



PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

**CAN A GENERAL-PURPOSE COMMONSENSE
ONTOLOGY IMPROVE PERFORMANCE OF
LEARNING-BASED IMAGE RETRIEVAL?**

RODRIGO ANDRÉS TORO ICARTE

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Advisor:

JORGE BAIER ARANDA

ÁLVARO SOTO ARRIAZA

Santiago de Chile, December 2015

© MMXV, RODRIGO ANDRÉS TORO ICARTE



PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

**CAN A GENERAL-PURPOSE COMMONSENSE
ONTOLOGY IMPROVE PERFORMANCE OF
LEARNING-BASED IMAGE RETRIEVAL?**

RODRIGO ANDRÉS TORO ICARTE

Members of the Committee:

JORGE BAIER ARANDA

ÁLVARO SOTO ARRIAZA

DENIS PARRA SANTANDER

PABLO ESPINACE RONDA

JORGE GIRONÁS LEÓN

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Santiago de Chile, December 2015

© MMXV, RODRIGO ANDRÉS TORO ICARTE

To Jorge Andrés Icarte Romo

ACKNOWLEDGEMENTS

In January of 2013, my AI professors came to me with a simple, but powerful idea: *“Deductive and inductive techniques must eventually intersect”*. However, they were not sure about *“how to do it”* or *“where to do it”*. For the next 3 years, answering those questions became my job. After two and a half years of research, we developed hundreds of different ways to intersect inductive and deductive techniques which drove to awful results. I was very frustrated. Suddenly, one cold night in Toronto, not so long ago, all my hard work finally paid off when we made a small improvement with a specific experiment. Since then, everything has been easier. Nevertheless, none of this could be possible if it were not for the several people who supported this process. In this section I wish to acknowledge the help provided by them.

First, I would like to thank my family, especially to Nancy Icarte, Guillermo Toro, Javier Chandía, Matías Smith, Oliva Toro, Marta Icarte, Gustavo Aranda, Rebeca Jorquera and Jorge Icarte. I would not be the same person without your comprehension, care, friendship, life lessons, support and love. Thanks for always being with me and for cheering me up any time an experiment went wrong. You gave me the calmness that I needed to continue believing in this project. You are the best family anyone could have.

Second, I would like to offer my special thanks to the computer science department of Universidad Católica. This department provided me with a perfect environment to develop in these works. The amenities included a nice office, a coffee machine, a computer cluster with a GPU and highly qualified professors. I am particularly grateful to Dr. Álvaro Soto, Dr. Jorge Baier and Dr. Cristian Ruz. Álvaro opened the research door for me, Jorge taught me the value of simple ideas and Cristian helped me to turn an idea into code lines. I am very grateful for having the opportunity to work with you.

Finally, I would like to thank Margarita Castro, one of the smartest people that I know. During my thesis, we had plenty of valuable discussions, from which we generated several

ideas to continue working on. In fact the main algorithm presented in this thesis was born from one of those discussions. I could not do it without you.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF FIGURES	viii
LIST OF TABLES	ix
ABSTRACT	x
RESUMEN	xi
1. INTRODUCTION	1
2. PREVIOUS WORK	4
3. PRELIMINARIES	6
3.1. Stemming	6
3.2. ConceptNet	6
3.3. Sentence-Based Image Retrieval	7
3.4. Visual Datasets	8
3.4.1. MS COCO	8
3.4.2. ESPGAME	8
4. A BASELINE FOR IMAGE RETRIEVAL	10
5. OUR METHOD: ENHANCING DETECTORS SETS USING CN	12
5.1. CN for Undetectable Words	12
5.2. The CN Score	14
6. RESULTS AND DISCUSSION	16
6.1. Comparing Variants of CN	16
6.2. Comparison to Other Approaches	18
6.3. From COCO 5K to COCO 22K	21

7. CONCLUSIONS AND PERSPECTIVES 23

References 24

LIST OF FIGURES

1.1	Left. An image and one of its associated sentences from the MS COCO dataset. Among its words, the sentence features <i>Chef</i> , for which there is not a visual detector available. Right. Related nodes to <i>chef</i> in ConceptNet <i>s.t.</i> we have visual detectors for them. Notice that several of those nodes would be informative if we wanted to detect <i>chef</i>	3
3.1	A sample of CN relations that involve the concept <i>sofa</i> , together with the English description provided by the CN team in their website.	7
3.2	An image with its 5 associated sentences in MS COCO.	8
3.3	An image with its associated tag list in ESPGAME.	9
6.1	Qualitative examples for our baseline “MIL STEM” and our method “CN_ESP_MAX” over COCO 5K. Blue words are stemming-detectable, whereas red words are only CN-detectable.	20
6.2	Qualitative examples for <i>tuxedo</i> retrieval. First image row contains our ground truth, the 4 examples where <i>tuxedo</i> was used to describe the image. The second row of images are the first 4 retrieved images from CN_ESP_MAX.	21

LIST OF TABLES

6.1	Subset of COCO 5K with sentences that contains an undetectable word. . . .	17
6.2	Image retrieval results over COCO 5K. Note our baselines outperform state-of-the-art results and our approach, CN_ESP_MAX, improved our baselines performance.	18
6.3	Image Retrieval for new word detectors over COCO 5K. We include a random baseline, and results for CN_ESP_MAX divided in 4 categories: Retrieving nouns, verbs, adjectives and all of them. The results show that is easier, for CN_ESP_MAX, to detect nouns than verbs or adjectives.	21
6.4	Image retrieval results for COCO 5K and 20K. In this table we compare our best baseline against a version of CN_ESP_MAX which only detect new noun words. Performance improvement increases when more images are considered. . . .	22

ABSTRACT

The knowledge representation community has invested great efforts in building general-purpose ontologies which contain large amounts of commonsense knowledge on various aspects of the world. Among the thousands of assertions contained in them, many express relations that can be regarded as relevant to visual inference; e.g., “a ball is *used by* a football player”, “a tennis player is *located at* a tennis court”. In general, current state-of-the-art approaches for visual recognition do not exploit these rule-based knowledge sources. Instead, they use learning techniques to learn recognition models directly from training examples (plenty of them). In this thesis, we study the question of whether or not general-purpose ontologies—specifically, MIT’s ConceptNet ontology—could play a role in state-of-the-art vision systems, as it seems plausible, in principle, that general-knowledge may be complementary to knowledge acquired from examples. As a testbed, we tackle the problem of sentence based image retrieval. As a starting point we make use of recent successful efforts on convolutional network models to develop a retrieval approach based only on a large pool of object detectors. Afterwards, we show how we can enhance this system using relevant relations from ConceptNet ontology leading to new state-of-the-art results on a common benchmark dataset.

Keywords: artificial intelligence, commonsense ontologies, machine learning, computer vision, visual recognition, image retrieval, ConceptNet, ESPGAME.

