient (%)

result. Practically, the difference between them is far from significance in our significance tests on WS.

Starting from above assessments and results from Table 5.1 and 5.2, we are justified to appoint the DM32K as the representative for all trimmed models in further experiments.

Figure 5.1 presents the learning curves of the combined models against others on WordSim, WS-Sim and WS-Rel data sets. It suggests that the multimodal semantic models perform better than independent ones on both semantic relatedness and semantic similarity. Interestingly, both text-based model and vision-based model do a better job in semantic similarity measurement than relatedness measurement. The reason could be the relations of pairs in WS-Sim seem to be more straight forward (i.e, in the same category level, for example: *plane - car*) while the relations in WS-Rel is more indirect (i.e, one concept could belong to the sub-category of the other, for example *planet - galaxy*), thus, the distributional model can capture them better in WS-Sim pairs. Additionally, we see multimodal models stay above the text-based one constantly on WordSim and both of its subsets. The same overall result patterns are mostly identical in all 3 tests. Thus, we are confident that analysing the result on WordSim is fine enough to predict what is obtained on WS-Sim - *semantically similar* (e.g., synonyms or coordinate terms) and WS-Rel - *semantically related* (e.g., meronyms or topically related concepts).

In all tests, the purely image-based model is having the worst performance in all settings. Although even the lowest image-based Spearman score (*0.29*) is significantly above chance ($p. < 0.05$), suggesting that the model does capture some

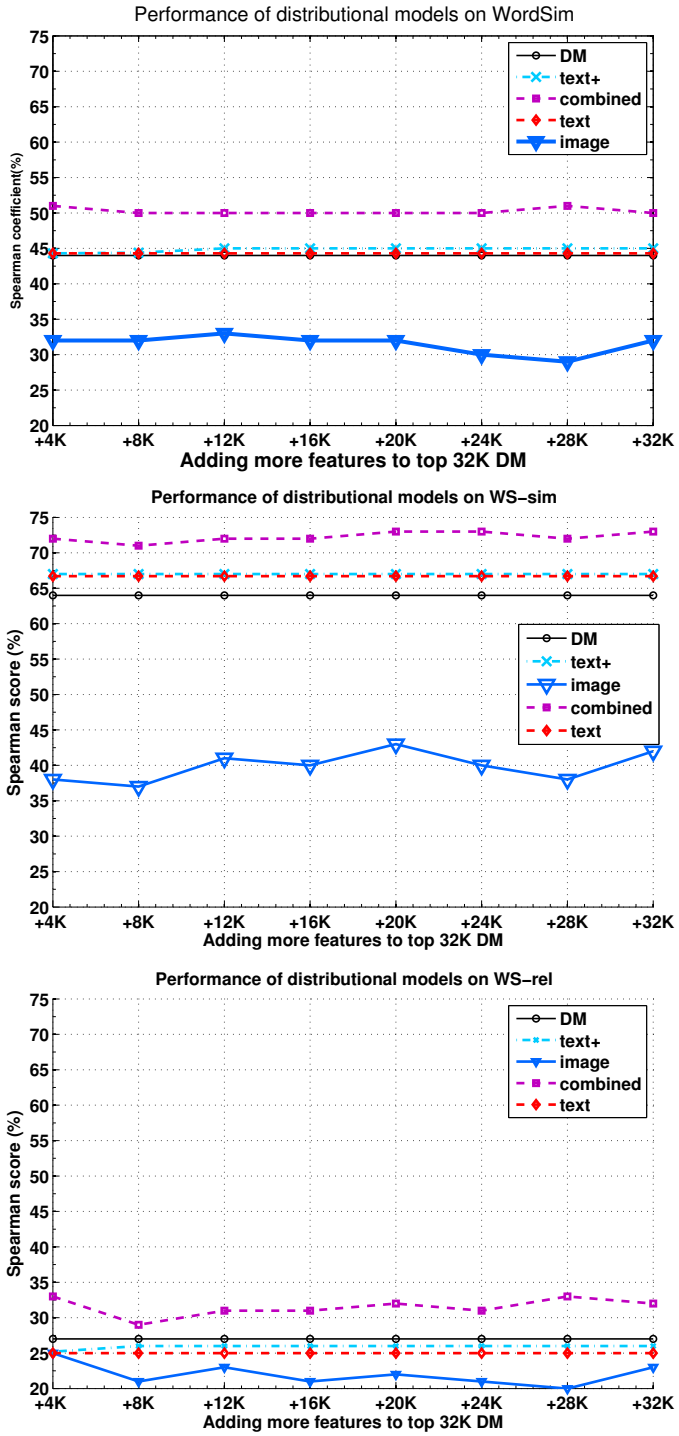

Figure 5.1: Performance of distributional models on WordSim

semantic information. Contrarily, adding image-based dimensions to a textual model (*combined*) consistently reaches the best performance, also better – for all choices of dimensionality – than adding an equal number of text features (*text+*) or using the full *DM* matrix..

Based on the results reported in Figure 5.1, further analyses will focus on the *combined* model with +20K image-based features, since performance of *combined* does not seem to be greatly affected by the dimensionality parameter, and performance around this value looks quite stable (it is better only at the boundary +4K value, and with +28K, where, however, there is a dip for the *image* model). The *text+* performance is not essentially affected by the dimensionality parameter, and we pick the +20K version for maximum comparability with *combined* (with WS, they share the same Spearman coefficient score).

The Table 5.2 shows the comparison of different models from above results.

model	*WordSim*	*model*	*WordSim*
DM	44	image	32
text	44	combined	50
text+	45	-	-

Table 5.2: WordSim Spearman coefficient of distributional models

The difference between *combined* and *text+*, although consistent, is not statistically significant according to a two-tailed paired permutation test [54] conducted on the results for the +20K versions of the models. Still, very interesting qualitative differences emerge. Table 5.3 reports those WordSim pairs (among the ones with above-median human-judged similarity) that have the highest and lowest *combined*-to-*text+* cosine ratios, i.e., pairs that are correctly treated as similar by *combined* but not by *text+*, and *vice versa*. Strikingly, the pairs characterizing the image-feature-enriched *combined* are all made of concrete, highly imageable concepts, whereas the *text+* pairs refer to very abstract notions. We thus see here the first evidence of the complementary nature of visual and textual information.

While general SIFTs features contribute well to the **text** model, other color channels of **color SIFT** does not show the same ability. The Table 5.4 shows the performance of models based on color SIFTs.

