Object Detection in Images using Knowledge Graphs

Double Blind

Abstract

Object detection in images is a crucial task in computer vision, with important applications ranging from security surveillance to autonomous vehicles. Existing state-of-the-art algorithms, including deep neural networks, only focus on utilizing features within an image itself, largely neglecting the vast amount of background knowledge about the real world. In this paper, we propose a novel framework of *knowledge-aware object detection*, which enables the integration of knowledge graphs with any detection algorithm. The framework employs the notion of semantic consistency to quantify and generalize knowledge, which improves object detection through an re-optimization to achieve better consistency with the background knowledge. Finally, empirical evaluation on two benchmark datasets show that our approach can significantly increase recall by up to 7.8 points without compromising mean average precision, when compared to the state-of-the-art baseline.

1 Introduction

Many computer vision tasks ultimately seek to interpret the world through data such as images and videos. While significant progress has been made in the past decade, there still exists a striking gap between how humans and machines learn. Although current machine learning approaches, including state-of-the-art deep learning algorithms, can effectively find patterns from the training data, they fail to leverage what an average person has at his or her disposal—the vast amount of background knowledge about the real world. Given that images and videos are reflections of the world, exploiting background knowledge can have a tremendous advantage towards interpreting these data.

Task and insight

In this paper, we study the key computer vision task of object detection [Felzenszwalb *et al.*, 2010]. Given an image, the goal is to identify a set of regions or bounding boxes, and to further classify each bounding box with one of the predefined object labels, as illustrated in Figure 1.

(a) Detecting `cat` and `table` (b) Detecting `bear`

Figure 1: Object detection on images from COCO15.

Recent advances in deep convolutional neural networks [Sermanet *et al.*, 2013; Girshick *et al.*, 2014], in particular Fast or Faster R-CNN [Girshick, 2015; Ren *et al.*, 2015], show great promise in object detection. However, like previous approaches, these methods only account for patterns present in the training images, without leveraging much of the knowledge an average person would have. For example, humans have the common sense or implicit knowledge that a domestic `cat` sometimes sits on a table, but a `bear` does not barring very rare circumstances. This background knowledge would natrually help reinforce the simultaneous detections of `cat` and `table` (e.g. in Figure 1a), even if none of the training images portrays a `cat` together with a `table`. On the other hand, if an image is predicted to contain both `bear` and `table`, which conflicts with our background knowledge, the detections are more prone to be false.

While such background knowledge appears random and difficult to organize, there have been extensive research and commercial efforts to encode it into machine readable forms often known as knowledge graphs [Paulheim, 2017]. A knowledge graph is a graph that models semantic knowledge, where each node is a real-world concept, and each edge represents a relationship between two concepts. For instance, Figure 2 showcases a toy knowledge graph. In particular, the relationship "`cat` sits on `table`" reinforces the detections of `cat` and `table` in Figure 1a. We note that knowledge graphs already demonstrate considerable success in other domains such as web search and social networks [Dong *et al.*, 2014]. Beyond a toy graph, large-scale knowledge graphs are often constructed through crowd sourcing or automated extraction from semi-structured and unstructured data, which are beyond the scope of this paper.

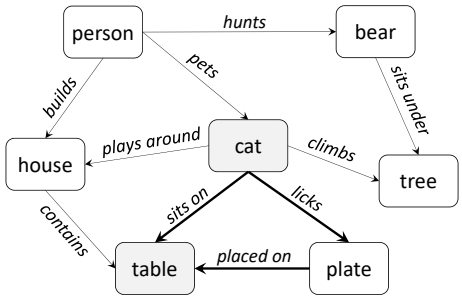


Figure 2: A toy knowledge graph modeling seven concepts as nodes (e.g., `cat` and `table`), as well as their relationships as edges (e.g., "`cat` sits on `table`").

Challenges and approach

Even with an existing knowledge graph, to effectively leverage its knowledge for object detection, two major technical challenges still remain.

First, how do we *quantify* and *generalize* knowledge? Quantification is the first step, as knowledge graphs entail symbolic representations but most object detection algorithms operate over numerical representations. Moreover, the quantification shall not only apply to images with contexts matching directly observed knowledge, but also generalize to images with new contexts. In our approach, for every pair of concepts on the knowledge graph, we compute a numerical degree of semantic consistency for them. For example, since the relationship "`cat` sits on `table`" is present on the knowledge graph, `cat` and `table` are semantically consistent concepts, but `bear` and `table` are not. Concepts can also be connected through a chain of indirect relationships, such as "`cat` licks `plate`" and "`plate` placed on `table`". This gives rise to the generalization ability—we can infer that `cat` and `table` tend to appear together without directly knowing "`cat` sits on `table`".