RESUMEN

La comunidad de representación del conocimiento ha invertido grandes esfuerzos en la creación de ontologías de sentido común. Ellas poseen miles de relaciones sobre distintos aspectos del mundo cotidiano, por ejemplo “todo hombre es persona” o “los libros son usados para leer”. Dentro de esta gran cantidad de relaciones, algunas de ellas contienen información relevante sobre el mundo visual. Sin embargo, hasta la fecha, ningún algoritmo (que sea el estado del arte en alguna tarea de visión por computador) ha incorporado este conocimiento en forma explícita. Dichos algoritmos suelen utilizar técnicas de aprendizaje de máquina para aprender modelos de reconocimiento a partir de ejemplos (miles de ellos). En esta tesis estudiamos si una ontología de propósito general, específicamente ConceptNet (la ontología del MIT), puede, o no, tener un rol en el estado del arte de visión por computador. Elegimos *sentence based image retrieval* (búsqueda de imágenes mediante oraciones) como escenario de pruebas. Nuestro punto de partida es una red convolucional profunda que nos permite generar un algoritmo de *image retrieval* basado en detectores de palabras. Luego de eso presentamos una variante que incorpora relaciones de sentido común provenientes de ConceptNet. Como resultado, obtuvimos una mejora al estado del arte para la base de datos MSCOCO 5K.

Palabras Claves: inteligencia artificial, ontologías de sentido común, aprendizaje de máquina, visión por computador, reconocimiento visual, búsqueda de imágenes, ConceptNet, ESPGAME.

1. INTRODUCTION

In the last few decades, feature-based techniques have become dominant at tackling visual recognition problems (Grauman & Leibe, 2010). While earlier approaches focused on the creation of hand-crafted visual features (Lowe, 1999; Belongie, Malik, & Puzicha, 2000; Dalal & Triggs, 2005), recent approaches focus on applying learning techniques to obtain suitable mid-level representations (Bourdev, Maji, Brox, & Malik, 2010; Yang, Yu, Gong, & Huang, 2009; S. Singh, Gupta, & Efros, 2012; Lobel, Vidal, & Soto, 2013) or deep hierarchical layers of composable features (Boureau, Cun, et al., 2008; Krizhevsky, Sutskever, & Hinton, 2012). A common denominator of those methods is that they mainly rely on visual appearance. Their goal is to uncover visual spaces where visual similarities carry enough information to achieve robust visual recognition.

Feature-based approaches do not incorporate high-level semantic knowledge. Indeed, they consist mainly of data-driven procedures that, besides a limited use of class labels, do not exploit further semantic information. Contextual-based approaches (Choi, Lim, Torralba, & Willsky, 2010; Rabinovich, Vedaldi, Galleguillos, Wiewiora, & Belongie, 2007; Galleguillos & Belongie, 2010) and approaches based on visual semantic attributes (Farhadi, Endres, Hoiem, & Forsyth, 2009; J. Liu, Kuipers, & Savarese, 2011; Parikh & Grauman, 2011) aim at closing the semantic gap, incorporating stronger sources of high-level knowledge. This knowledge is typically obtained through manual labeling (Wolf & Bileschi, 2006), direct mining of nonaccidental visual relations (Desai, Ramanan, & Fowlkes, 2011; Peralta, Espinace, & Soto, 2012), or statistical analysis of text corpora (Rabinovich et al., 2007; Espinace, Kollar, Roy, & Soto, 2013). Recently, new massive sources of structured visual data, such as ImageNet (Deng et al., 2009), have been used to augment appearance-based visual similarities using semantic relations (Deselaers & Ferrari, 2011; C. Fang & Torresani, 2012). However, to the best of our knowledge, to date no approach has attempted to exploit publicly available, large commonsense knowledge repositories for visual recognition.

The knowledge representation community has recognized for years that large commonsense knowledge bases are needed to reason in the real world. Consequently, a number of projects have been conducted to build such a knowledge base or ontology. Cyc (Lenat, 1995) and ConceptNet (Havasi, Speer, & Alonso, 2007) are two well-known examples of large, publicly available commonsense knowledge bases. In particular, in this work we focus on ConceptNet (CN), a large hyper-graph containing information about million of world concepts and their relations, such as, “books can be made from paper“. CN has been used successfully for tasks that require rather complex commonsense reasoning. This includes a recent study that showed that the information in CN may be used to score as good as a four-year old in an IQ test (Ohlsson, Sloan, Turán, & Urasky, 2013).

In this work, we study the question of whether or not large publicly available commonsense knowledge repositories could play a role in state-of-the-art vision systems. We understand commonsense knowledge as a collection of facts about the world that most people possess, such as, “a piano is used for performing music“. This also includes commonsense relations such as “used for”, “capable of”, or “caused by”, which are not considered by current visual recognition systems. In principle, these knowledge sources can provide strong constraints and a richer representation to achieve a deeper understanding of the visual world. Furthermore, they can be complementary to current state-of-the-art visual recognition approaches mainly based on mining large sources of training examples. Our own visual perceptual system illustrates the power of commonsense reasoning. As an example, we use knowledge about friendships or familial relationships to be able to recognize people in low resolution images.

As a testbed, we tackle the problem of unconstrained sentence based image retrieval. This is a challenging and relevant task that requires a suitable mapping between natural language and the visual world. Recent works that fuse text and images have highlighted the need to provide this task with suitable sources of high level knowledge (Karpathy & Li, 2015; Vinyals, Toshev, Bengio, & Erhan, 2015). Our proposed approach is based on using CN to augment or complement the information that we can extract from a set of object

detectors operating on the input image. As an example, Figure 1 shows an illustrative situation, where an image retrieval query contains the word *Chef*, for which there is not a visual detector available. However, the information contained in the nodes directly connected to the concept *Chef* in CN provides key information to active related visual detectors, such as *Person*, *Dish* and *Kitchen* that results key to retrieve the intended image.

Briefly, our proposed method consists of using the information contained in the words of the query sentence, and a set of related concepts and relations extracted from CN, to selectively activate a bank of visual object detectors. These detectors correspond to a large set previously trained using a recent successful model based on convolutional neural networks. The level of activations of the selected detectors in each of the images in the dataset allow us then to rank the images according to their relevance to the textual query. As an additional step, we mine complementary sources of knowledge, such as the ESPGAME dataset, to filter out noisy or irrelevant relations provided by CN. We evaluate our approach over standard 5K- and 22K- COCO image datasets showing that it improves the performance of a state-of-the-art approach to image retrieval.