The results of color VMs are much lower than what is delivered from general SIFT although they are still above the chance. They don't act marvelously in the combining tasks with text-based models and result in worse Spearman coefficient

multimodal model	*text+*
tennis/racket	physics/proton
planet/sun	championship/tournament
closet/clothes	profit/loss
king/rook	registration/arrangement
cell/phone	mile/kilometer

Table 5.3: WordSim pairs with highest (first column) and lowest (second column) *combined*-to-*text+* cosine ratios

name	*combined*	*color-VM*
Opposite-SIFT	36	19
RGB-SIFT	37	19
rgSIFT	37	19
HSV-SIFT	39	21

Table 5.4: Performance of color SIFTs feature-based models

scores than the ones of the text-based model. The reason could be *colorSIFTs* place a heavy emphasis on different separated color-channels while our image data set doesn't bring rich enough color-oriented information about semantic relations (typically, the ESP data is in poor definition as well as resolution). So for further experiments, we more concentrate on the general SIFT and its respective vision-based models, nevertheless, from our point of view, it does not mean color SIFTs can't capture semantic relation among images. We just put a starting brick in the problem of exploring various types of visual features in semantic representation task. As said, we would like to keep our framework simple as a prototype and leave the job of employing sophisticated computer vision studies for further researches.

Weighted linear combination We pick the **text** model for the combining test with all of our vision-based models (again, only those from general SIFT). The result is presented in the Figure 5.2. It is pretty clear that combined models own better results than independent ones (purely image-based models is obtained when $\alpha = 0.0$ and pure text-based model is obtained when $\alpha = 1.0$). The models reach the highest scores when α stays around in the middle of the range [0.0 - 1.0]. Based on that results, we claim that α should be the best for our multimodal vector space at 0.5. Its overall results are higher than results at other values

except 0.4, but when $\alpha = 0.4$, combinations result are less stable (We can figure out from the Figure 5.2 that the variation of results at 0.4 is much larger than that at 0.5). That result is very attractive for us because it means vision-based model and text-based model share the even importance level in the multimodal space.

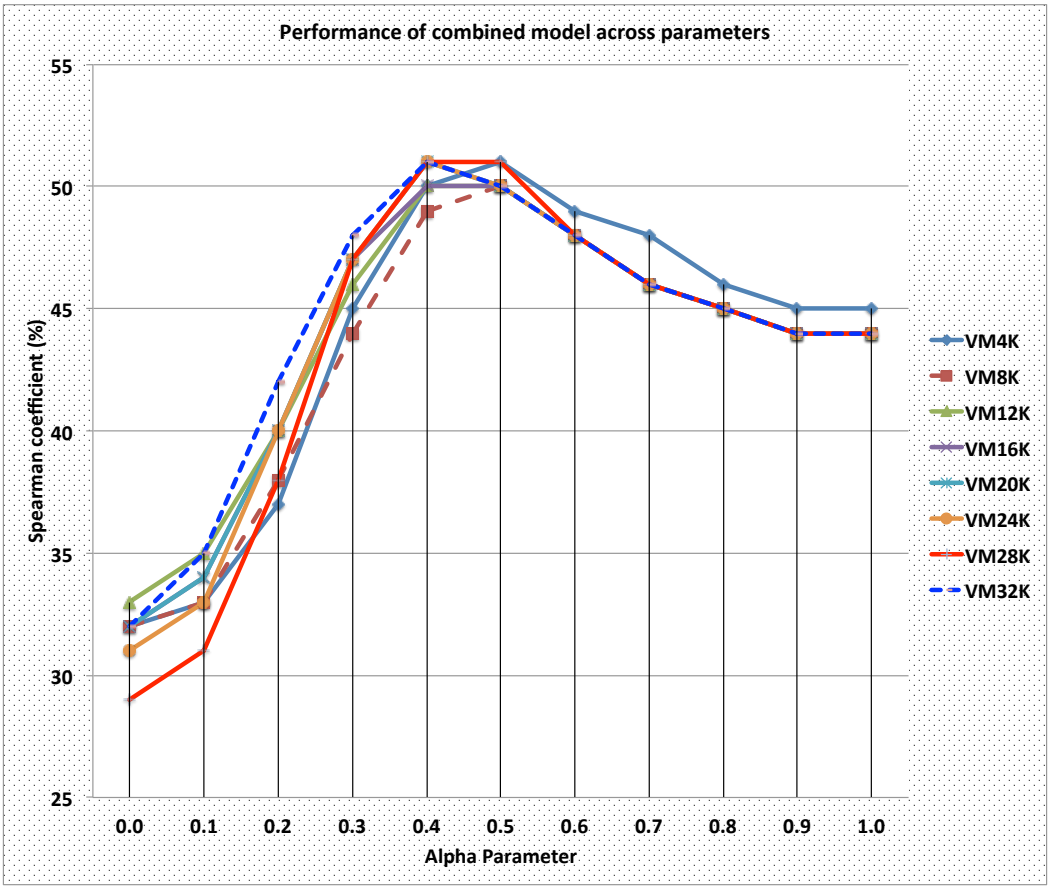

Figure 5.2: Performance of distributional models on WordSim

Combination with parameter estimation (PE) algorithm PE algorithm works pretty well in the task of similarity measurement on WordSim with much higher results than what attained by other linear techniques. The Table 5.5 shows its overall Spearman coefficient scores against our previous models, where the model coming from parameter estimation algorithm is marked as **PE**. The distance between our PE-model and other models is not bad with 14% higher than *text+* and 9% higher than *combined*. It suggests us that weighting the features relying on their cumulative scores is indeed effective for our model. We

will observe that judgement clearer when we apply the parameters' value tuned from ESP50K on WordSim on other image data sets and other tests later.

Increasing number of images From tuning results we got from ESP50K, we do experiments with ESP100K where the size of image data is double. The Table 5.5 shows the results of the models coming from ESP100K in comparison with other models[9].

model	*WS*	model	WS
DM	44	Strube and Ponzetto (2006) Wikipedia	19-48
text	44	jarmasz (2003) Roget's	55
text+	45	Hughes and Ramage (2007) Wordnet	55
image	32	Agirre et al. (2009) Wordnet	56
combined	50	Farrington (2010) Web corpus LSA	56
PE	**59**	Harrington (2010) Sem. Network	62
image-esp100k	33	Agirre et al. (2009) Web corpus	66
combined-esp100k	55	Agirre et al. (2009) Wordnet + gloss	66
PE-esp100k	**65**	Gabrilovich and Markovitch (2007) Wikipedia	75

Table 5.5: WordSim Spearman coefficient experiment of distributional models

As remarked in earlier results on ESP50K, images can capture semantic relations among words/concepts and adding further visual features is at least as good as adding more text features. They also show capacity in enhancing semantic measurement between concrete concepts. That statement is stronger with the results relying on large image data set. Generally, it is fascinating that **image** models all stay above chance and combined models bring about better results than **text+** models, but it is even more interesting that increasing number of images dramatically affect the quality of multimodal semantic space. From the Table 5.5, the *combined-esp100k* earns 5% more than *combined (esp50k)* and the *PE-esp100k* earn 10% more than *PE (esp50k)*. Comparing to purely text-based models, the difference reaches 20 %. This leads us to concretely confirm our previous conclusion: on WordSim, images actually capture semantic relations and adding visual features is better than adding further text features from the state of the art text-based model.

