How a General-Purpose Commonsense Ontology can Improve Performance of Learning-Based Image Retrieval*

Rodrigo Toro Icarte[†], **Jorge A. Baier**[‡,§], **Cristian Ruz**[‡], **Alvaro Soto**[‡]
[§]Chilean Center for Semantic Web Research,
[‡]Pontificia Universidad Católica de Chile,
[†]University of Toronto
rntoro@cs.toronto.edu, {jabaier, cruz, asoto}@ing.puc.cl

Abstract

The knowledge representation community has built general-purpose ontologies which contain large amounts of commonsense knowledge over relevant aspects of the world, including useful visual information, e.g.: "a ball is *used by* a football player", "a tennis player is *located at* a tennis court". Current state-of-the-art approaches for visual recognition do not exploit these rule-based knowledge sources. Instead, they learn recognition models directly from training examples. In this paper, we study how general-purpose ontologies—specifically, MIT's ConceptNet ontology—can improve the performance of state-of-the-art vision systems. As a testbed, we tackle the problem of sentence-based image retrieval. Our retrieval approach incorporates knowledge from ConceptNet on top of a large pool of object detectors derived from a deep learning technique. In our experiments, we show that ConceptNet can improve performance on a common benchmark dataset. Key to our performance is the use of the ESPGAME dataset to select visually relevant relations from ConceptNet. Consequently, a main conclusion of this work is that general-purpose commonsense ontologies improve performance on visual reasoning tasks when properly filtered to select meaningful visual relations.

1 Introduction

The knowledge representation community has recognized that commonsense knowledge bases are needed for reasoning in the real world. Cyc [Lenat, 1995] and ConceptNet (CN) [Havasi *et al.*, 2007] are two well-known examples of large, publicly available commonsense knowledge bases.

CN has been used successfully for tasks that require rather complex commonsense reasoning, including 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 *et al.*, 2013]. CN also contains many assertions that seem visually relevant; such as *"a chef is (usually) located at the kitchen"*.

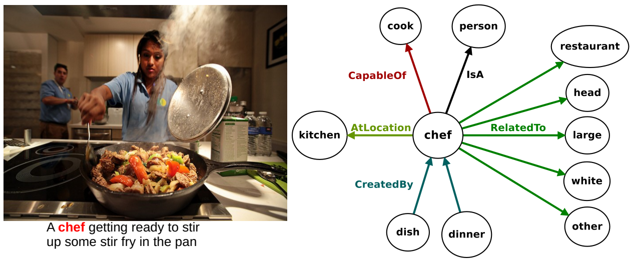

Figure 1: **Left.** An image and one of its associated sentences from the MS COCO dataset. Among its words, the sentence features the word *Chef*, for which there is not a visual detector available. **Right.** Part of the hypergraph at distance 1 related to word *Chef* in ConceptNet. In the list of nodes related to the concept *Chef*, there are several informative concepts for which we have visual detectors available.

State-of-the-art approaches to visual recognition tasks are mostly based on learning techniques. Some use mid-level representations [Singh *et al.*, 2012; Lobel *et al.*, 2013], others deep hierarchical layers of composable features [Ranzato *et al.*, 2008; Krizhevsky *et al.*, 2012]. Their goal is to uncover visual spaces where visual similarities carry enough information to achieve robust visual recognition. While some approaches exploit knowledge and semantic information [Liu *et al.*, 2011; Espinace *et al.*, 2013], none of them utilize large-scale ontologies to improve performance.

In terms of CN, previous works have suggested that incorporating CN knowledge to visual applications is nontrivial [Le *et al.*, 2013; Xie and He, 2013; Snoek *et al.*, 2007]. Indeed, poor results in [Le *et al.*, 2013] and [Xie and He, 2013] can be attributed to a non-negligible rate of noisy relations in CN. The work in [Snoek *et al.*, 2007] helps to support this claim: "...manual process (of CN relations) guarantees high quality links, which are necessary to avoid obscuring the experimental results."

In this paper we study the question of how large and publicly available general-purpose commonsense knowledge repositories, specifically CN, can be used to improve state-of-the-art vision techniques. We focus on the problem of sentence-based image retrieval. We approach the problem by assuming that we have visual detectors for a number of words,

*This work was partially funded by grants FONDECYT 1151018 and 1150328. Rodrigo also gratefully acknowledges funding from CONICYT (Becas Chile). We thank the IJCAI-17 reviewers and Margarita Castro for providing valuable insights into this research.

and describe a CN-based method to enrich the existing set of detectors. Figure 1 shows an illustrative example: An image retrieval query contains the word *Chef*, for which there is not a visual detector available. In this case, the information contained in the nodes directly connected to the concept *Chef* in CN provides key information to trigger related visual detectors, such as *Person*, *Dish*, and *Kitchen* that are highly relevant to retrieve the intended image.