Second, how do we incorporate quantified knowledge for *knowledge-aware object detection*? We hinge on the key constraint that more semantically consistent concepts are more likely to occur in an image with comparable probability. For instance, letting (`o`, p) denote a bounding box containing object `o` with probability p, it is more plausible to have two bounding boxes (`cat`, 0.8) and (`table`, 0.8) rather than (`bear`, 0.8) and (`table`, 0.8) in the same image. In particular, for the latter, it is more plausible to have (`bear`, 0.8) and (`table`, 0.1) or (`bear`, 0.1) and (`table`, 0.8) instead. We cast such a constraint as an optimization problem.

Contribution

We make three major contributions in this paper. First, we advocate incorporating knowledge graphs into the object detection task, an emerging paradigm still limited in visual tasks. Second, we formulate a knowledge-aware framework that quantifies knowledge graphs in a generalizable manner, and re-optimizes object detection to achieve semantic consistency. Last, we conduct extensive evaluation on two benchmark datasets, which significantly improves recall by up to 7.8 points while keeping the same level of mean average precision.

2 Related Work

In recent years, deep convolutional neural networks (CNNs) have become the de-facto baseline for computer vision tasks such as image classification and object detection. Their strong performance stems from the ability to learn high-level image features [Krizhevsky *et al.*, 2012; Simonyan and Zisserman, 2014; Szegedy *et al.*, 2015; He *et al.*, 2016]. For object detection, earlier work such as Regions with CNN features (R-CNN) [Girshick *et al.*, 2014] and its fast variant [Girshick, 2015] uses CNNs to classify objects, but depends on precomputed region proposals for object localization. Subsequently, networks such as Overfeat [Sermanet *et al.*, 2013] and Faster R-CNN [Ren *et al.*, 2015] leverages CNNs for not only object classification but also object localization. Faster R-CNN in particular introduces a region proposal network that efficiently shares convolutional features for both region proposal and classification. More recent work such as Region-based Fully Convolutional Network (R-FCN) [Dai *et al.*, 2016] further makes the entire network convolutional and avoids the generation of sub-networks in the object classification stage. Using contextual information from the entire image has also been explored to improve object detection, by generating a context feature to enhance classification of individual regions [Bell *et al.*, 2016].

There is also an emerging trend to exploit information outside of one specific image, i.e., external background knowledge including natural language texts and knowledge graphs, for certain computer vision tasks such as visual motivation prediction [Vondrick *et al.*, 2016], question answering [Zhu *et al.*, 2015; Wu *et al.*, 2016] and relationship extraction [Lu *et al.*, 2016]. However, to date, using external background knowledge has received limited attention for the task of object detection. An early work [Rabinovich *et al.*, 2007] introduces a conditional random field (CRF) model to maximize the agreement of labels and semantic contexts learnt from the training data, as well as an external service called Google Sets[1] which returned a set of similar objects from a few examples. However, this method cannot generalize to images with contexts not observed in their training or external data, while our knowledge graph-based approach can generalize better. Furthermore, their model is computationally intractable and thus relies on heavy approximation.

The use of knowledge graphs have become widespread and largely successful in many data-driven applications including web search and social networks [Dong *et al.*, 2014]. Extensive efforts have been spent to construct large-scale knowledge graphs [Paulheim, 2017], which often require continuous expansion and refinement. Typically, knowledge graphs are built through human curation [Lenat, 1995], crowdsourced contribution [Liu and Singh, 2004], as well as automatic extraction from semi-structured data [Auer *et al.*, 2007; Suchanek *et al.*, 2007] or text data [Carlson *et al.*, 2010]. More recently, knowledge has also been systematically harvested from multimodal data including images [Zhu *et al.*, 2015; Krishna *et al.*, 2016].

[1]The product was discontinued in 2011.

3 Proposed Approach

We describe our knowledge-aware framework in this section, starting with the problem statement and notations, followed by the methods of quantifying and integrating knowledge into object detection.