[9]Results of existed models in literature listed in this table are from Wojtinnek et al. (2010)

Staying apart from the successful story about image itself, we can see that our combination techniques greatly help the multimodal space models, especially with the parameter estimation (PE) algorithm where the relatively improvement is much better than with the linear combination.

Taking a look at the state of the art league in the Table 5.5, we see that our model is comparable to the top corpus-based methods in spite of the fact that they are based on much large corpus than which DM is based on. It suggests a promising result can be achieved by exploiting visual information as an alternative source.

5.2.2 Rubenstein-Goodeneough

The Figure 5.3 provides the outline of the performance of our multimodal models on our version of RG (to remind, we cover 73 % of the original RG). It reports all combination of *text* with vision-based models coming from ESP50k and ESP100K.

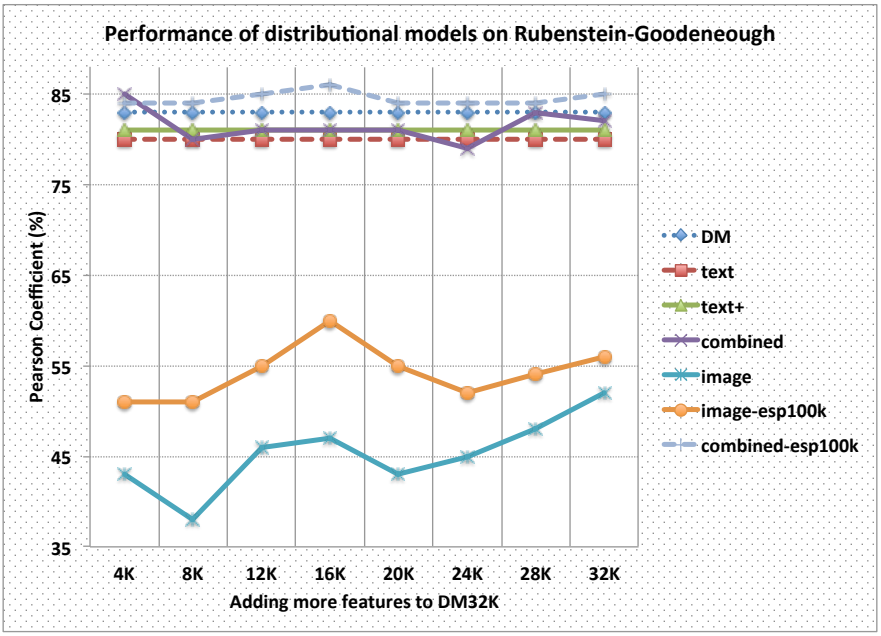

Figure 5.3: Performance of distributional models on Rubenstein-Goodeneough

The foremost observation we get here is that the result is pretty similar to what we got with WordSim. It gets us more confident to confirm the conclusion above. However, we just have a very limited number of pairs participating on

this test (i.e, 47 pairs covered, much smaller than what we have in WordSim with 260 pairs covered). That explains why the performance of models is less stable that what we get in the WordSim. From the overall result pattern, we can see that the *text* and *text+* have a similar result and it is a bit worse than the original DM models. Also, the multimodal models are not worse than all of them (especially *combined-esp100k* is constantly above all of vision-based only models and text-based only models)

Once more, It is clear that the image can capture semantic relations, that is, vision-based only models stays above chance and with more images, the results get higher. Homogeneously, the more images, the more semantic relations the multimodal DSM can capture. We can see that vividly from the Table 5.6[10] which presents results of all representative models (to remind, we also pick up the combination *text* +20K vision-based model for multimodal model construction)

model	*RG*	*model*	*RG*
DM	83	Chen et al. (2006) DoubleCheck	85
text	80	Herdagdelen et al. (2009) SVD-09	80
text+	81	Pado and Lapata (2007) DV-07	62
image	43	Pado and Lapata (2007) cosDV-07	47
combined	81	-	-
PE	**87**	-	-
image-esp100k	55	-	-
combined-esp100k	84	-	-
PE-esp100k	81	-	-

Table 5.6: RG Pearson coefficient experiment of distributional models

Based on results on the Table 5.6, again, the highest result belongs to our *PE* model obtained by parameter settings tuning on WordSim. However, the *PE* model related to esp100k has a lower result than that of linear *combined* model, even it is still equal to *text+*. The reason, as we emphasized from beginning, might be that our version of RG is too small with only 47 pairs and normally makes our models unstable like what happened in the beginning part of this section. That makes us less trustful in our version of RG than WordSim, but still the RG results are very interesting. Although we don't have a full cover for

[10]The results of existed models in literature are from Baroni and Lenci (2010)

RG but relatively comparing to existed models, we believe our multimodal model can stay in the top of the state of the art line, for example, DM provides quite similar result on the full RG (see more details in [8]); and our best model (PE) is comparable to the DoubleCheck, which is holding one of the best scores on RG and even though it is an unstructured system that relies on Web queries (and thus on a much larger corpus) [8].

We do a further analysis in how different the multimodal model (PE) and *text+* model are by repeating the qualitative analysis we did with WordSim. The result is indicated in the Table 5.7

multimodal model	*text+*
autograph/signature	magician/oracle
cook/rooster	serf/slave
cushion/pillow	brother/monk
bird/crane	magician/wizard

Table 5.7: WordSim pairs with highest (first column) and lowest (second column) *PE*-to-*text+* cosine ratios

Another time, the multimodal model list is made of more concrete concepts than what in the *text+* list like what we observed with WordSim. The difference here is the *text+* pairs look more concrete than the *text+* pairs in the experiments with WordSim above, but it is because RG doesn't contains purely abstract concepts. It emphasizes that our visual features indeed contribute more to the concrete concepts' semantic representation.

5.2.3 Concept Categorization

Table 5.8[11] reports percentage purities in the AP and Battig clustering tasks for full *DM* and the representative models discussed above.