Given a word w for which we do not have a visual detector available, we propose various probabilistic-based approaches to use CN's relations to estimate the likelihood that there is an object for w in a given image. Key to the performance of our approach is an additional step that uses a complementary source of knowledge, the ESPGAME dataset [Von Ahn and Dabbish, 2004], to filter out noisy and non-visual relations provided by CN. Consequently, a main conclusion of this work is that filtering out relations from CN is very important for achieving good performance, suggesting that future work that attempts to integrate pre-existing general knowledge with machine learning techniques should put close attention to this issue.

The rest of the paper is organized as follows: Section 2 reviews related work; Section 3 describes the elements used in this paper; Sections 4 and 5 motivate and describe our proposed method; Section 6 presents qualitative and quantitative experiments on standard benchmark datasets; finally, Section 7 presents future research directions and concluding remarks.

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 [Biederman, 1972]. In computer vision, the main focus has been on using contextual relations in the form of object co-occurrences, and geometrical and spatial constraints. Due to space constraints, we refer the reader to [Marques *et al.*, 2011] for an in-depth review about these topics. As a common issue, these methods do not employ high-level semantic relations as the one included in CN.

Knowledge acquisition 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 [Deng *et al.*, 2009]. As an alternative, recently, Chen *et al.* [2013] and Divvala *et al.* [2014] present 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 commonsense 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 by Von Ahn and Dabbish [2004], public information is biased. In particular, commonsense knowledge is so obvious that it is commonly 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.

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

Figure 2: A sample of CN relations that involve the concept *sofa*, together with the English description provided by the CN team in their website.

Recently, work on automatic image captioning has made great advances to integrate image and text data [Karpathy and Fei-Fei, 2015; Vinyals *et al.*, 2015; Klein *et al.*, 2015]. 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, work by Karpathy and Fei-Fei; Vinyals *et al.* [2015; 2015] share similar ideas which follow initial work by Weston *et al.* [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. [Klein *et al.*, 2015], on the other hand, proposes to use the Fisher Vector as a sentence representation instead of recurrent neural networks. In contrast to our 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] builds a custom ontology to perform visual object recognition. [Ordonez *et al.*, 2015] uses 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] uses Markov Logic Networks and a custom ontology to identify several properties related to object affordance in images. In contrast to our work, these methods target different applications. Furthermore, they do not exploit the type of commonsense relations that we want to extract from CN.

3 Preliminaries

ConceptNet ConceptNet (CN) [Havasi *et al.*, 2007] is a commonsense-knowledge semantic network which represents knowledge in a hypergraph structure. Nodes in the hypergraph correspond to a concept represented by a word or a phrase. In addition, hyperarcs 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 knowledge base "containing lots of things computers should know about the world, especially when understanding text written by people."

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 (see Figure 2). 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*).

Even though CN's assertions are reasonably accurate [Singh *et al.*, 2002] and that it can be used to score as good as a four-year-old in an IQ test [Ohlsson *et al.*, 2013], it 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.

Stemming A standard Natural Language Processing technique used below is *stemming*. The stemming of word w is an English word resulting from stripping a suffix out of w. It is a heuristic process that aims at returning 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 $st(W) = \{st(w) : w \in W\}$.

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 [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. 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 [Fang *et al.*, 2015] to train the detectors). Mathematically,

$$\mathrm{MIL}(t, I) = \prod_{w \,\in\, V \,\cap\, S_t} \hat{P}_V(w|I), \tag{1}$$

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

The main assumption behind Equation 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 Section 6, we show that, although simple, this score outperforms previous works (*e.g.*, [Klein *et al.*, 2015; Karpathy and Fei-Fei, 2015]).

5 CN-based Detector Enhancement

The MIL score has two limitations. First, it considers the text query as a set of independent words, ignoring their semantic 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 be 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.

5.1 CN for Undetectable Words

Our goal is to provide a score to each image analogous to that defined in Equation 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: for most words representing a concept c, CN "knows" a number of concepts related to c that share related visual characteristics. For example, if w is *tuxedo*, then *jacket* may provide useful information since "tuxedo–*IsA*→ jacket" is in CN.