3.1 Notation and problem

Consider a set of pre-defined concepts or object labels $\mathcal{L} = \{1, 2, \ldots, L\}$. We assume an existing object detection algorithm that outputs a set of bounding box $\mathcal{B} = \{1, 2, \ldots, B\}$ for each image, and predicts a label $\ell \in \mathcal{L}$ on each bounding box $b \in \mathcal{B}$ with probability $p(\ell|b)$. For each image, these probabilities can be encoded by a $B \times L$ matrix P, such that $P_{b,\ell} = p(\ell|b)$.

Our goal is to produce a new matrix $\widehat{P}$ by integrating knowledge into the initial matrix P. In other words, $\widehat{P}$ is a knowledge-aware enhancement of P. Ultimately, the new matrix $\widehat{P}$ enables us to improve object detection, such that a bounding box b is assigned a potentially new label $\widehat{\ell} = \arg\max_\ell \widehat{P}_{b,\ell}$. The overall framework is summarized in Fig. 3.

3.2 Knowledge quantification

Knowledge is fundamentally symbolic and logical. However, most state-of-the-art algorithms function on numerical representations. Thus, towards a knowledge-aware framework, the first step is to quantify such knowledge, especially in a manner that can generalize to images with unobserved contexts. To this end, we propose to measure a numerical degree of *semantic* consistency between each pair of concepts. A high degree of semantic consistency implies that two concepts are likely to appear together in the same image.

Formally, let S be an $L \times L$ matrix such that $S_{\ell,\ell'}$ is defined as the degree of semantic consistency between concepts (i.e., labels) ℓ and ℓ', $\forall(\ell, \ell') \in \mathcal{L}^2$. Naturally, S shall be symmetric (i.e., $S_{\ell,\ell'} = S_{\ell',\ell}$). Moreover, when $\ell = \ell'$, $S_{\ell,\ell'}$ captures the self-consistency, which is a meaningful measure since multiple instances of the same concept can appear in the same image.

In other words, additional background knowledge about various concepts can be quantified and modeled by the matrix S. In the following, we describe two alternatives of constructing S from additional knowledge: one using simple frequency, and the other based on a knowledge graph.

Frequency-based quantification

To compute semantic consistency, one immediate approach is to count the frequency of co-occurrences for each pair of concepts. Such co-occurrences can be identified from given background data, which can be potentially multi-modal including text corpora and photo collections.

In particular, let $f(\ell, \ell')$ denote the frequency of co-occurrences for concepts ℓ and ℓ', and $f(\ell)$ denote the frequency of ℓ. Then, we define the semantic consistency as the Jaccard index below.

$$S_{\ell,\ell'} = \frac{f(\ell, \ell')}{f(\ell) + f(\ell') - f(\ell, \ell')} \tag{1}$$

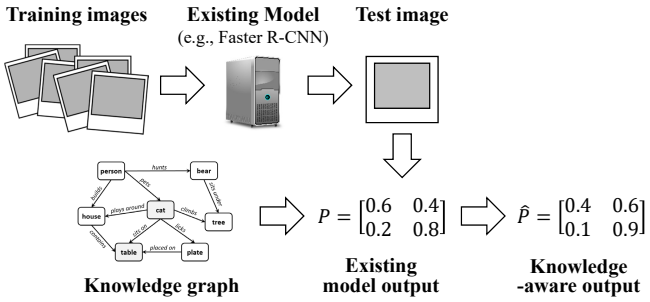


Figure 3: Overview of our knowledge-aware framework.

While it is straightforward to compute Eq. (1), there is a major drawback. The resulting matrix S only works for known co-occurrences in the background data, but does not generalize to unseen co-occurrences in new images. In other words, if two concepts never co-occur in the background data, their semantic consistency would be exactly zero, and thus are not helpful to new images containing these two concepts. Note that this problem is also inherent to the early CRF-based model that explores semantic contexts [Rabinovich *et al.*, 2007].

Knowledge graph-based quantification

Next, we consider a knowledge graph for modeling semantic consistency. Unlike the toy example in Figure 2, a typical off-the-shelf knowledge graph often captures at least millions of concepts and their complex relationships, providing immense background knowledge external to the images.