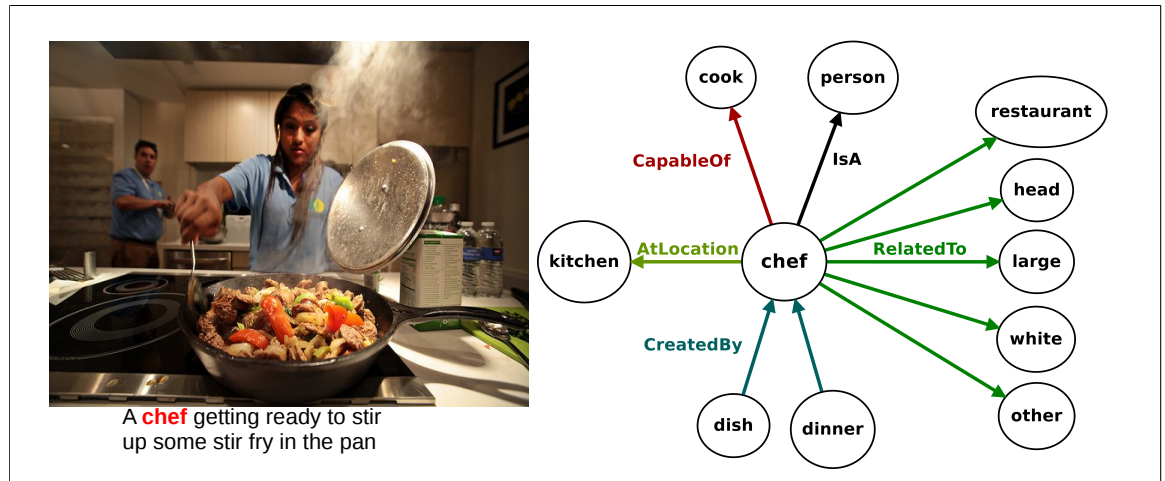


FIGURE 1.1. **Left.** An image and one of its associated sentences from the MS COCO dataset. Among its words, the sentence features *Chef*, for which there is not a visual detector available. **Right.** Related nodes to *chef* in ConceptNet *s.t.* we have visual detectors for them. Notice that several of those nodes would be informative if we wanted to detect *chef*.

2. PREVIOUS WORK

The relevance of contextual or semantic information to visual recognition has long been acknowledged and studied by the cognitive psychology and computer vision communities (Peissig & Tarr, 2007; Biederman, 1972). In computer vision, the main focus has been on using contextual relations to boost recognition by learning models about the organization of the visual world. In particular, following Biederman (1972), contextual information in the form of object co-occurrences, and geometrical and spatial constraints, has been successfully applied to improve object and action recognition performance (Choi et al., 2010; Desai et al., 2011; Wang, Chen, & Wu, 2011). Adaptive forms on contextual cues has also been proposed (Peralta et al., 2012). Due to space constraints, we derive the reader to Galleguillos and Belongie (2010) and Marques et al. (2011) for in-depth reviews about these topics. As a common issue of these methods, they do not employ high-level semantic relations.

Several works have pointed out the relevance of enhancing appearance-based similarity metrics with semantic information. Deselaers and Ferrari (2011) and C. Fang and Torresani (2012) used the structure of ImageNet and nearest neighbors learning techniques to create distance metrics based on appearance and semantic object-category similarities. Following Bush (1979), Malisiewicz and Efros (2009) posed visual recognition as a search operation in a large relational graph, the so-called Visual Memex. Malisiewicz and Efros (2009) focused on relations based on visual similarity and object co-occurrence. However, they do not use relations between object-context, object-action or object-attribute, which are available in ConceptNet.

Knowledge adquisition is one of the main challenges of using a semantic based approach to object recognition. One common approach to obtain this knowledge is via text mining (Rabinovich et al., 2007; Espinace et al., 2013) or crowd sourcing (Von Ahn & Dabbish, 2004a; Deng et al., 2009). As an alternative, recently, Chen et al. (2013) and Divvala et al. (2014) presented bootstrapped approaches where an initial set of object detectors and relations is used to mine the web in order to discover new object instances and

new common sense relationships. The new knowledge is in turn used to improve the search for new classifiers and semantic knowledge in a never ending process. While this strategy opens new opportunities, unfortunately, as it has been pointed out for Von Ahn and Dabish (2004a), public information is biased. In particular, common sense knowledge is so obvious that it is generally tacit and not explicitly included in most information sources. Furthermore, unsupervised or semi-supervised semantic knowledge extraction techniques often suffer from semantic drift problems, where slightly misleading local association are propagated to lead to wrong semantic inference.

Recently, work on automatic image captioning has made great advances to integrate image and text data (Karpathy & Li, 2015; Vinyals et al., 2015; H. Fang et al., 2015; Donahue et al., 2014). These approaches use datasets consisting of images as well as sentences describing their content, such as the Microsoft COCO dataset (Lin et al., 2014). Coincidentally, all these methods share similar ideas which follow initial work in (Weston, Bengio, & Usunier, 2011). Briefly, these works employ deep neural network models, mainly convolutional and recurrent neural networks, to infer a suitable alignment between sentence snippets and the corresponding image region that they describe. In contrast to our proposed approach, these methods do not make explicit use of high level semantic knowledge.

In terms of works that use ontologies to perform visual recognition, Maillot and Thonnat (2008) built a custom ontology to perform visual object recognition. Ordonez et al. (2015) used Wordnet and a large set of visual object detectors to automatically predict natural nouns that people will use to name visual object categories. Zhu et al. (2014) used Markov Logic Networks and a custom ontology harvested from different sources, to identify in images several properties related to object affordance. In contrast to our work, these methods target different applications. Furthermore, they do not exploit the type of common sense relations that we want to extract from ConceptNet.

3. PRELIMINARIES

Below we present an overview of the elements we use in this thesis, including the ontology and visual datasets we use, and background on the image retrieval task.

3.1. Stemming

A standard technique in Natural Language Processing that we use below is *stemming*. The stemming of a word w is a legal (*e.g.*, English) word that results from stripping a suffix out of w . Stemming is a heuristic process that aims at returning a word that is closest to the “root” of a word. For example, the stemming of the words *run*, *runs*, and *running* all return the word *run*.

For a word w , we denote its stemmed version as $st(w)$. If W is a set of words, then $st(W) = \{st(w) \mid w \in W\}$.

3.2. ConceptNet

ConceptNet (CN) (H. Liu & Singh, 2004) is a commonsense-knowledge semantic network which represents knowledge in a hypergraph structure. Each node in the hypergraph corresponds to a concept represented by a *stemmed* word or phrase. Hyperarcs, on the other hand, represent relations between nodes, and are associated with a weight, that expresses the confidence in such a relation. As stated in its webpage, CN is a knowledgebase “containing lots of things computers should know about the world, especially when understanding text written by people”.

CN relations are commonsensical in the sense that they represent knowledge that is standard for most humans (see Figure 3.1 for a sample). One of the design principles behind CN is that knowledge is automatically generated from sentences coming from the *Open Mind Commonsense Corpus* (P. Singh et al., 2002; H. Liu & Singh, 2004), which are natural-language sentences created by humans. Online games (*e.g.*, *Verbosity* von Ahn, Kedia, & Blum, 2006) have also been used to massively collect relations for CN.

ConceptNet relation	ConceptNet’s description
sofa <i>–IsA→</i> piece of furniture	<i>A sofa is a piece of furniture</i>
sofa <i>–AtLocation→</i> livingroom	<i>Somewhere sofas can be is livingroom</i>
sofa <i>–UsedFor→</i> read book	<i>A sofa is for reading a book</i>
sofa <i>–MadeOf→</i> leather	<i>sofas are made from leather</i>

FIGURE 3.1. A sample of CN relations that involve the concept *sofa*, together with the English description provided by the CN team in their website.