We define $cn(w)$ as the set of concepts that are directly 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 $w' \in cn(w)$:

$$P(w|I) = P(w|w', I)P(w'|I) + P(w|\neg w', I)P(\neg w'|I), \tag{2}$$

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 2 can be re-stated in terms of estimations. $P(w|I)$ can be estimated by $\hat{p}(w', w, I)$, which is defined by

$$\hat{p}(w', w, I) \stackrel{\text{def}}{=} \hat{P}(w|w', I)\hat{P}(w'|I) + \hat{P}(w|\neg w', I)\hat{P}(\neg w'|I). \tag{3}$$

However the drawback 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 words in V whose detectors can be used to detect w after stemming. Now we define $\hat{P}(w'|I)$ as:

$$\hat{P}(w'|I) = \max_{w \in stDet(w', V)} \hat{P}_V(w|I), \tag{4}$$

i.e., to estimate how likely it is that w' is in I, we look for a word w in the set of 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. 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 approximate $\hat{P}(w|w', I)$ by $P(w|w')$; 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 [Von Ahn and Dabbish, 2004], which contains tags for many images. 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 3 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:

$$\hat{P}_{\mathcal{F}}(w|I) = \mathcal{F}_{w' \in cnDet(w,V)} \hat{p}(w', w, I), \qquad (5)$$

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,

$$\text{MILSTEM}(t, I) = \text{MIL}(t, I) \times \prod_{w \in W'_t} \hat{P}(w|I), \qquad (6)$$

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 $stDet(w, V) \neq \emptyset$), and where $\hat{P}(w|I)$ is defined by Equation 4.

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.* $stDet(w, V) = \emptyset$), but (3) they have at least one related concept in CN for which there is a detector (*i.e.*, $cnDet(w, V) \neq \emptyset$). Then we define:

$$\text{CN}_{\mathcal{F}}(t, I) = \text{MILSTEM}(t, I) \times \prod_{w \in W''_t} \hat{P}_{\mathcal{F}}(w|I), \qquad (7)$$

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

Code Our source code is publicly available at the following repository: https://bitbucket.org/RToroIcarte/cn-detectors.

6 Results and Discussion

We evaluate our algorithm over the MS COCO image database [Lin *et al.*, 2014]. Each image in this set contains 5 natural-language descriptions. Following [Karpathy and Fei-Fei, 2015] and [Klein *et al.*, 2015] we use a specific subset of 5K images and evaluate the methods on the union of the sentences for each image. We refer to this subset as COCO 5K.

We report the mean and median rank of the *ground truth* image; that is, the one that is 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 is found among the top k results.

Recall that we say that a word w is detectable when there is a detector for w. In this section we use Fang *et al.*'s detectors, which is comprised by 616 detectors for nouns, 176 for verbs, and 119 for adjectives. 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 is 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 Section 5. We consider that a comparison to this method is important because we want to evaluate the impact of using an ontology with general-purpose knowledge versus using a crowd-sourced, mainly visual knowledge such as that in ESPGAME.

Table 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 and ESP_OP (with OP = MIN (minimum), MEAN_G (geometric mean), MEAN_A (arithmetic mean) and MAX (maximum)). Results show that algorithms based on CN perform better in all the reported metrics, including the median rank. Overall, the MAX version of CN seems to obtain the best results and thus we focus our analysis on it.

Figure 3 shows qualitative results for 3 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_MAX. Green words on the query are stem-detectable and red words are CN-detectable but not stem-detectable.

Query 1 shows an example 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 and, as a result, the correct image is under

Database ⊂ COCO 5K	r@1	r@5	r@10	median rank	mean rank
Our baselines					
MIL	13.2	33.4	45.2	13	82.2
MILSTEM	13.5	33.8	45.7	13	74.6
Without CN					
ESP_MIN	12.6	30.7	41.1	17	122.4
ESP_MEAN_G	13.5	34.0	46.0	13	70.5
ESP_MEAN_A	13.6	34.2	46.2	13	69.0
ESP_MAX	13.5	33.7	45.7	13	66.2
Using CN					
CN_MIN	14.3	34.6	46.6	**12**	68.3
CN_MEAN_G	14.5	35.2	47.3	**12**	64.3
CN_MEAN_A	**14.6**	35.6	48.0	**12**	61.2
CN_MAX	14.3	**35.9**	**48.2**	**12**	**60.6**

Table 1: Subset of COCO 5K with sentences that contain at least one undetectable word.