Using a large-scale knowledge graph has a significant advantage—it can better generalize to a pair of concepts even if they are not connected by any edge. In particular, when two concepts are not involved in a direct relationship, potentially we can still establish a chain of relationships between them. For instance, `people` and `plate` in Figure 2 are not directly connected. This does not necessarily mean that they are not semantically consistent. Quite the contrary, they should enjoy a fair degree of semantic consistency based on human knowledge. Nonetheless, despite a missing edge between them, there is still a chain of edges "`person` pets `cat`" and "`cat` licks `plate`") to indicate that they are semantically consistent to some extent. Furthermore, multiple direct relationships or chains of relationships can exist between two concepts. In Figure 2, `cat` and `table` can be related through the edge "`cat` sits on `table`", and a chain of edges "`cat` licks `plate`" and "`plate` placed on `table`". Each relationship or chain is called a *path* from `cat` to `table`. Different paths between the two concepts complement each other for increased robustness.

To quantify semantic consistency on a knowledge graph, we employ *random walk with restart* [Jeh and Widom, 2003; Tong *et al.*, 2006]. Starting from a node v_0 on the graph, we move to a random neighboring nodes of v_0, and record it as v_1. Once at v_1, we repeat this process. In general, when we are at v_t, we move to one of its neighbor randomly, and denote the new node we have just arrived as v_{t+1}. In addition, to avoid being trapped in a small locality, at each move, there is a probability of α to restart the random walk by "teleport-

ing" to the starting node v_0, instead of moving to one of the neighbors. Formally, a random walk is a sequence of nodes $\langle v_0, v_1, v_2 \ldots, v_t \rangle$, and $p(v_t = \ell' | v_0 = \ell; \alpha)$ represents the probability of reaching the concept ℓ' in t steps given that we start from ℓ, $\forall (\ell, \ell') \in \mathcal{L}^2$.

This probability can be used to formulate semantic consistency—a larger probability from ℓ to ℓ' implies that they are more semantically consistent. Intuitively, when the number of paths from ℓ to ℓ' increases or the length of these paths decreases, the semantic consistency between ℓ and ℓ' becomes larger, so does the probability of reaching ℓ' from ℓ. Interestingly, as we take longer random walks, this probability eventually converges to a unique steady state as follows.

$$R_{\ell,\ell'} = \lim_{t \to \infty} p(r_t = \ell' | r_0 = \ell; \alpha). \tag{2}$$

Note that $R_{\ell,\ell'}$ is not symmetric in general. Thus, in Eq. (3) we define a symmetric matrix S based on the geometric mean. The geometric mean has a roundtrip random walk interpretation, and has been shown to be superior than the arithmetic or harmonic means [Fang *et al.*, 2013]. The matrix S can be efficiently computed even on a very large knowledge graph [Zhu *et al.*, 2013; Fang *et al.*, 2013].

$$S_{\ell,\ell'} = \sqrt{R_{\ell,\ell'} R_{\ell',\ell}} \tag{3}$$

One caveat is the huge effort required to build and refine a large-scale knowledge graph, which itself is an active research area. Fortunately, a suite of off-the-shelf solutions are available, many of which offer open datasets or APIs. For a thorough discussion on this matter, we refer the reader to a survey paper [Paulheim, 2017] and the citations therein. In our experiments, we adopt MIT ConceptNet [Liu and Singh, 2004], a crowdsourced knowledge graph with more than 4 million concepts and 9 million relationships.

3.3 Knowledge integration

Given a matrix S that quantifies the semantic consistency between pairwise concepts, we need to further integrate it with an existing solution to enable knowledge-aware detection. In the following, we formulate a cost function based on S, and discuss its efficient optimization.

Cost function

The key intuition is that two concepts with a higher degree of semantic consistency are more likely to appear in the same image with comparable probability. That is, for two different bounding boxes b and b' in one image, $P_{b,\ell}$ and $P_{b',\ell'}$ should not be too different when $S_{\ell,\ell'}$ is large. This constraint can be formalized by minimizing the cost function in Eq. (4), where $\{P_{b,\ell} : b \in \mathcal{B}, \ell \in \mathcal{L}\}$ represent the detections from any existing algorithm, and $\{\widehat{P}_{b,\ell} : b \in \mathcal{B}, \ell \in \mathcal{L}\}$ represent our proposed knowledge-aware detections.

$$E(\widehat{P}, P) = (1-\epsilon) \sum_{b=1}^{B} \sum_{\substack{b'=1 \\ b' \neq b}}^{B} \sum_{\ell=1}^{L} \sum_{\ell'=1}^{L} S_{\ell,\ell'} \left(\widehat{P}_{b,\ell} - \widehat{P}_{b',\ell'} \right)^2 + \epsilon \sum_{b=1}^{B} \sum_{\ell=1}^{L} B \|S_{\ell,*}\|_1 \left(\widehat{P}_{b,\ell} - P_{b,\ell} \right)^2 \tag{4}$$

On the one hand, the first term of Eq. (4) captures the constraint on the semantic consistency. For a pair of detected bounding boxes b and b', if $S_{\ell,\ell'}$ is large, minimizing the objective function would force $P_{b,\ell}$ and $P_{b',\ell'}$ to become smaller; if $S_{\ell,\ell'}$ is small, $P_{b,\ell}$ and $P_{b',\ell'}$ are less constrained and can become very different.