Among the set of relation types in CN, a number of them can be regarded as “visual”, in the sense that they correspond to relations that are important in the visual world. These include relations for spatial co-occurrence (e.g., *LocatedNear*, *AtLocation*), visual properties of objects (e.g., *PartOf*, *SimilarSize*, *HasProperty*, *MadeOf*), and actions (e.g. *UsedFor*, *CapableOf*, *HasSubevent*).

CN 5.3, the version we use in our experiments, contains over four million nodes and over 13 million relations. Many of these relations are truthful. Indeed Singh *et al.* (P. Singh et al., 2002) reported that human evaluators have rated 75% of items as “largely true”, 82% as “largely objective”, and 85% as “largely making sense”.

CN has been used successfully for tasks that require rather complex commonsense reasoning. A recent study showed that the information in CN may be used to score as good as a four-year-old person in an IQ test (Ohlsson et al., 2013).

Unfortunately however, CN contains a number of so-called *noisy relations*, which are relations that do not correspond to a true statement about the world. Two examples of these for the concept *pen* are “pen *–AtLocation→* pen”, “pig *–AtLocation→* pen”. The existence of these relations is an obvious hurdle when utilizing this ontology. We discuss later some of the implications of this for visual reasoning.

3.3. Sentence-Based Image Retrieval

Given a set of images \mathcal{I} and a natural-language query string t , the objective of image retrieval is to rank the images in \mathcal{I} according to their “relevance” with respect to t . As such,

the problem reduces to building a function that returns a numeric score to each image in $I \in \mathcal{I}$ given t that correlates with the relevance of I with respect to t .

Two recent state-of-the-art approaches to image retrieval are those by Klein *et al.* (Klein, Lev, Sadeh, & Wolf, 2015), based on Mixture Models, and that by Karpathy and Li (Karpathy & Li, 2015), based on a bidirectional recurrent neural network.

3.4. Visual Datasets

In the next chapters we refer to two visual datasets: MS COCO and ESPGAME.

3.4.1. MS COCO

Microsoft COCO (Lin et al., 2014) is a new captioning dataset. It is divided in 3 sets with over 80K, 40K and 40K images for training, validating, and testing, respectively. Each image in the training and validating set is associated with, at least, 5 natural language descriptions. Figure 3.2 shows an example image with its descriptions.

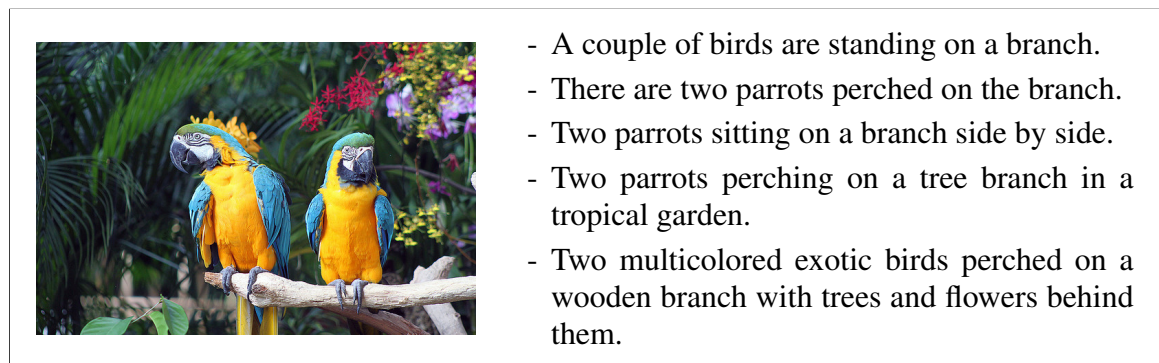
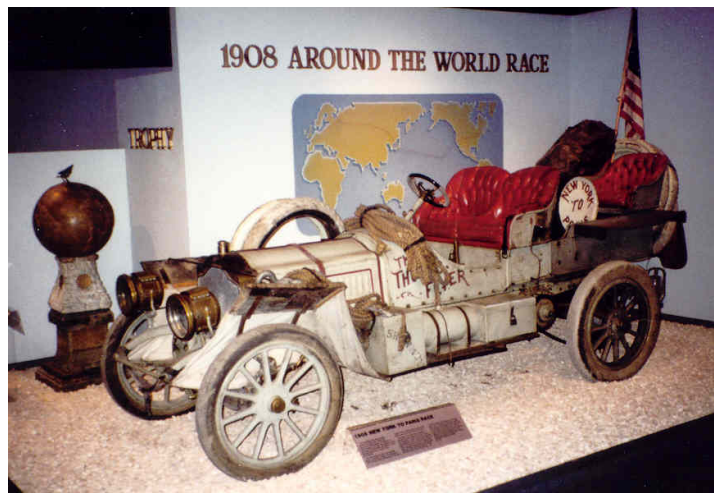


FIGURE 3.2. An image with its 5 associated sentences in MS COCO.

3.4.2. ESPGAME

ESPGAME (Von Ahn & Dabbish, 2004b) is a well known tagging database. It contains 100001 images with meaningful tags (English word) for each of them. Figure 3.3 shows an example image with its tag list.



Tags: old, 1908, wheels, red, car, auto, race, museum, wheel, flag, around, automobile, the, world.

FIGURE 3.3. An image with its associated tag list in ESPGAME.

4. A BASELINE FOR IMAGE RETRIEVAL

To evaluate our technique for image retrieval, we chose as a baseline a simple approach based on a large set of visual word detectors (H. Fang et al., 2015). These detectors, that we refer to as Fang *et al.*’s detectors, were trained over the MS COCO image dataset (Lin et al., 2014). Each image in this dataset is associated with 5 natural-language descriptions (for more details, see section 3.4.1). Fang *et al.*’s detectors were trained to detect instances of words appearing in the sentences associated to MS COCO images. As a result, they obtain a set of visual word detectors for a vocabulary V , which contains the 1000 most common words used to describe images on the training dataset.

Given an image I and a word w , Fang *et al.*’s detector outputs a score between 0 and 1. With respect to training data, such a score can be seen as an estimate of the probability that image I has been described with word w . Henceforth, we denote such a score by $\hat{P}_V(w \mid I)$.

A straightforward, but effective way of applying these detectors to image retrieval is by simply multiplying the scores. Specifically, given a text query t and an image I we run the detectors on I for words in t that are also in V and multiply their output scores. We denote this score by MIL (after Multiple Instance Learning, the technique used in H. Fang et al., 2015 to train the detectors). Mathematically,

$$MIL(t, I) = \prod_{w \in V \cap S_t} \hat{P}_V(w|I), \quad (4.1)$$

where S_t is the set of words in the text query t .

The main assumption behind Equation 4.1 corresponds to an independence assumption among word detectors given an image I . This is similar to the Naive Bayes assumption used by the classifier with the same name (Mitchell, 1997). In chapter 6, we show that, although simple, this score outperforms state-of-the-art approaches (*e.g.*, Klein et al., 2015; Karpathy & Li, 2015). While these detectors were trained using MS COCO dataset, we still

found—surprisingly—that they lead to state-of-the-art performance when they are used in a sentence-based image retrieval task using this dataset.