The *image* model alone is not at the level of the text models, although both its AP and Battig purities are significantly above chance ($p < 0.05$ based on simulated distributions for random cluster assignment). Thus, even alone, image-based vectors do capture aspects of meaning. For AP, adding image features does improve performance, although the best multimodal model *PE* is not significantly

[11]The results of existed models in literature are from Baroni and Lenci (2010)

model	*AP*	*Battig*
DM	81	96
text	79	83
text+	80	86
image	25	43
combined	78	96
PE	**83**	**97**
image-esp100k	25	52
combined-esp100k	69	91
PE-esp100k	74	95
—	—	—
Rothenhausler and Schutze (2009) DepPath	79	NA
Almuhared and Poesio (2005) AttrValue-05	71	NA
Herdagdelen et al. (2009) VSM	70	NA
Baroni et al. (2010) Strudel	NA	91
Baroni et al. (2010) DV-10	NA	79
Baroni et al. (2010) AttrValue	NA	45

Table 5.8: Percentage AP and Battig purities of distributional models

better than the *text+* (by a two-tailed paired permutation test of difference between *text+* and *PE*). For Battig, adding visual features (refer to *combined* and *PE*) improves on the purely text-based models based on a comparable number of features (although the difference between *text+* and *PE* is not significant), reaching a modestly better performance than the one obtained with the full *DM* model (that in these categorization tests is slightly above that of the trimmed models). Intriguingly, the Battig test is entirely composed of concrete concepts, so the difference in performance for *combined* might be related to its preference for concrete things we already observed for WordSim.

Strangely, the ESP100K-based multimodal models have worse performance than the ESP50K-based models on both AP and Battig noun categorizing test. It is opposite with the results we obtained from WordSim and RG experiments in the previous experiments, that is, increasing number of images does not enhance the model pretty much. Since we still believe in our assumption that: images can captures semantic relations - so more images, we should have better performance, we think a bottleneck problem here is our techniques in extracting visual features.

With more images, we are supposed to use difference settings instead of those tuned in ESP50K (twice smaller data set).

However, the best multimodal model based on ESP100K (*PE-esp100k*) is still not worse than the *text+*. Its result is lower in AP but higher in Battig than that of the *text+*. All significance tests between their results shows that the difference is still far from significance.

Looking at the combination techniques, our PE algorithm again is the winner as it carries out best job done in both AP and Battig. The linear combination is not bad but still not on the same line with the PE.

To sum up, the concept categorization experiments do not confirm the tendency on WordSim for increasing the number of images to improve performance. Nonetheless, the experiments support what we observed on WordSim, that is, visual features do not harm performance. The results show that the multimodal model is slightly better than the purely text-based models, even though they are not statistically significance different. The next test will give us some insights on how visual features affect the behaviour of the models, independently of performance.

5.2.4 BLESS

The BLESS distributions of text-based models (including *multimodal model (DSM)*) are similar while those of the image-based only models are largely different so we use here the full *DM* model as representative of the text-based set and the *PE-esp50k* model as representative of the all combined models in Figure 5.4 and – their boxplots are then compared to the ones of the purely *image*-based models in Figure 5.5.

We see that purely text-based *DM* cosines capture a reasonable scale of taxonomic similarity among nominal neighbours (coordinates then hypernyms then meronyms then random nouns), whereas verbs and adjectives are uniformly very distant, whether they are related or not. This is not surprising because the *DM* links mostly reflect syntactic patterns, that will be disjoint across parts of speech (e.g., a feature like *subject_kill* will only apply to nouns, save for parsing errors).

Those relations are also captured by the multimodal model. The only difference between the multimodal model and DM lies on the scale of *attributes/events, random nouns/adjectives/verbs*. That suggests the multimodal model is a bit better than DM at differentiating related attributes/events/random adjectives/ran-

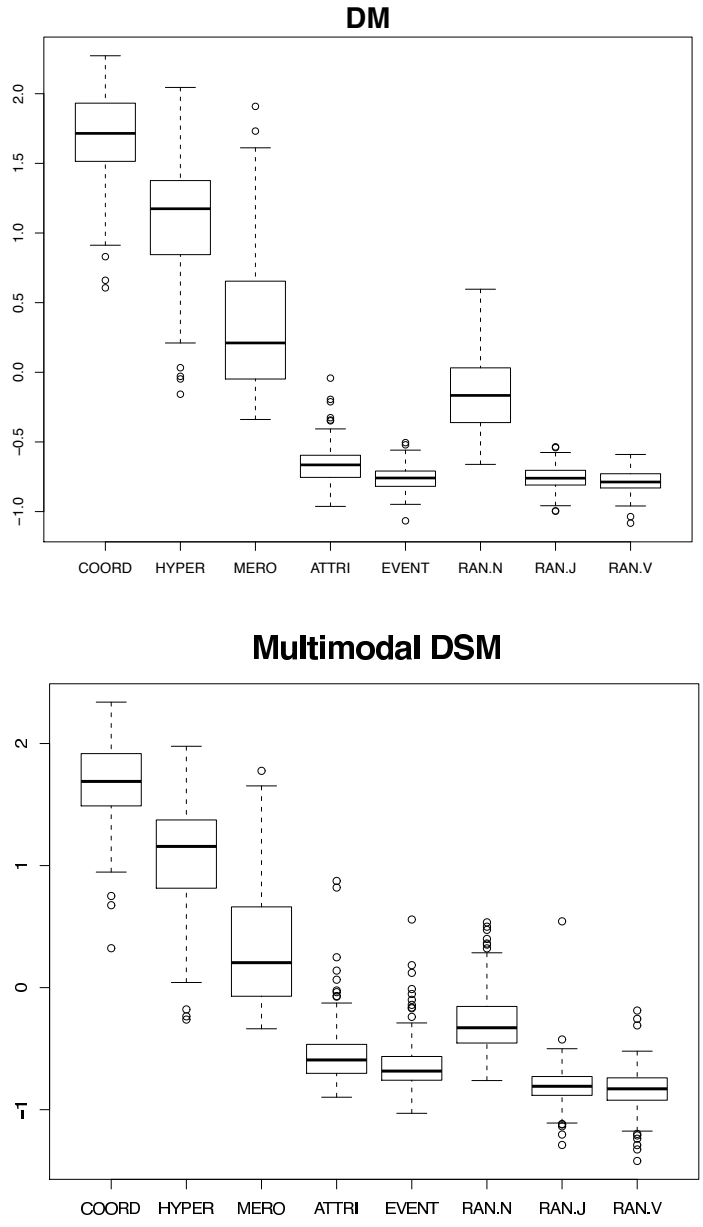

Figure 5.4: Distribution of z-normalized cosines of words instantiating various relations across BLESS concepts of DM and multimodal DSM.