Algorithm CN_MAX	r@1	r@5	r@10	median rank	mean rank
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

Table 2: Image Retrieval for new word detectors over COCO 5K. We include a random baseline, and results for CN_MAX divided in 4 categories: Retrieving nouns, verbs, adjectives and all of them. The results show that it is easier to detect nouns than verbs or adjectives.

Database COCO 5K	r@1	r@5	r@10	median rank	mean rank
Other approaches					
NeuralTalk	6.9	22.1	33.6	22	72.2
GMM+HGLMM	10.8	28.3	40.1	17	49.3
BRNN	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_MAX	**16.2**	**39.1**	**51.9**	**10**	**44.4**

Table 3: Image retrieval results over COCO 5K. References: NeuralTalk [Vinyals *et al.*, 2015], BRNN [Karpathy and Fei-Fei, 2015], and GMM+HGLMM [Klein *et al.*, 2015].

r@25. Query 2 shows an example where we have both stem-detectable words and CN-detectable words (that are not stem-detectable). In these cases, CN_MAX is able to detect "*bagel*" using the "*doughnut*" and "*bread*" detectors (among others), improving the ranking of the correct image. The last query is a case for which the CN-score is detrimental. For Query 3, the word "*resort*" is highly related with "*hotel*" in both CN and ESPGAME, thus the "*hotel*" detector became more relevant than "*looking*" and "*hills*".

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 design the following experiment: we consider the set of words from the union of text tags that are only CN-detectable, and we interpret those as one-word queries. An image is ground truth in this case if any of its tags contains w.

Results in Table 2 are disaggregated for word type (Nouns, Verbs, Adjectives). As a reference of the problem difficulty, we add 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_MAX.

We observe that numbers are lower than in Table 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 4 shows an example of this phenomenon for the word *tuxedo*.

Visual knowledge from ESPGAME is key We experiment with three alternative ways to compute CN-detectable scores which do not yield good results. First, we use CN considering $\hat{P}(w|w', I) = 1$ and $\hat{P}(w|\neg w', I) = 0$ for Equation 3. We also try estimating $\hat{P}(w|w', I)$ using CN weights. Finally, we use the similarity measure of Word2Vector [Mikolov *et al.*, 2013] as an alternative to ESPGAME. All those variants perform worse than ESP_MEAN_G.

Considering the relation types We explore using different subsets of relation types, but the performance always decreases. To our surprise, even removing the "Antonym" relation decreases the overall performance. We study this phenomenon and discover that a high number of antonym relationships are visually relevant (e.g. "cat –*Antonym*→ dog"). As CN is built by people, we believe that even non-visual relation types, such as "Antonym", are biased towards concepts that are somehow related in our mind. This might partially explain why the best performance is obtained by letting ESPGAME to choose visually relevant relation instances without considering their type. Nonetheless, we believe that relation types are a valuable information that we have not discovered how to properly use.

6.2 Comparison to Other Approaches

We compare with NeuralTalk[1] [Vinyals *et al.*, 2015], BRNN [Karpathy and Fei-Fei, 2015], and GMM+HGLMM (the best algorithm in [Klein *et al.*, 2015]) over COCO 5K. To reduce the impact of noisy relations in ConceptNet in CN_MAX, we only consider relationships with CN confidence weight ≥ 1 (this threshold is defined by carrying out a sensitivity analysis). As we can see on table 3, MIL outperforms previous approaches to image retrieval. Moreover, adding CN to detect new words improves the performance in almost all metrics.

6.3 From COCO 5K to COCO 22K

We test 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 best at noun detection (*c.f.*, Table 2), we design a version of CN_MAX, called CN_MAX (NN), whose CN-score only focuses on undetectable nouns.

[1] https://github.com/karpathy/neuraltalk.