On the other hand, the second term requires that knowledge-aware detections should not depart too much from detections of existing algorithms. Existing algorithms use features specific to each image which form the basis of knowledge-aware detections. Note that the squared error has a coefficient $B\|S_{\ell,*}\|_1$ in order to balance different concepts. Without this coefficient, the cost function would give more importance to the first term over summations involving $P_{b,\ell}, \forall b \in \mathcal{B}$ when $\|S_{\ell,*}\|$ is larger. The overall trade-off between the two terms is controlled by a hyperparameter $\epsilon \in (0, 1)$, which can be selected on a validation set.

Optimization

To minimize Eq. (4), we find its stationary point where its gradient w.r.t. $\widehat{P}_{b,\ell}$ is zero, $\forall b \in \mathcal{B}, \ell \in \mathcal{L}$.

$$\frac{\partial E(\widehat{P}, P)}{\partial \widehat{P}_{b,\ell}} = 4(1-\epsilon) \sum_{\substack{b'=1 \\ b' \neq b}}^{B} \sum_{\ell'=1}^{L} S_{\ell,\ell'} \left(\widehat{P}_{b,\ell} - \widehat{P}_{b',\ell'} \right) + 4\epsilon B \|S_{\ell,*}\|_1 \left(\widehat{P}_{b,\ell} - P_{b,\ell} \right) \tag{5}$$

Setting the above to zero, we obtain below an equivalent configuration over optimal $\widehat{P}_{b,\ell}, \forall b \in \mathcal{B}, \ell \in \mathcal{L}$.

$$\widehat{P}_{b,\ell} = (1-\epsilon) \frac{\sum_{b'=1, b' \neq b}^{B} \sum_{\ell'=1}^{L} S_{\ell,\ell'} \widehat{P}_{b',\ell'}}{\sum_{b'=1, b' \neq b}^{B} \sum_{\ell'=1}^{L} S_{\ell,\ell'}} + \epsilon P_{b,\ell} \tag{6}$$

It can be shown that the exact solution to Eq. (6) is the limit of the series in Eq. (7) for $i \in \{1, 2, \ldots\}$. In particular, for any arbitrary initialization $\widehat{P}^{(0)}_{b,\ell}$, $\widehat{P}^{(i)}_{b,\ell}$ always converges to the same solution as $i \to \infty$.

$$\widehat{P}^{(i)}_{b,\ell} = (1-\epsilon) \frac{\sum_{b'=1, b' \neq b}^{B} \sum_{\ell'=1}^{L} S_{\ell,\ell'} \widehat{P}^{(i-1)}_{b',\ell'}}{\sum_{b'=1, b' \neq b}^{B} \sum_{\ell'=1}^{L} S_{\ell,\ell'}} + \epsilon P_{b,\ell} \tag{7}$$

Note that the solution can be computed in polynomial time. The theoretical complexity is $O(B^2 L^2 I)$, where I is the number of iterations. Convergence typically happens very fast in fewer than 20 iterations. To further speed up the computation, we could iterate b' over only a subset of B_k bounding boxes (e.g., those of the smallest distance to b), and iterate ℓ' over only a subset of L_k labels (e.g., those with the largest semantic consistency to ℓ). In practice, we find out that $B_k \sim 50$ and $L_k \sim 10$ are already enough to achieve a near-perfect approximation. Thus, the practical complexity is only $O(BL)$, assuming that I, B_k, L_k are small constants.

4 Evaluation

We empirically evaluate the proposed approach on two benchmark datasets. Results of our knowledge-aware detection is promising, significantly outperforming the baseline method in recall while maintaining the same level of mean average precision.