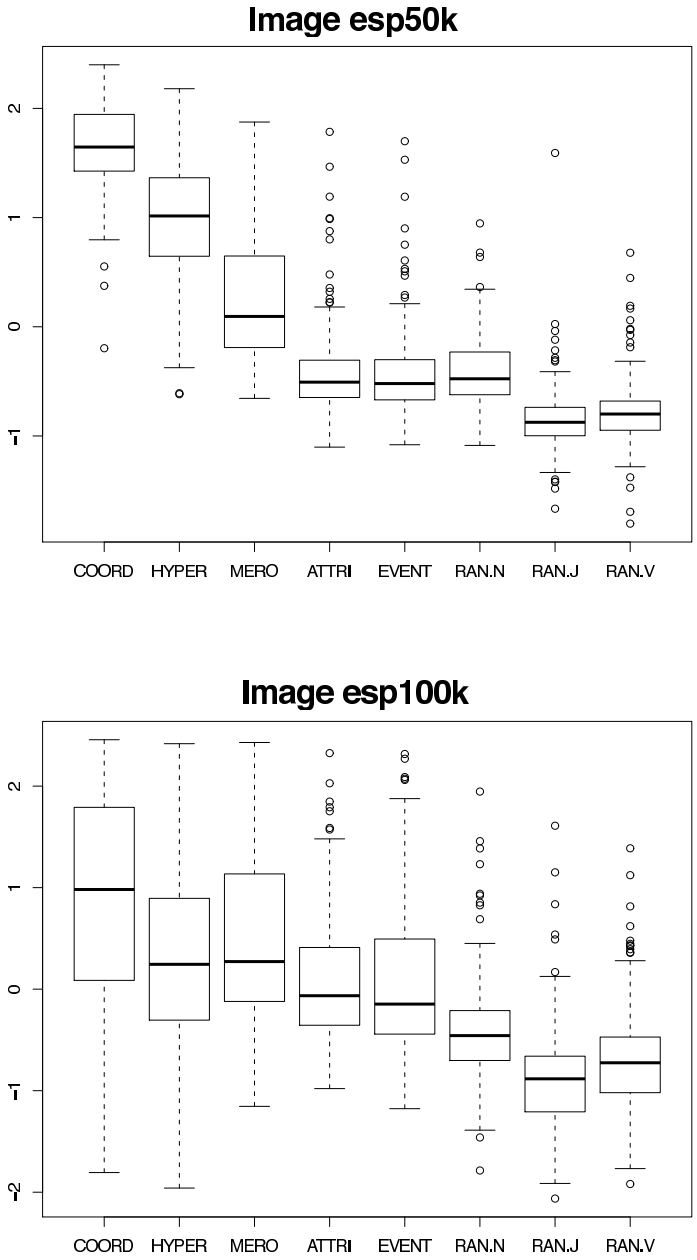

Figure 5.5: Distribution of z-normalized cosines of words instantiating various relations across BLESS concepts of image-base models.

dom verbs and a bit worse at avoiding random nouns. While DM is only able to discriminate between attributes and random verbs, the multimodal model can additionally distinguish attributes from random adjectives more.

Looking at the *image-only models*, we first observe that the *esp50k image* based only model can capture differences between related attributes/events and random adjectives/verbs (according to a Tukey HSD test for all pairwise comparisons, these differences are highly significant, whereas *DM* only significantly distinguishes attributes from random verbs). In this respect, *image esp50k* is arguably the "best" model on BLESS. However, perhaps more interestingly, the *image-esp50k* model also shows a biased for nouns, capturing the same taxonomic hierarchy found for *DM*. This suggests that image analysis is providing a decomposition of concepts into attributes shared by similar entities, that capture ontological similarity beyond mere syntagmatic co-occurrence in an image description. What is really astonishing is that the esp100k image-based only model can be the "worst" model at all, regarding some discussed aspects above. It can not make a distinction of coordinates, hypernyms then meronyms like others. It is also bias to nouns an can't capture the taxonomic hierarchies. Once more, we observe the poorer quality of the vision-based model coming from esp100k. It leads us suspect (more) that perhaps the settings tuned on ESP50k data set are not suitable for a twice-time bigger dataset. However, the *esp100k model* still can discriminate the difference of random verbs and random adjectives among relations.

Basing on merits of 2 vision-based (only) models (*image esp50k and image esp100k*), we can conclude that vision information still admits a potential capacity to capture semantic differences among concepts.

Considering that the *image esp50k* model outperforms the *image esp100k* model, we choose the *image esp50k* model as the representative for generally purely image-based models and use it for our next experiments (so, we name it *image* model). To support the conclusion from the paragraph above, we counted the average number of times that the related terms picked by the *image* model directly co-occur with the target concepts in an ESP-Game label. It turns out that this count is higher for both attributes (10.6) and hypernyms (7.5) than for coordinates (6.5). So, the higher similarity of coordinates in the image model demonstrates that its features do generalize across images, allowing us to capture "attributional" or "paradigmatic" similarity in visual space. More in general, we find that, among all the related terms picked by the *image* model that have

an above-average cosine with the target concept, almost half (41%) *never* co-occur with the concept in the image set, again supporting the claim that, by our featured analysis, we are capturing visual properties of similar concepts beyond their co-occurrence as descriptions of the same image.

A final interesting point pertains to the specific instances of each (non-random) relation picked by the textual and visual models: of 870 related term pairs in total, almost half (418) differ between *DM* and *image*, suggesting that the boxplots in Figure 5.4 hide larger differences in what the models are doing. The randomly picked examples of mismatches in top attributes from Table 5.9 clearly illustrate the qualitative difference between the models, and, once more, the tendency of *image*-based representations to favour (not surprisingly!) highly visual properties such as colours and shapes, vs. the well-known tendency of text-based models to extract systemic or functional characteristics such as *powerful* or *elegant* [7]. By combining the two sources of information, we should be able to develop distributional models that come with more well-rounded characterizations of the concepts they describe.

concept	*DM*	*image*	*concept*	*DM*	*image*
ant	small	black	potato	edible	red
axe	powerful	old	rifle	short	black
cathedral	ancient	dark	scooter	cheap	white
cottage	little	old	shirt	fancy	black
dresser	new	square	sparrow	wild	brown
fighter	fast	old	squirrel	fluffy	brown
fork	dangerous	shiny	sweater	elegant	old
goose	white	old	truck	new	heavy
jet	fast	old	villa	new	cosy
pistol	dangerous	black	whale	large	gray

Table 5.9: Randomly selected cases where nearest attributes picked by DM and *image* differ.