5. OUR METHOD: ENHANCING DETECTORS SETS USING CN

The MIL score has two limitations. First, it considers the text query as a set of independent words, ignoring their semantics relations and roles in the sentence. Second, it is limited to the set V of words the detector has been trained for. While the former limitation may also present in other state-of-the-art approaches to image retrieval, the latter is inherent to any approach that employs a set of visual word detectors for image retrieval.

Henceforth, given a set of words V for which we have a detector, we say that word w is *undetectable* with respect to V iff w is not in V , and we say w is *detectable* otherwise.

In the rest of the chapter we describe a technique to enhance a detector-based approach for image retrieval with undetectable words.

5.1. CN for Undetectable Words

Our goal is to provide a score to each image analogous to that defined in Equation 4.1, but including undetectable words. A first step is to define a score for an individual undetectable word w . Intuitively, if w is an undetectable word, we want an estimate analogous to $\hat{P}_V(w|I)$. Formally, the problem we address can be stated as follows: given an image I and a word w which is undetectable wrt. V , compute the estimate $\hat{P}(w|I)$ of the probability $P(w|I)$ of w appearing in I .

To achieve this, we are inspired by the following fact: for most words representing a concept c , CN “knows” a number of other concepts that are related to c that may share related visual characteristics. As an illustration, if w is, *e.g.* *tuxedo*, then we may consider the concept *jacket* as one that could provide useful visual information because “tuxedo—*IsA*→ *jacket*” is a relation in CN. Likewise, other visual CN relations such as *AtLocation* or *MadeOf* may also provide useful information.

Let us define $cn(w)$ as the set of concepts (each represented by an English word) that appear related to the stemmed version of w , $st(w)$, in CN. We propose to compute $\hat{P}(w|I)$ based on the estimation $\hat{P}(w'|I)$ of words w' that appear in $cn(w)$. Specifically, by using

standard probability theory we can write the following identity about the *actual* probability function P and every $w' \in cn(w)$:

$$P(w|I) = P(w|w', I)P(w'|I) + P(w|\neg w', I)P(\neg w'|I), \quad (5.1)$$

where $P(w|w', I)$ is the probability that there is an object in I associated to w given that there is an object associated to w' in I . Likewise $P(w|\neg w', I)$ represents the probability that there is an object for word w in I , given that no object associated to word w' appears in I .

Equation 5.1 can be re-stated in terms of estimations, thus for *any* $w' \in cn(w)$, $P(w|I)$ can be estimated by $\hat{p}(w', w, I)$, which is defined by

$$\hat{p}(w', w, I) \stackrel{\text{def}}{=} \hat{P}(w|w', I)\hat{P}(w'|I) + \hat{P}(w|\neg w', I)\hat{P}(\neg w'|I). \quad (5.2)$$

However the problem with such an approach is that it does not tell us which w' to use. Below, we propose to aggregate \hat{p} over the set of all concepts w' that are related to w in CN. Before stating such an aggregation formally, we focus on how to compute $\hat{P}(w|w', I)$ and $\hat{P}(w'|I)$.

Let us define the set $stDet(w, V)$ as the set of words in the set V such that when stemmed are equal to $st(w)$; i.e., $stDet(w, V) = \{w' \in V : st(w) = st(w')\}$. Intuitively, $stDet(w, V)$ contains all the words in V whose detectors can be used to detect w after applying stemming.

Now we define $\hat{P}(w'|I)$ as:

$$\hat{P}(w'|I) = \max_{w \in stDet(w', V)} \hat{P}_V(w|I), \quad (5.3)$$

that is, to estimate how likely it is that w' is in I , we look for a word w in the set of detectors V whose stemmed version matches the stemmed version of w' , and that maximizes $\hat{P}_V(w|I)$.

Now we need to define how to compute $\hat{P}(w|w', I)$. We tried two options here (we report both in Section 6). The first is to assume $\hat{P}(w|w', I) = 1$, for every w, w' . This is

because for some relation types it is correct to assume that $\hat{P}(w|w', I)$ is equal to 1. For example $\hat{P}(\text{person}|\text{man}, I)$ is 1 because there is a CN relation “man-*IsA*→ person”. While it is clear that we should use 1 for the *IsA* relation, it is not clear whether or not this estimate is correct for other relation types. Furthermore, since CN contains noisy relations, using 1 might yield significant errors.

Our second option, which yielded better results, is to make $\hat{P}(w|w', I)$ equal to an estimate of $P(w|w')$; i.e., the probability that an image that contains an object for w' contains an object for word w . $P(w|w')$ can be estimated from the ESPGAME database, which contains tags for many images (for more details, see section 3.4.2). After stemming each word, we simply count the number of images tagged in which both w and w' occur and divide it by the number of images tagged by w' .

Now we are ready to propose a CN-based estimate for $P(w|I)$ when w is undetectable. As discussed above, $P(w|I)$ could be estimated by the expression of Equation 5.2 for any concept $w' \in cn(w)$. As it is unclear which w' to choose, we propose to aggregate over $w' \in cn(w)$ using three aggregation functions. Consequently, we identify three estimates of $P(w|I)$ that are defined by:

$$\hat{P}_{\mathcal{F}}(w|I) = \mathcal{F}_{w' \in cnDet(w, V)} \hat{p}(w', w, I), \quad (5.4)$$

where $cnDet(w, V) = \{w' \in cn(w) : stDet(w', V) \neq \emptyset\}$ is the set of concepts related to w in CN for which there is a stemming detector in V , and $\mathcal{F} \in \{\min, \max, \text{mean}\}$.

5.2. The CN Score

With a definition in hand for how to estimate the score of an individual undetectable word w , we are ready to define a CN-based score for a complete natural-language query t . For any w in t , what we intuitively want is to use the set of detectors whenever w is detectable and $\hat{P}_{\mathcal{F}}(w|I)$ otherwise.

To define our score formally, a first step is to extend the MIL score with stemming. Intuitively, we want to resort to detectors in V as much as possible, therefore, we will attempt

to stem a word and use a detector before falling back to our CN-based score. Formally,

$$\text{MILSTEM}(t, I) = \text{MIL}(t, I) \times \prod_{w \in W'_t} \hat{P}(w|I), \quad (5.5)$$

where W'_t is the set of words in t that are undetectable wrt. V but that are such that they have a detector via stemming (*i.e.*, such that $\text{stDet}(w, V) \neq \emptyset$), and where $\hat{P}(w|I)$ is defined by Equation 5.3.

Now we define our CN score which depends on the aggregation function \mathcal{F} . Intuitively, we want to use our CN score with those words that remain to be detected after using the detectors directly and using stemming to find more detectors. Formally, let W''_t be the set of words in the query text t such that (1) they are undetectable with respect to V , (2) they have no stemming-based detector (*i.e.*, $\text{stDet}(w, V) = \emptyset$), but (3) they have at least one related concept in CN for which there is a detector (*i.e.*, $\text{cnDet}(w, V) \neq \emptyset$). Then we define:

$$\text{CN}_{\mathcal{F}}(t, I) = \text{MILSTEM}(t, I) \times \prod_{w \in W''_t} \hat{P}_{\mathcal{F}}(w|I), \quad (5.6)$$

for $\mathcal{F} \in \{\min, \max, \text{mean}\}$.