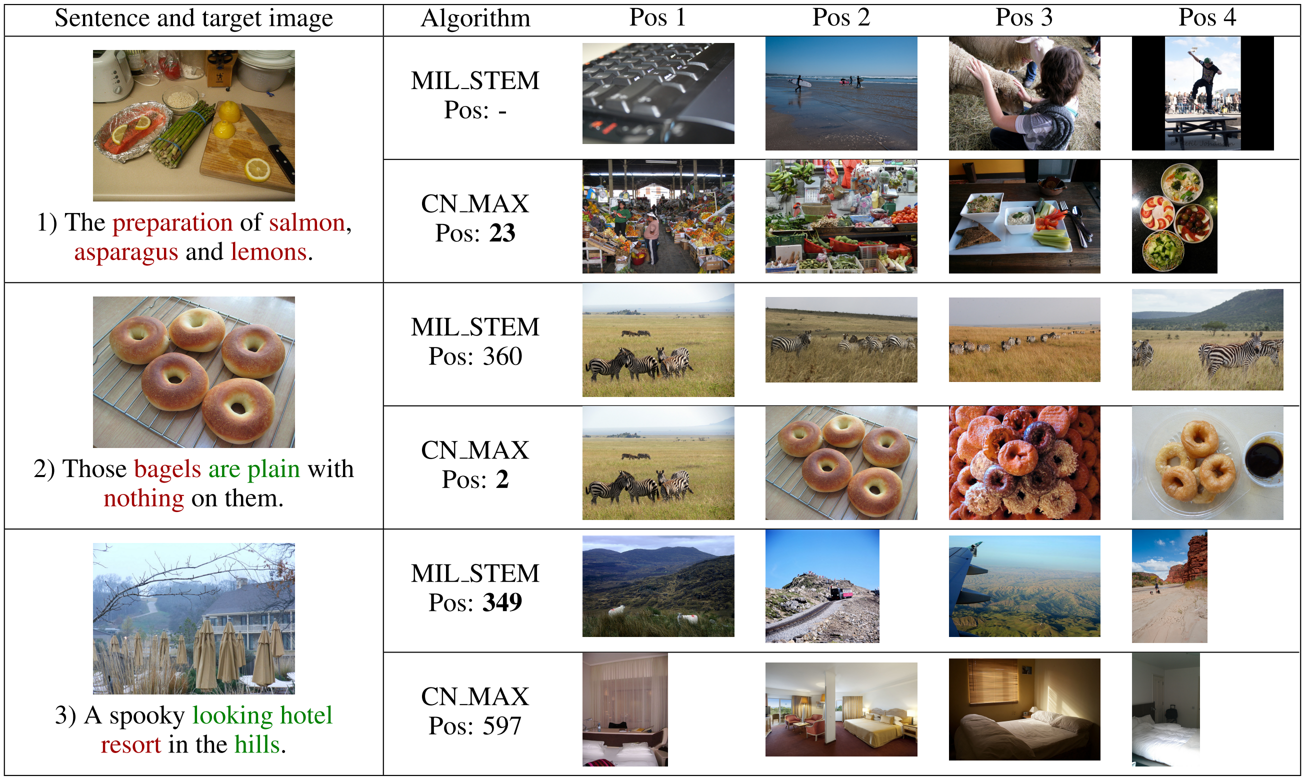
Sentence and target image
Algorithm
Pos 1
Pos 2
Pos 3
Pos 4
1) The preparation of salmon, asparagus and lemons.
MIL_STEM Pos: -
CN_MAX Pos: 23
2) Those bagels are plain with nothing on them.
MIL_STEM Pos: 360
CN_MAX Pos: 2
3) A spooky looking hotel resort in the hills.
MIL_STEM Pos: 349
CN_MAX Pos: 597

Figure 3: Qualitative examples for our baseline "MIL_STEM" and our method "CN_MAX" over COCO 5K. Green words are stemming-detectable, whereas red words are only CN-detectable.

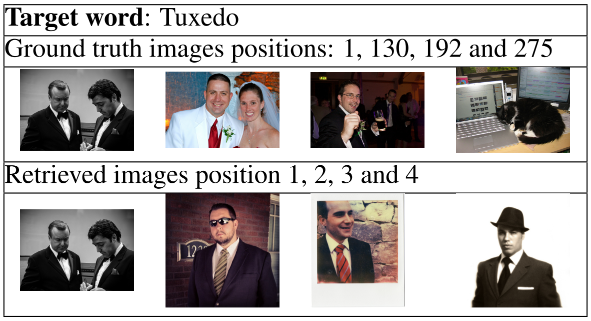
Target word: Tuxedo
Ground truth images positions: 1, 130, 192 and 275
Retrieved images position 1, 2, 3 and 4

Figure 4: 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_MAX.