5.3 Summary

In conclusion, we have carefully evaluated tasks ranging from word similarity judgement, concept clustering to BLESS semantic relationship differentiating in

order to measure how good our multimodal model is. The overall results are high with 65 % Spearman coefficient score in WS, 87% Pearson coefficient score in RG for similarity measurement, 83% and 97% respectively for AP and Battig clustering tasks. The multimodal model can also recognize/differentiate most of semantic relationship in BLESS (we don't have a relative comparison to other results on BLESS yet because it is newly introduced). Different to almost all published models that are designed and tuned for a task at hand, our model works well on all tests gold-standard tests (we do experiments on all of available tests to evaluate semantic representation, which are possible for our model's architecture, except TOEFL test of synonym detection). The result indicates that our multimodal model can (or very closely) stay in the top state of the art level.

More interestingly, the results indicate that visual information certainly captures semantic relations among words/concepts. The purely image-based models, while worse than the more advanced models, still achieve above-chance performance. Especially, in the BLESS test, purely image-based model is robust in differentiating the semantic relationships, is even better than the state of the art text-based models. To confirm our judgement, we did the qualitative analyses and successfully recognized the evidence that visual information can capture better semantic relations among *concrete concepts*.

Our various experiments show that adding further visual information is comparable to adding further text information. In similarity measurement and concept clustering tests, our multimodal model earns higher results than the text-based model of the same feature dimensionality. Additionally, the BLESS test indicates the multimodal model can distinguish attributes from random adjectives, while no purely text-based model can.

Considering all factors above, we are safe to claim that the two sources of information (text and images) are complementary in semantic representation.

Chapter 6

Conclusion and Future Work

6.1 Thesis Contribution

We proposed the first framework to integrate a state of the art text-based semantic model and a vision-based semantic model to create a multimodal semantic model. The framework is designed as an open prototype for further studies / analysis in both computer vision and computational linguistics.

We proposed an effective method to augment a state-of-the-art text-based distributional semantic model with information extracted from image analysis. The method is based on the famous bag-of-visual-words representation of images in computer vision where a "visual" word is a cluster of keypoint descriptor SIFT. The image-based distributional profile of a word is encoded in a vector of co-occurrences with "visual words", that we concatenate with a text-based co-occurrence vector.

We showed evidences that the image-based semantic model tends to capture semantic relations among concrete concepts while the text-based semantic model tends to capture semantic relations among more abstract concepts.

6.2 Result Summary

In all result reports, adding image-based features earns higher absolute scores than adding further text-based features, but the statistical significance tests show that the differences between them are not close to significance level. However, the

result is still very interesting because we can safely claim that adding image-based features is at least not damaging, when compared to adding further text-based features, and largely possible beneficial.

Especially, the experiments in the Chapter 5 briefly explored an interesting aspect of the semantic relations image-based features are capturing. We find that image-based features lead to interesting qualitative differences in performance: Models including image-based information are more oriented towards capturing similarities between concrete concepts, and focus on their more imageable properties, whereas the text-based features are more geared towards abstract concepts and properties. Our experiments show a preliminary evidence for an integrated view of semantics where the more concrete aspects of meaning derive from perceptual experience, whereas verbal associations mostly account for abstraction.

In addition, we observe that the general SIFT visual features work better than color oriented SIFT visual features such as *HSV-SIFT*, *OpponentSIFT*, *RGB-SIFT*, *rgSIFT* in semantic similarity measurement task. It does not mean the color SIFTs are bad for our task but it suggests visual features which reflect only one isolated channel of color information might not capture semantic relations well, in spite of the fact that image-based representations tend to characterize visual or color properties of objects in images.

Through all of our experiments, our proposed algorithm namely "Parameter estimation" (PE) works productively and outperforms the naive linear combination method in the task of creating a multimodal semantic model by integrating text-based and vision-based features. The algorithm is based on the idea of extracting top features and weighting them differently, instead of treating them evenly. Our best model is the multimodal model coming from applying the PE algorithm on the ESP50K-based visual model and DM. We haven't successfully proved that increasing number of images will improve the quality of multimodal model but we believe it is just the matter of computer vision techniques. We will leave that proof task with more sophisticated techniques from computer vision community for our next studies.

6.3 Future Work

In future work, we plan first of all to improve performance, by focusing on visual word extraction and on how the text- and image-based vectors are combined

(possibly using supervision to optimize both feature extraction and integration with respect to semantic tasks). However, the most exciting direction we intend to follow next will concern evaluation, and in particular devising new benchmarks that address the special properties of image-enhanced models directly. For example, Baroni and Lenci (2008) observe that text-based distributional models are seriously lacking when it comes to characterize physical properties of concepts such as their colors or parts. These are exactly the aspects of conceptual knowledge where image-based information should help most, and we will devise new test sets that will focus specifically on verifying this hypothesis.

Last but not least, we come back to our discussion in the embodied literature reviews: perceptual information from images indeed captures semantic relations. In a nutshell, by exploiting visual distributional information, we may break the limit: the description of objects based on text may loose its accuracy (e.g, "green banana", or "red light", ...), and that will help us in widely applied scopes of applications: from solving general tasks mentioned in the recent survey of Turney and Pantel (2010), such as query expansion, information retrieval or word sense disambiguation, more ambitiously, to constructing a learning system that is comparable to the human learner.

Bibliography

[1] Andrews, M., Vigliocco, G. 2009. Learning Semantic Representations with Hidden Markov Topic Models. Proceedings of the 31st Annual Meeting of the Cognitive Science Society.

[2] Andrews, M., Vigliocco, G. 2010. The Hidden Markov Topic Model: A Probabilistic Model of Semantic Representation. Topics in Cognitive Science, Vol. 2, pp. 101-113.

[3] Vigliocco, G., Meteyard, L., Andrews, M., Kousta S. 2009. Toward a Theory of Semantic Repre- sentation. Language and Cognition, Vol. 1(2).

[4] Andrews, M., Vigliocco, G., Vinson, D. 2009. Integrating Experiential and Distributional Data to Learn Semantic Representations. Psychological Review, Vol. 116(3), pp 463-498.

[5] Almuhareb, Abdulrahman and Poesio, Massimo . 2004. Attribute-based and value-based clustering: An evaluation. In Proceedings of EMNLP, pages 158165, Barcelona, Spain.