6. RESULTS AND DISCUSSION

We evaluated our algorithm over the MS COCO image database (section 3.4.1). Following Karpathy and Li (2015) and Klein et al. (2015) we used a specific subset of 5K images (from the validation set) and evaluated the methods on the union of the sentences for each image. Henceforth, we refer to this subset as COCO 5K.

In our tables we report the mean and median rank of the *ground truth* image; that is, the one that was tagged by the query text being used in the retrieval task. We report also the k -recall ($r@k$), for $k \in \{1, 5, 10\}$, which corresponds to the percentage of times the correct image was found among the top k results.

Recall that we say that word w is detectable when there is a detector for word w ; in this chapter we use Fang *et al.*'s detectors (H. Fang et al., 2015) which is comprised by 616 detectors for nouns, 176 detectors for verb and 119 detectors for adjectives. We do not use detectors for other word types (such as prepositions or pronouns) because they decrease the performance. In addition, we say that a word w is stemming-detectable if it is among the words considered by the MILSTEM score, and we say a word is CN-detectable if it is among the words included in the CN-score.

6.1. Comparing Variants of CN

The objective of our first experiment was to compare the performance of the various versions of our approach that use different aggregation functions. Since our approach uses data from ESPGAME we also compare to an analogous approach that uses only ESPGAME data, without knowledge from CN. This is obtained by interpreting that word w is related to a word w' if both occur on the same ESPGAME tag, and using the same expressions presented in chapter 5. We considered a comparison to this method was important because we wanted to evaluate the impact of using an ontology with general-purpose knowledge versus using a crowded-sourced, mainly visual knowledge such as that in ESPGAME.

TABLE 6.1. Subset of COCO 5K with sentences that contains an undetectable word.

Database \subset COCO 5K	r@1	r@5	r@10	median rank	mean rank
Our baselines					
MIL	13.2	33.4	45.2	13	82.2
MIL_STEM	13.5	33.8	45.7	13	74.6
CN only					
CN_MIN	12.2	31.4	43.4	15	77.0
CN_MEAN	13.2	33.7	46.0	13	66.3
CN_MAX	12.2	32.1	44.1	14	73.0
ESPGAME only					
ESP_MIN	12.6	30.7	41.1	17	122.4
ESP_MEAN	13.6	34.2	46.2	13	69.0
ESP_MAX	13.5	33.7	45.7	13	66.2
CN + ESPGAME					
CN_ESP_MIN	14.3	34.6	46.6	12	68.3
CN_ESP_MEAN	14.6	35.6	48.0	12	61.2
CN_ESP_MAX	14.3	35.9	48.2	12	60.6

Table 6.1 shows results over the maximal subset of COCO 5K such that a query sentence has a CN-detectable word that is not stemming-detectable, including 9,903 queries. The table shows results for our baselines, CN_OP, ESP_OP and CN_ESP_OP (with OP = MIN (minimum), MEAN (arithmetic mean) and MAX (maximum)). CN_OP considers $\hat{P}(w|w', I) = 1$ and $\hat{P}(w|\neg w', I) = 0$ for Equation 5.2, while CN_ESP_OP uses ESPGAME to estimate those probabilities. To reduce the impact of noisy relations in ConceptNet in CN_OP and CN_ESP_OP, we only consider relationships with confidence weight ≥ 1 (this threshold was defined by carrying out a sensitivity analysis). The last method, ESP_OP uses ESPGAME without ConceptNet. Results show that algorithms based on CN + ESPGAME perform better in all the reported metrics, including the median rank. Overall, the MAX version of CN_ESP seems to obtain the best results and thus we select it as the method that we compare to other approaches below.

TABLE 6.2. Image retrieval results over COCO 5K. Note our baselines outperform state-of-the-art results and our approach, CN_ESP_MAX, improved our baselines performance.

Database	r@1	r@5	r@10	median	mean
COCO 5K				rank	rank
Other approaches					
NeuralTalk (Karpathy & Li, 2015; Vinyals et al., 2015)	6.9	22.1	33.6	22	72.2
GMM+HGLMM (Klein et al., 2015)	10.8	28.3	40.1	17	49.3
BRNN (Karpathy & Li, 2015)	10.7	29.6	42.2	14	NA
Our baselines					
MIL	15.7	37.8	50.5	10	53.6
MIL_STEM	15.9	38.3	51.0	10	49.9
Our method					
CN_ESP_MAX	16.2	39.1	51.9	10	44.4

6.2. Comparison to Other Approaches

We compared with NeuralTalk¹ (Karpathy & Li, 2015; Vinyals et al., 2015), BRNN(Karpathy & Li, 2015), and GMM+HGLMM (the best algorithm in Klein et al., 2015) over COCO 5K. As we can see on the Table 6.2, MIL outperforms previous approaches to image retrieval. Moreover, adding ConceptNet to detect new words improves the performance in almost all metrics.

Figure 6.1 shows qualitative results for 6 example queries. The first column describes the target image and its caption. The next columns show the rank of the correct image and the top-4 ranked images for MIL_STEM and CN_ESP_MAX. Blue words on the query are stem-detectable but not detectable, and red words are CN-detectable but not stem-detectable.

Queries 1 and 2, show examples of images for which no detectors can be used and thus the only piece of available information comes from CN. Rank for MIL_STEM is, therefore, arbitrary but, as a result of using CN with ESPGAME, the correct image are under r@25 in CN_ESP_MAX. Queries 3 and 4 show examples where we have both stem-detectable words

¹We used the source code from <https://github.com/karpathy/neuraltalk>.

and CN-detectable words (that are not stem-detectable). In both cases, using CN improves the position of the target image by more than 100. For Query 3, CN_ESP_MAX is able to detect “*bagel*” using the “*doughnut*” and “*bread*” detectors (among others), improving the ranking of the correct image. Similarly, with Query 4, our approach detects “*sculpture*” using detectors for “*statue*”, “*metal*”, and “*stone*”, among others. The last two queries are cases for which the CN-score is detrimental. For Query 5, the word “*resort*” is highly related with “*hotel*” in both CN and ESPGAME, thus the “*hotel*” detector became more relevant than “*looking*” and “*hills*”. For Query 6, it turns out that the word “*soul*” is related in CN to words that would seem helpful, like “*person*”, “*man*”, and “*body*” however it is also related to detectable words like “*church*”, “*live*” and “*many*”, which do not seem informative for this particular image.

Finally, we wanted to evaluate how good is the performance when focusing only on those words that are CN-detectable but not stemming-detectable. To that end, we designed the following experiment: we took the set of words from the union of text captions that were only CN-detectable, and we interpreted those as one-word queries. An image is ground truth in this case if any of its captions contains w .