Databases From 5K to 22K	r@1	r@5	r@10	median rank	mean rank
COCO 5K					
MIL_STEM	15.9	38.3	51.0	10	49.9
CN_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_MAX	7.1	19.1	27.1	42	198.8
CN_MAX (NN)	7.1	19.2	27.2	41	199.7
Improvement (%)	1.4	**2.7**	**2.3**	**5**	**46.7**

Table 4: Image retrieval results for COCO 5K and 22K. In this table we compare our best baseline against a version of CN_MAX which only detects new noun words. Performance improvement increases when more images are considered.

Table 4 shows the results for MIL_STEM and CN_MAX (NN). We show results over COCO 5K and 22K. Interestingly, the improvement of CN_MAX over MIL_STEM *increases* when we add more images. Notably, we improve upon the median score, which is a good measure of a significant improvement.

7 Conclusions and Perspectives

This paper 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 across different metrics and experimental settings. 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 conclusion of 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. This suggests that future works in the intersection of knowledge representation and vision may require special attention to relevance and knowledge base filtering.

References

[Biederman, 1972] Irving Biederman. Perceiving real-world scenes. *Science*, 177(4043):77–80, 1972.

[Chen *et al.*, 2013] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, pages 1409–1416, 2013.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Divvala *et al.*, 2014] Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, pages 3270–3277, 2014.

[Espinace *et al.*, 2013] Pablo Espinace, Thomas Kollar, Nicholas Roy, and Alvaro Soto. Indoor scene recognition by a mobile robot through adaptive object detection. *Robotics and Autonomous Systems*, 61(9):932–947, 2013.

[Fang *et al.*, 2015] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015.

[Havasi *et al.*, 2007] Catherine Havasi, Robert Speer, and Jason Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *RANLP*, pages 27–29, 2007.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[Klein *et al.*, 2015] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, pages 4437–4446, 2015.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Le *et al.*, 2013] Dieu-Thu Le, Jasper RR Uijlings, and Raffaella Bernardi. Exploiting language models for visual recognition. In *EMNLP*, pages 769–779, 2013.

[Lenat, 1995] Douglas B Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[Liu *et al.*, 2011] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR*, pages 3337–3344, 2011.

[Lobel *et al.*, 2013] Hans Lobel, René Vidal, and Alvaro Soto. Hierarchical joint max-margin learning of mid and top level representations for visual recognition. In *ICCV*, pages 1697–1704, 2013.

[Maillot and Thonnat, 2008] Nicolas Eric Maillot and Monique Thonnat. Ontology based complex object recognition. *Image and Vision Computing*, 26(1):102–113, 2008.

[Marques *et al.*, 2011] Oge Marques, Elan Barenholtz, and Vincent Charvillat. Context modeling in computer vision: techniques, implications, and applications. *Multimedia Tools and Applications*, 51(1):303–339, 2011.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mitchell, 1997] Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.

[Ohlsson *et al.*, 2013] Stellan Ohlsson, Robert H Sloan, György Turán, and Aaron Urasky. Verbal iq of a four-year old achieved by an ai system. In *AAAI*, pages 89–91, 2013.

[Ordonez *et al.*, 2015] Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. Predicting entry-level categories. *International Journal of Computer Vision*, 115(1):29–43, 2015.

[Rabinovich *et al.*, 2007] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *ICCV*, pages 1–8, 2007.

[Ranzato *et al.*, 2008] Marc'aurelio Ranzato, Y lan Boureau, and Yann L. Cun. Sparse feature learning for deep belief networks. In *NIPS*, pages 1185–1192, 2008.

[Singh *et al.*, 2002] Push Singh, Thomas Lin, Erik Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. *On the move to meaningful internet systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237, 2002.

[Singh *et al.*, 2012] Saurabh Singh, Abhinav Gupta, and Alexei Efros. Unsupervised discovery of mid-level discriminative patches. *ECCV 2012*, pages 73–86, 2012.

[Snoek *et al.*, 2007] Cees GM Snoek, Bouke Huurnink, Laura Hollink, Maarten De Rijke, Guus Schreiber, and Marcel Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on multimedia*, 9(5):975–986, 2007.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[Von Ahn and Dabbish, 2004] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *SIGCHI*, pages 319–326, 2004.

[Weston *et al.*, 2011] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, pages 2764–2770, 2011.

[Xie and He, 2013] Lexing Xie and Xuming He. Picture tags and world knowledge: learning tag relations from visual semantic sources. In *ACM-MM*, pages 967–976, 2013.

[Zhu *et al.*, 2014] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, pages 408–424, 2014.