4.1 Experimental setup

Datasets

We use benchmark data COCO15 [Lin *et al.*, 2014] and VOC07 [Everingham *et al.*, 2010]. A summary of their statistics is presented in Table 1. For COCO15, we combine their training and validation sets for training the baseline method, except for a subset of 5000 images named *minival*. We further split minival into two subsets with 1000 and 4000 images, named *minival-1k* and *minival-4k* respectively. We use *minival-1k* to choose hyperparameters in our approach, and *minival-4k* for offline evaluation. Online evaluation is performed on *test-dev* and *test-std* sets, where the latter only allows for limited submissions. For VOC07, we use their training set for training the baseline method, validation set for choosing our hyperparameters, and test set for evaluation.

Model training

We employ the state-of-the-art Faster R-CNN and VGG-16 as the baseline [Simonyan and Zisserman, 2014; Ren *et al.*, 2015], using the public Python Caffe implementation[2]. We simply call this baseline **FRCNN** hereafter. Models on both datasets are trained using stochastic gradient descent with a momentum of 0.9, a mini-batch size of 2 and a weight decay of 5e-4. On COCO15, we use a learning rate of 1e-3 for the first 350K iterations, followed by 1e-4 for another 140K iterations. On VOC07, we use a learning rate of 1e-3 for 50K iterations, and 1e-4 for the next 10K iterations. Layer weights for both models are initialized from a VGG-16 model pre-trained on ImageNet. New layers defined by Faster R-CNN are randomly initialised from a Gaussian distribution with a standard deviation of 0.01.

For our proposed knowledge-aware approach, we re-optimize the detections generated by FRCNN. We only retain bounding boxes of high confidence, requiring their scores to be at least 1e-5. We choose the hyperparameter ϵ in Eq. (4) on the validation set for $\epsilon \in \{0.25, 0.5, 0.75\}$. We run Eq. (7) for 10 iterations, which already show convergence. To construct the frequency-based semantic consistency, we use the training set as the background data for each dataset, respectively. For the knowledge graph-based method, we employ MIT ConceptNet 5[3] as our knowledge graph. We only use its subset in English, and filter out "negative" relationships (NotDesires, NotHasProperty, NotCapableOf, NotUsedFor, Antonym, DistinctFrom and ObstructedBy) and self-loops. The resulting graph has 1.3 million concepts and 2.7 million relationships. For the random walk restarting probability, we set $\alpha = 0.15$ which is a typical value shown to be stable [Fang *et al.*, 2013]. We call the two knowledge-aware alternatives **K-FREQ** and **K-CNET**, respectively.

Metrics.

The main metrics computed are mean average precision (MAP) and recall at top 100. A bounding box is judged correct only if its intersection over union (IoU) w.r.t. the ground truth is greater than a certain threshold. We vary the IoU threshold over $\{0.50, 0.55, \ldots, 0.95\}$, and report the average performance. As our knowledge-aware approach increases recall significantly while maintaining MAP, we examine the recall performance in more details. Specifically, on COCO15 we also report recall at top 10 as well as recall by the object areas (small, medium and large); on VOC07 we further report recall by concepts.

[2]https://github.com/rbgirshick/py-faster-rcnn

[3]http://conceptnet-api-1.media.mit.edu/

Dataset	# Concepts (i.e., labels)	# Images training	validation	test
COCO15	80	83K	41K	20K (dev/std)
VOC07	20	2.5K	2.5K	5.0K

Table 1: A summary of datasets.

	MAP @100	Recall @10	Recall @100	Recall@100 by area small	medium	large
minival-4k						
FRCNN	24.5	35.9	37.0	15.7	42.9	56.5
K-FREQ	24.6	36.5	40.9	18.0	47.7	62.1
K-CNET	24.5	**37.3**	**42.8**	**18.9**	**50.0**	**64.3**
test-dev						
FRCNN	24.2	34.0	34.6	12.0	38.5	54.4
K-FREQ	24.3	35.8	37.5	13.8	42.1	58.3
K-CNET	24.3	**36.2**	**38.6**	**14.1**	**43.3**	**60.1**
test-std (limited submission)						
FRCNN	24.2	34.1	34.7	11.5	38.9	54.4
K-CNET	24.1	**36.2**	**38.6**	**13.9**	**43.4**	**59.9**

Table 2: Comparison of our knowledge-aware approaches K-FREQ and K-CNET with baseline FRCNN on COCO15.

4.2 Main results

We report the results on COCO15 in Table 2. Both knowledge-aware approaches K-FREQ and K-CNET significantly increase recall@100 over the baseline method FRCNN by up to 3.9 and 5.8 points, respectively. Other recall metrics, including at top 10 and by areas, also show significant improvement up to 5.6 and 7.8 points, respectively. At the same time, both approaches do not compromise MAP.