[6] Baroni, Marco and Alessandro Lenci. 2008. Concepts and properties in word spaces. Italian Journal of Linguistics, 20(1):5588.

[7] Baroni, Marco and Eduard Barbu, Brian Murphy, and Massimo Poesio. 2010. Strudel: A distributional semantic model based on properties and types. Cognitive Science, 34(2):222254.

[8] Baroni, Marco and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. Computational Linguistics 36 (4): 673-721.

[9] Barsalou, Lawrence; and Ava Santos, Kyle Simmons, and Christine Wilson, 2008. Language and Simulation in Conceptual Processing, chapter 13, pages 245283. Oxford University Press, USA, 1 edition.

[10] Bloom, P. (2000). How Children Learn the Meanings of Words. The MIT Press. Cambridge, Mass.

[11] Bosch, A. and A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In Proceedings of the ACM International Conference on Image and Video Retrieval, pages 401408, Amsterdam, The Netherlands, 2007.

[12] Bullinaria, John and Joseph Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, 39:510526.

[13] Bruni, Elia. and Tran, Giang Binh. and Baroni, Marco. 2011. Distributional semantics from text and images. EMNLP GEMS Workshop. Edinburgh - UK, July.

[14] Chen, L.D., Mooney, J.R. 2008. Learning to Sportscast: A Test of Grounded Language Acquisition In Proceedings of the 25th International Conference on Machine Learning (ICML) , Helsinki, Finland.

[15] Chen, L.D., Kim, J., Mooney, J.R. 2010. Training a Multilingual Sportscaster: Using Perceptual Context to Learn Language. Journal of Artificial Intelligence Research 37 (2010), 397-435.

[16] Curran, James and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition, pages 5966, Philadelphia, PA, USA.

[17] Csurka, Gabriella and Christopher Dance, Lixin Fan, Jutta Willamowski,andCe dricBray. 2004. Visualcategorization with bags of keypoints. In In Workshop on Statistical Learning in Computer Vision, ECCV, pages 122.

[18] Collins, A. M., Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240-248.

[19] Erk, Katrinand and Sebastian Pado.2008.A structured vector space model for word meaning in context. In Proceedings of EMNLP, pages 897906, Honolulu, HI, USA.

[20] Evert, Stefan. 2005. The Statistics of Word Cooccurrences. Dissertation, Stuttgart University.

[21] Evert Stefan, Distributional Semantics Models, NAACL-HLT Tutorial Lecture, NAACL-HLT, 2010.

[22] Feng, Y. and Lapata, M. 2008. Automatic Image Annotation Using Auxiliary Text Information. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 272-280. Columbus, OH.

[23] Feng, Y. and Lapata, M.2010. Visual Information in Semantic Representation. In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 91-99. Los Angeles, CA.

[24] Feng, Y. and Lapata, M. 2010. Topic Models for Image Annotation and Text Illustration. In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 831-839. Los Angeles, CA.

[25] Finkelstein, Lev and Evgeniy Gabrilovich and Yossi Matias and Ehud Rivlin and Zach Solan and Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept Revisited. ACM Transactions on Information Systems, 20(1):116131.

[26] Glenberg, Arthur. 1997. What memory is for. Behav Brain Sci, 20(1), March.

[27] Grauman, Kristen and Trevor Darrell. 2005. The pyramid match kernel: Discriminative classification with sets of image features. In In ICCV, pages 14581465.

[28] Griffiths L.T., and Joshua B. Tenenbaum and Mark Steyvers. 2007. Topics in semantic representation. Psychological Review, 114:211244.

[29] Gupta, S., Mooney, J.R. 2010. Using Closed Captions as Supervision for Video Activity Recognition In Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, GA, pp. 1083- 1088.

[30] Harnad, S. 1990. The symbol grounding problem. Physica D: Nonlinear Phenomena, 42:335-346

[31] Harrington, B. A semantic network approach to measuring semantic relatedness. In COLING, 2010.

[32] Harrington, B. and S. Clark, Asknet: automated semantic knowledge network. In AAAI, 2007.

[33] Harris, Zellig. 1954. Distributional structure. Word, 10(2-3):14561162.

[34] Herzog, G. and P. Wazinski. 1994. VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions. In: Artificial Intelligence Review, 8(2/3):175-187.

[35] Hinton, J., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. Psychological Review, 98, 74-95.

[36] Hinrich Schutze. Dimensions of meaning. In proceedings of Supercomputing. 1992

[37] Hinrich Schutze. Word space. In Stephen Jose Hanson, Jack D. Cowan, and C. Lee Giles, editors, Advances in Neural Information Processing Systems 5, pages 895902. Morgan Kaufmann Publishers, San Mateo CA, 1993b

[38] Hinrich Schutze. Ambiguity Resolution in Language Learning. CSLI Publications / University of Chicago Press, 1997

[39] Huiskes M. J. and M. S. Lew (2008). The MIR Flickr Retrieval Evaluation. ACM International Conference on Multimedia Information Retrieval (MIR'08), Vancouver, Canada.

[40] Hawkins, Jeff. 2008. Hierarchical Memory: Computing Beyond Turing, Keynote Speech at the Artificial Intelligence Conference, RSA 2008, San Francisco, USA.

[41] Kelleher, J.D. and Costello, F. and van Genabith, J. 2005. Dynamically Updating and Interrelating Representations of Visual and Linguistic Discourse. Artificial Intelligence 167 62-102.

[42] Islam, A. and Inkpen, D. 2008. Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Discov. Data. 2, 2, Article 10 (July 2008).

[43] Lund, Kevin and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, 28:203208.

[44] Landauer, Thomas and Susan Dumais. 1997. A solution to Platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2):211240.

[45] Lazebni S., Schmid C., Ponce J. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In the Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR06). pp 2169-2178

[46] Lowe, David G. "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.

[47] Manning, Christopher D. and Hinrich Schutze. 1999. Review of "Foundations of statistical natural language processing". The MIT Press 1999.. Comput. Linguist. 26, 2 (June 2000), 277-279.

[48] Mitchell, Jeff and Mirella Lapata, Composition in Distributional Models of Semantics. Journal of Cognitive Science, 2010.

[49] Roy, D. K. and Pentland, A. P. 2002, Learning words from sights and sounds: a computational model. Cognitive Science, 26: 113146. doi: 10.1207/s15516709cog2601_4

[50] Roy, Deb and Peter Gorniak. 2004. Grounded Semantic Composition for Visual Scenes. Journal of Artificial Intelligence Research, 21: 429-470.