Results presented in Table 6.3 disaggregate for word type (Nouns, Verbs, Adjectives). As a reference of the problem difficulty, we added a random baseline. The results suggest that CN yields more benefits for nouns, which may be easier to detect than verbs and adjectives by CN_ESP_MAX.

We observe that numbers are lower than in Table 6.1. In part this is due to the fact that in this experiment there is more than one correct image, therefore the k recall has higher chances of being lower than when there is only one correct image. Furthermore, a qualitative look at the data suggests that sometimes top-10 images are “good” even though ground truth images were not ranked well. Figure 6.2 shows an example of this phenomenon for the word *tuxedo*, which is not stemming-detectable.




























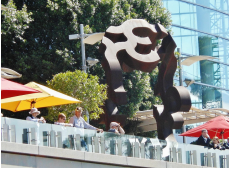




















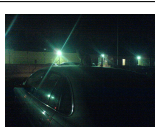

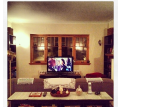



Sentence and target image	Algorithm	Pos 1	Pos 2	Pos 3	Pos 4
 1) Seamen inside a navy vessel communicate over the radio .	MIL STEM Pos: 5000				
	CN_ESP_MAX Pos: 15				
 2) The preparation of salmon , asparagus and lemons .	MIL STEM Pos: 5000				
	CN_ESP_MAX Pos: 23				
 3) Those bagels are plain with nothing on them.	MIL STEM Pos: 360				
	CN_ESP_MAX Pos: 2				
 4) People stand near a large modern art sculpture .	MIL STEM Pos: 149				
	CN_ESP_MAX Pos: 1				
 5) a spooky looking hotel resort in the hills .	MIL STEM Pos: 349				
	CN_ESP_MAX Pos: 597				
 6) Here is a soul in the image alone.	MIL STEM Pos: 1831				
	CN_ESP_MAX Pos: 2602				

FIGURE 6.1. Qualitative examples for our baseline “MIL STEM” and our method “CN_ESP_MAX” over COCO 5K. Blue words are stemming-detectable, whereas red words are only CN-detectable.

TABLE 6.3. Image Retrieval for new word detectors over COCO 5K. We include a random baseline, and results for CN_ESP_MAX divided in 4 categories: Retrieving nouns, verbs, adjectives and all of them. The results show that is easier, for CN_ESP_MAX, to detect nouns than verbs or adjectives.

Algorithm	r@1	r@5	r@10	median rank	mean rank
CN_ESP_MAX					
Random	0.02	0.1	0.2	2500.5	2500.50
All	0.4	1.8	3.3	962.0	1536.8
Noun	0.5	2.1	3.7	755.0	1402.7
Verb	0.2	1.1	1.9	1559.5	1896.2
Adjective	0.1	0.7	1.9	1735.5	1985.2

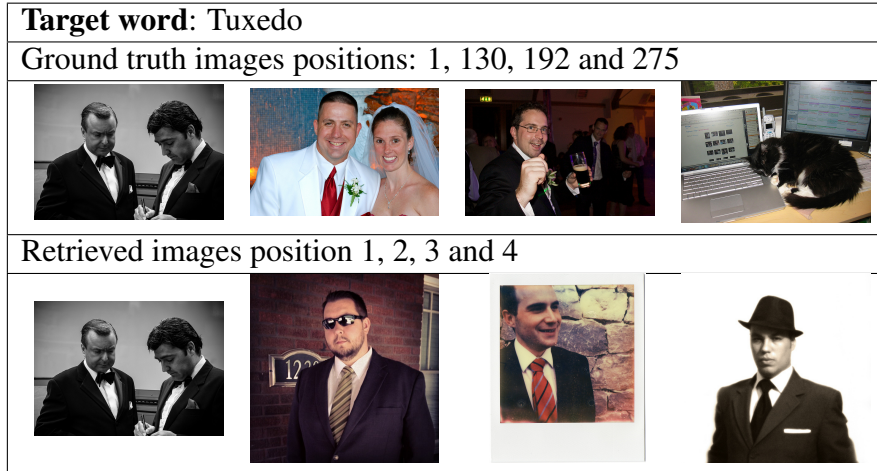


FIGURE 6.2. Qualitative examples for *tuxedo* retrieval. First image row contains our ground truth, the 4 examples where *tuxedo* was used to describe the image. The second row of images are the first 4 retrieved images from CN_ESP_MAX.

These results prompted us to experiment with a version of our CN score that tries to detect only undetectable nouns (but not verbs or adjectives). As we see below, better results are produced by this variant.

6.3. From COCO 5K to COCO 22K

We tested our method on 22K images of the MS COCO database. With more images, the difficulty of the task increases. Motivated by the fact that CN seems to be

TABLE 6.4. Image retrieval results for COCO 5K and 20K. In this table we compare our best baseline against a version of CN_ESP_MAX which only detect new noun words. Performance improvement increases when more images are considered.

Databases	r@1	r@5	r@10	median	mean
From 5K to 22K				rank	rank
COCO 5K					
MIL_STEM	15.9	38.3	51.0	10	49.9
CN_ESP_MAX (NN)	16.3	39.2	51.9	10	44.5
Improvement (%)	2.5	2.4	1.8	0	-10.8
COCO 22K					
MIL_STEM	7.0	18.7	26.6	43	224.6
CN_ESP_MAX	7.1	19.1	27.1	42	198.8
CN_ESP_MAX (NN)	7.1	19.2	27.2	41	199.7
Improvement (%)	1.4	2.7	2.3	-5	-46.7

best at noun detection (*c.f.*, Table 6.3), we designed a version of CN_ESP_MAX, called CN_ESP_MAX(NN), whose CN-score only focuses on undetectable nouns.

Table 6.4 shows the results for MIL_STEM and CN_ESP_MAX (NN). We show results over COCO 5K and 22K. Interestingly, the improvement of CN_ESP_MAX over MIL_STEM *increases* when we added more images. Notably, we improve upon the median score, which is a good measure of a significant improvement.

7. CONCLUSIONS AND PERSPECTIVES

This thesis presented an approach to enhancing a learning-based technique for sentence-based image retrieval with general-purpose knowledge provided by ConceptNet, a large commonsense ontology. Our experimental data, restricted to the task of image retrieval, shows improvements leading to state-of-the-art performance. This suggests a promising research area where the benefits of integrating the areas of knowledge representation and computer vision should continue to be explored.

An important lesson learned while carrying out this work is that integration of a general-purpose ontology with a vision approach is not straightforward. This is illustrated by the experimental data that showed that information in the ontology alone did not improve performance, while the *combination* of an ontology and crowd-sourced visual knowledge (from ESPGAME) did.

We believe there are many perspectives for future work. For example, an issue we did not address is one related to the polysemy and synonymy issues. For example, in ConceptNet the node *fly* has relations for the *verb fly* and for the *insect fly*; meanwhile, *computer* has relations to concepts that *pc* does not. The stemming process—that we needed to use in order to apply CN—may sometimes introduce ambiguity; for example, by transforming an undetectable word such as *bowling* into *bowl*. Our approach also can be further optimized for the task of image retrieval. For example, currently to detect *tuxedo*, we look for images with high scores for *jacket*, *black*, etc., but we do not constrain these detectors to fire at the same location.