We further report the results on VOC07 in Table 3. Likewise, both K-FREQ and K-CNET outperforms FRCNN in terms of recall@100 by 3.2 and 2.4 points, respectively. Moreover, they both outperform the baseline in 17 concepts out of 20.

Note that K-CNET is consistently better than K-FREQ on COCO15, whereas K-FREQ achieves slightly better results on VOC07. We hypothesize that the discrepancies are caused by the different complexity of the two benchmarks. In particular, COCO15 features many more concepts than VOC07: The former contains an average of 3.5 concepts and 7.7 object intances per image, whereas the latter contains less than 2 concepts and 3 object instances per image. Thus, the more complex scenes in COCO15 would require more generalization than the simpler scenes in VOC07. To validate this hypothesis, we compare the shift in concept co-occurrences from training to testing images. For a pair of concepts ℓ and ℓ', we define its shift as

$$|S^{\text{train}}_{\ell,\ell'} - S^{\text{test}}_{\ell,\ell'}| / \max(S^{\text{train}}_{\ell,\ell'}, S^{\text{test}}_{\ell,\ell'}) \times 100\%, \qquad (8)$$

	MAP @100	Recall@100 by concepts																				
		all	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
FRCNN	66.5	81.9	76.1	89.0	74.3	73.4	64.6	89.7	85.8	90.5	69.0	88.9	85.4	91.6	92.0	85.2	82.4	60.8	83.1	89.1	84.4	82.1
K-FREQ	66.6	**85.1**	80.0	**93.2**	**80.8**	**79.5**	**66.7**	**92.5**	**88.8**	**92.5**	68.0	**95.1**	**87.4**	**94.1**	**94.0**	**89.8**	81.1	**66.0**	**89.3**	88.7	**90.4**	**83.1**
K-CNET	66.1	84.3	**80.4**	90.8	78.6	76.8	66.1	90.1	**88.8**	91.6	**71.8**	92.6	85.0	92.6	90.5	88.9	**87.2**	65.8	87.6	**90.4**	89.0	82.1

Table 3: Comparison of our knowledge-aware approaches K-FREQ and K-CNET with baseline FRCNN on VOC07.

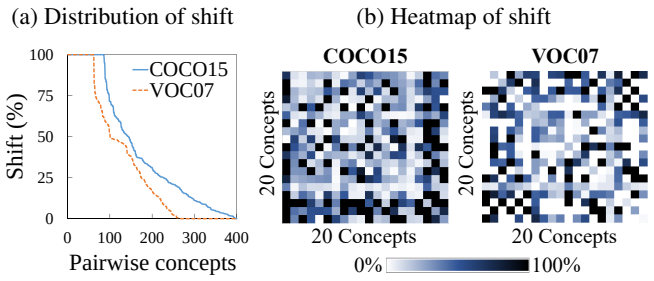


Figure 4: Shift in concept co-occurrences from training to testing images on 20 concepts common to both benchmarks.

where S^{train} and S^{test} are the co-occurrence matrices (Eq. 1) computed from the training and testing data, respectively. (For COCO15, the split *minival-4k* is used as the testing data since the groundtruth for *test-dev* and *test-std* is not publicly available.) Larger shifts imply that the training and testing images have more different co-occurrence patterns, and thus requires more generalization. The shift statistics are visualized in Fig. 4, which show that COCO15 have larger shifts and thus require more generalization. Not surprisingly, K-CNET performs significantly better on COCO15 due to its ability to generalize, whereas K-FREQ performs slightly better on VOC07 which requires less generalization.

Furthermore, the improvement of both approaches on COCO15 is larger than on VOC07, which could be caused by the same reason. We believe that knowledge-aware methods are able to beneifit more from the semantically richer scenes in COCO15.

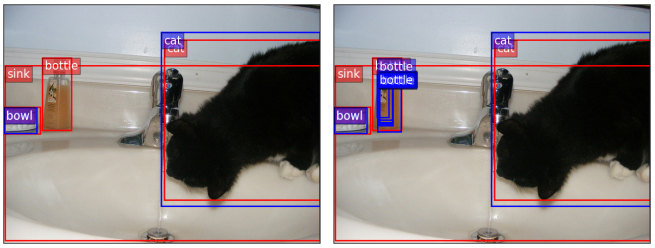
4.3 Case study

Finally, we showcase the power of the knowledge graph-based approach K-CNET on two real images from COCO15.