[51] Roy, Deb and Ehud Reiter. 2005. Connecting language to the world. Artificial Intelligence, 167(1-2): 1-12

[52] Michael Fleischman and Deb Roy. 2007. Situated Models of Meaning for Sports Video Retrieval. HLT/ACL 2007, Rochester, NY.

[53] Miller, George and Walter Charles. 1991. Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1):128.

[54] Moore, David and George McCabe. 2005. Introduction to the Practice of Statistics. Freeman, New York, 5 edition.

[55] Nister, David and Henrik Stewenius. 2006. 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 21612168.

[56] Nixon S., Mark and Aguado S. Alberto. 2008. Feature Extraction and Image Processing. ISBN: 978-0-12372-538-7. Second Edition. Elsevier Academic Press.

[57] Piantadosi Steven T. and Noah D. Goodman, Benjamin A. Ellis, Joshua B. Tenenbaum. 2008. A Bayesian Model of the Acquisition of Compositional Semantics. Cognitive Science 2008.

[58] van de Weijer J ., T. Gevers, and A. Bagdanov. Boosting color iency in image feature detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1):150156, 2006.

[59] Yang Jun and Yu-Gang Jiang and Alexander G. Hauptmann and Chong-Wah Ngo. 2007. Evaluating bag-of- visual-words representations in scene classification. In James Ze Wang, Nozha Boujemaa, Alberto Del Bimbo, and Jia Li, editors, Multimedia Information Retrieval, pages 197206. ACM.

[60] Pado, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. Computational Linguistics, 33(2):161199.

[61] Rapp, Reinhard. 2003. Word sense discovery based on sense descriptor dissimilarity. In Proceedings of the 9th MT Summit, pages 315322, New Orleans, LA, USA.

[62] Roy, D. and Reiter, E. 2005. Connecting language to the world. Artificial Intelligence, 167(1-2): 1-12.

[63] Mavridis, Nick and Deb Roy. 2005. Grounded Situation Models for Robots: Bridging language, Perception, and Action. AAAI Workshop on Modular Construction of Human-Like Intelligence, pages 32-39.

[64] Rothenhausler, Klaus and Hinrich Schutze. 2009. Unsupervised classification with dependency based word spaces. In Proceedings of the EACL GEMS Workshop, pages 1724, Athens, Greece.

[65] Salton, G. 1971. The SMART Retrieval SystemExperiments in Automatic Document Processing. PrenticeHall, Inc., Upper Saddle River, NJ, USA.

[66] Sivic, Josef and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the International Conference on Computer Vision, volume 2, pages 14701477, October.

[67] Steyvers, Mark and Joshua B. Tenenbaum, The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. Journal of Cognitive Science, 2005.

[68] Smith, E. E., & Medin, D. L. (1981). Categories and concepts. Cambridge, MA, Harvard University Press.

[69] Siskind, J.M. and Morris, Q., A Maximum-Likelihood Approach to Visual Event Classification. 1996. Proceedings of the Fourth European Conference on Computer Vision (ECCV) , pp. 347-60, April 1996.

[70] Szeliski, Richard. Computer Vision: Algorithms and Applications. Springer, New York, 2010.

[71] Thomason, Rich. 1996. What is Semantics? Lectures at University of Michigan, Ann Arbor, MI.

[72] Turney, P. and Pantel, P. 2010. From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research (JAIR), 37(1):141-188. AI Access Foundation.

[73] Turney, P. D. (2006). Similarity of semantic relations. Computational Linguistics, 32 (3), 379-416.

[74] Turney, Peter. 2006b. Similarity of semantic relations. Computational Linguistics, 32(3):379416.

[75] Uijlings J.R.R. , A.W.M. Smeulders and R.J.H. Scha. 2010. Real-Time Visual Concept Classification. IEEE Transactions on Multimedia, 99.

[76] van de Sande, Koen E. A., Gevers T., Snoek C.G.M. 2008. A Comparison of Color Features for Visual Concept Classification CIVR08, July 79, 2008, Niagara Falls, Ontario, Canada.

[77] van de Sande, Koen E. A., Gevers T., Snoek C.G.M. 2010. Evaluating Color Descriptors for Object and Scene Recognition. In the Transactions of Pattern analysis and Machine Intelligence.

[78] von-Ahn, Luis and Laura Dabbish. 2004. Labeling images with a computer game. In Proceedings of the SIGCHI conference on Human factors in computing systems, CHI 04, pages 319326, New York, NY, USA. ACM.

[79] Wojtinnek PiaRamona and Brian Harrington and Sebastian Rudolph and Stephen Pulman. 2010. Conceptual Knowledge Acquisition Using Automatically Generated LargeScale Semantic Networks. In Proceedings of the 18th International Conference on Conceptual Structures, 2010.

[80] Wittgenstein, Ludwig. 1953. Philosophical Investigations. Blackwell, Oxford. Translated by G.E.M. Anscombe.

[81] Zwaan, Rolf. 2004. The immersed experiencer: Toward an embodied theory of language comprehension. Psychology of Learning and Motivation: Advances in Research and Theory, Vol 44, 44.

Giang Binh Tran

Combining text-based and vision-based semantics

We present an innovative (and first) framework for creating a multimodal distributional semantic model from state of the art text-and image-based semantic models. We evaluate this multimodal semantic model on simulating similarity judgements, concept clustering and the newly introduced BLESS benchmark. We also propose an effective algorithm, namely Parameter Estimation, to integrate text- and image-based features in order to have a robust multimodal system. By experiments, we show that our technique is very promising. Across all experiments, our best multimodal model claims the first position. By relatively comparing with other text-based models, we are justified to affirm that our model can stay in the top line with other state of the art models. We explore various types of visual features including SIFT and other color SIFT channels in order to have preliminary insights about how computer-vision techniques should be applied in the natural language processing domain. Importantly, in this thesis, we show evidences that adding visual features (as the perceptual information coming from images) is comparable (and possibly better) than adding further text features to the advanced text-based model; and more interestingly, the visual features can capture the semantic characteristics of (especially concrete) concepts and they are complementary with respect to the characteristics captured by textual features.

Partial results of this thesis are published as:

†Giang Binh Tran and Elia Bruni and Marco Baroni. 2011. Convergence of text- based and vision-based semantics, Social Media Retrieval Summer School, Poster section, Antalya - Turkey, June.

†Elia Bruni and Giang Binh Tran and Marco Baroni. 2011. Distributional semantics from text and images. EMNLP - GEMS Workshop Edinburgh - UK, July.

Printed and bound by Giang B.Tran