References

- Belongie, S., Malik, J., & Puzicha, J. (2000). Shape context: A new descriptor for shape matching and object recognition. In *Nips* (Vol. 2, p. 3).
- Biederman, I. (1972). Perceiving real-world scenes. *Science*, 177(4043), 77–80.
- Bourdev, L., Maji, S., Brox, T., & Malik, J. (2010). Detecting people using mutually consistent poselet activations. In *European conference on computer vision (ECCV)* (pp. 168–181).
- Boureau, Y.-l., Cun, Y. L., et al. (2008). Sparse feature learning for deep belief networks. In *Advances in neural information processing systems* (pp. 1185–1192).
- Bush, V. (1979). As we may think. *ACM SIGPC Notes*, 1(4), 36–44.
- Chen, X., Shrivastava, A., & Gupta, A. (2013). Neil: Extracting visual knowledge from web data. In *Computer vision (iccv), 2013 ieee international conference on* (pp. 1409–1416).
- Choi, M. J., Lim, J. J., Torralba, A., & Willsky, A. S. (2010). Exploiting hierarchical context on a large database of object categories. In *Computer vision and pattern recognition (cvpr), 2010 ieee conference on* (pp. 129–136).
- Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer vision and pattern recognition, 2005. cvpr 2005. ieee computer society conference on* (Vol. 1, pp. 886–893).
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 248–255).

- Desai, C., Ramanan, D., & Fowlkes, C. C. (2011). Discriminative models for multi-class object layout. In (Vol. 95, pp. 1–12). Springer.
- Deselaers, T., & Ferrari, V. (2011). Visual and semantic similarity in imagenet. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on* (pp. 1777–1784).
- Divvala, S. K., Farhadi, A., & Guestrin, C. (2014). Learning everything about anything: Webly-supervised visual concept learning. , 3270–3277.
- Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2014). Long-term recurrent convolutional networks for visual recognition and description..
- Espinace, P., Kollar, T., Roy, N., & Soto, A. (2013). Indoor scene recognition by a mobile robot through adaptive object detection. *Robotics and Autonomous Systems*, 61(9), 932–947.
- Fang, C., & Torresani, L. (2012). Measuring image distances via embedding in a semantic manifold. In *Computer vision–eccv 2012* (pp. 402–415). Springer.
- Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., . . . others (2015). From captions to visual concepts and back. In *Computer vision and pattern recognition (cvpr), 2015 ieee conference on*.
- Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 1778–1785).
- Galleguillos, C., & Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6), 712–722.
- Grauman, K., & Leibe, B. (2010). *Visual object recognition* (No. 11). Morgan & Claypool Publishers.

- Havasi, C., Speer, R., & Alonso, J. (2007). Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing* (pp. 27–29).
- Karpathy, A., & Li, F. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on computer vision and pattern recognition (CVPR)* (pp. 3128–3137). IEEE.
- Klein, B., Lev, G., Sadeh, G., & Wolf, L. (2015). Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on computer vision and pattern recognition (CVPR)* (pp. 4437–4446). IEEE.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).
- Lenat, D. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), 33–38.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision—eccv 2014* (pp. 740–755). Springer.
- Liu, H., & Singh, P. (2004). Conceptnet a practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4), 211–226.
- Liu, J., Kuipers, B., & Savarese, S. (2011). Recognizing human actions by attributes. In *Computer vision and pattern recognition (cvpr), 2011 IEEE conference on* (pp. 3337–3344).

- Lobel, H., Vidal, R., & Soto, A. (2013). Hierarchical joint max-margin learning of mid and top level representations for visual recognition. In *Computer vision (iccv), 2013 ieee international conference on* (pp. 1697–1704).
- Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. the proceedings of the seventh ieee international conference on* (Vol. 2, pp. 1150–1157).
- Maillot, N. E., & Thonnat, M. (2008). Ontology based complex object recognition. *Image and Vision Computing*, 26(1), 102–113.
- Malisiewicz, T., & Efros, A. (2009). Beyond categories: The visual memex model for reasoning about object relationships. In *Advances in neural information processing systems* (pp. 1222–1230).
- Marques, O., Barenholtz, E., & Charvillat, V. (2011). Context modeling in computer vision: techniques, implications, and applications. *Multimedia Tools and Applications*, 51(1), 303–339.
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- Ohlsson, S., Sloan, R. H., Turán, G., & Urasky, A. (2013). Verbal IQ of a four-year old achieved by an AI system. In *Proceedings of the 27th aai Conference on Artificial Intelligence (AAAI)*. Bellevue, WA.
- Ordonez, V., Liu, W., Deng, J., Choi, Y., Berg, A. C., & Berg, T. L. (2015). Predicting entry-level categories. *International Journal of Computer Vision*, 1–15.
- Parikh, D., & Grauman, K. (2011). Relative attributes. In *Computer vision (iccv), 2011 ieee international conference on* (pp. 503–510).
- Peissig, J., & Tarr, M. (2007). Visual object recognition: do we know more now than we did 20 years. *Annual Review of Psychology*, 50, 75-96.

- Peralta, B., Espinace, P., & Soto, A. (2012). Adaptive hierarchical contexts for object recognition with conditional mixture of trees. In *Bmvc*.
- Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In *Computer vision, 2007. iccv 2007. ieee 11th international conference on* (pp. 1–8).
- Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002). Open mind common sense: Knowledge acquisition from the general public. In *On the move to meaningful internet systems 2002: Coopis, doa, and odbase* (pp. 1223–1237). Springer.
- Singh, S., Gupta, A., & Efros, A. (2012). Unsupervised discovery of mid-level discriminative patches. In (pp. 73–86). Springer.
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Computer vision and pattern recognition (cvpr), 2015 ieee conference on*.
- Von Ahn, L., & Dabbish, L. (2004a). Labeling images with a computer game. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 319–326).
- Von Ahn, L., & Dabbish, L. (2004b). Labeling images with a computer game. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 319–326).
- von Ahn, L., Kedia, M., & Blum, M. (2006). Verbosity: a game for collecting common-sense facts. In *Proceedings of the 2006 Conference on human factors in computing systems (CHI)* (p. 75-78).

- Wang, J., Chen, Z., & Wu, Y. (2011). Action recognition with multiscale spatio-temporal contexts. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on* (pp. 3185–3192).
- Weston, J., Bengio, S., & Usunier, N. (2011). Wsabie: Scaling up to large vocabulary image annotation. In *Ijcai* (Vol. 11, pp. 2764–2770).
- Wolf, L., & Bileschi, S. (2006). A critical view of context. *International Journal of Computer Vision*, 69(2), 251–261.
- Yang, J., Yu, K., Gong, Y., & Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 1794–1801).
- Zhu, Y., Fathi, A., & Fei-Fei, L. (2014). Reasoning about object affordances in a knowledge base representation. In *Computer vision–eccv 2014* (pp. 408–424). Springer.