First, Figure 5a depicts a bathroom scene, containing ground-truth objects `bow` and `bottle`, among others. Although the baseline misses the `bottle`, it is successfully picked up by K-CNET after re-optimization. The reason is that the probability of `bottle` is promoted to become more similar to that of `bowl`, since the two concepts are semantically consistent. Indeed, their semantic consistency is 7.1 times of the median value among all pairwise concepts.

Second, Figure 5b depicts a sports scene, containing ground-truth objects `person` and `sports ball`. The `sports ball` is missed by the baseline but is correctly identified by K-CNET. In particular, the semantic consistency between the two concepts is 4.3 times of the median value.

(a) Bathroom scene: FRCNN (left) fails to detect `bottle`, but K-CNET (right) does due to the presence of `bowl`.



(b) Sports scene: FRCNN (left) fails to detect `sports ball`, but K-CNET (right) does due to the presence of `person`.

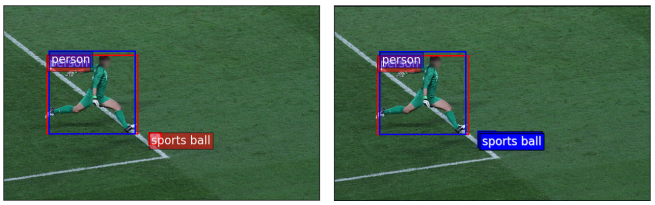


Figure 5: Two scenes from COCO15. In each scene, the *left* image contains the output of baseline FRCNN, whereas the *right* image contains the output of our proposed K-CNET. Ground-truth objects are marked orange, and correct detections (of IoU at least 0.75) in top 100 are marked blue.

5 Conclusion

In this paper, we study the problem of object detection in a novel knowledge-aware framework. Compared to existing algorithms which only focus on features within an image, we propose to leverage the vast amount of background knowledge. Towards this goal, we address the challenge of knowledge quantification that generalizes well on unseen co-occurrences, as well as the challenge of knowledge integration that aligns detections better with the background knowledge. Finally, we demonstrate the superior performance of our approach through extensive experiments.

References

[Auer *et al.*, 2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. DBpedia: A nucleus for a web of open data. In *ISWC-ASWC*, pages 722–735, 2007.

[Bell *et al.*, 2016] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross B. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, pages 2874–2883, 2016.

[Carlson *et al.*, 2010] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[Dai *et al.*, 2016] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016.

[Dong *et al.*, 2014] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *KDD*, pages 601–610, 2014.

[Everingham *et al.*, 2010] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[Fang *et al.*, 2013] Yuan Fang, Kevin Chen-Chuan Chang, and Hady Wirawan Lauw. Roundtriprank: Graph-based proximity with importance and specificity. In *ICDE*, pages 613–624, 2013.

[Felzenszwalb *et al.*, 2010] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010.

[Girshick *et al.*, 2014] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[Girshick, 2015] Ross B. Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Jeh and Widom, 2003] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *WWW*, pages 271–279, 2003.

[Krishna *et al.*, 2016] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, arXiv:1602.07332, 2016.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[Lenat, 1995] Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):32–38, 1995.

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

[Liu and Singh, 2004] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.

[Lu *et al.*, 2016] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *ECCV, Part I*, pages 852–869, 2016.

[Paulheim, 2017] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508, 2017.

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

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[Sermanet *et al.*, 2013] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, arXiv:1312.6229, 2013.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, arXiv:1409.1556, 2014.

[Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: a core of semantic knowledge. In *WWW*, pages 697–706, 2007.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[Tong *et al.*, 2006] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. In *ICDM*, pages 613–622, 2006.

[Vondrick *et al.*, 2016] Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, and Antonio Torralba. Predicting motivations of actions by leveraging text. In *CVPR*, pages 2997–3005, 2016.

[Wu *et al.*, 2016] Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, pages 4622–4630, 2016.

[Zhu *et al.*, 2013] Fanwei Zhu, Yuan Fang, Kevin Chen-Chuan Chang, and Jing Ying. Incremental and accuracy-aware personalized pagerank through scheduled approximation. *PVLDB*, 6(6):481–492, 2013.

[Zhu *et al.*, 2015] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base for visual question answering. *CoRR*, arXiv:1507.05670, 2015.