Leveraging Captions in the Wild to Improve Object Detection

Mert Kilickaya, Nazli Ikizler-Cinbis, Erkut Erdem and Aykut Erdem
Department of Computer Engineering
Hacettepe University
Beytepe, Ankara
kilickayamert@gmail.com
{nazli, erkut, aykut}@cs.hacettepe.edu.tr

Abstract

In this study, we explore whether the captions in the wild can boost the performance of object detection in images. Captions that accompany images usually provide significant information about the visual content of the image, making them an important resource for image understanding. However, captions in the wild are likely to include numerous types of noises which can hurt visual estimation. In this paper, we propose data-driven methods to deal with the noisy captions and utilize them to improve object detection. We show how a pre-trained state-of-the-art object detector can take advantage of noisy captions. Our experiments demonstrate that captions provide promising cues about the visual content of the images and can aid in improving object detection.

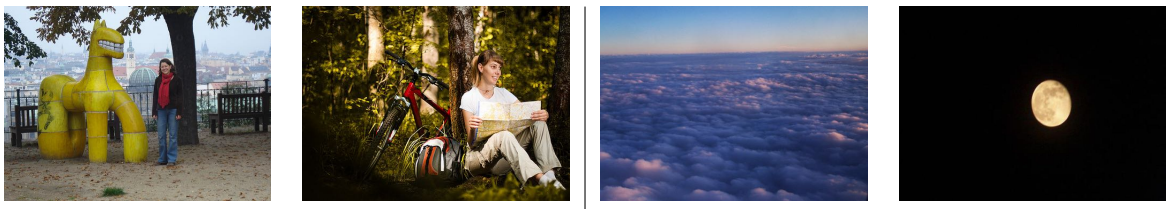
1 Introduction

Visual data on the Internet is generally coupled with descriptive text such as tags, keywords or captions. While tags and keywords are typically composed of single words, or phrases, and generally depict the main entities in an image (e.g. objects, places, etc.), a caption is a complete sentence which is intended to describe the image in a holistic manner. It can reveal information about not just the existing objects or the corresponding event but also the relationships between the objects/scene elements, their attributes or the actions in a scene (Figure 1). In this respect, captions provide a much richer source of information in order to understand the image content. This has recently motivated researchers to automate the task of describing images in natural languages using captions (Raffaella et al., 2016). However, most of these studies employ carefully collected image descriptions which are obtained by services like Amazon's Mechanical Turk (Rashtchian et al., 2010a; Hodosh et al., 2013; Young et al., 2014a; Lin et al., 2014). Little has been done on utilizing captions in the wild, i.e. the captions that accompany images readily available on the Web.

Although captions are rich, there are some challenges that limit their use in computer vision, and related language tasks. First, a caption may not be a visual depiction of the scene, but rather a sort of comment not directly related to the visual content of the image (Figure 1). The users might avoid explaining the obvious, but talk about more indirect aspects, abstract concepts and/or feelings. Third, the caption may be poorly written, which makes it difficult to understand the meaning of the text associated with the image.

On the other hand, there is also a major advantage in having image-caption pairs on the Web; billions of them are freely available online. Collectively considering image-caption pairs associated with a certain query image may allow to eliminate noisy information. Researchers have used this idea to collect a large scale images-captions dataset consisting of clean, descriptive texts paired with images (Chen et al., 2015). When noisy captions are eliminated, the rest can serve as an excellent source of information for what is available in the visual world.

In this paper, we investigate whether we can leverage captions in the wild to improve object detection. Object detection has seen some significant advances in recent years thanks to convolutional neural networks (LeCun et al., 2015). But in some cases, even state-of-the-art object detectors may fail to accurately locate objects or may produce false positives (see Figure 2). For such situations, we propose to utilize captions as an alternative source of information to determine what



the big yellow horse in Prague

beautiful young woman sitting near her bicycle under the tree in forest with map in her hands

above clouds airplane window (10)

Hey diddle diddle... the cat and the fiddle... the cow jumped over the M00N!!

Figure 1: Left: Examples of good captions, carrying rich information about the visual content of the image such as existence, sizes, attributes of objects, or their spatial organization. Right: Examples of noisy captions, where the mentioned objects may not exist visually (magenta for existing, red for non-existing objects).

is present in the image. Due to the reasons stated above, however, leveraging captions directly may result in errors. Therefore, we suggest to use data-driven methods which can eliminate the noise in the captions and inform about which objects are available in the image.

For our purpose, we first consider a constrained scenario where we assume access to test image captions and run detectors for objects mentioned in the caption, as previously motivated by (Ordonez et al., 2015). Then, we proceed to explore a more general setting where we observe captions only at training stage and infer possible objects within the test image using similar training images and their captions. In finding similar images/captions, we propose to use three different approaches, based on nearest neighbors, 2-view Canonical Correlation Analysis (CCA) and 3-view CCA. When the visual input is combined with caption information, these approaches not only help us to eliminate the noise in the captions, but also to infer about possible objects not even mentioned in the caption of a test image (see Figure 2). Our experimental results show that utilizing noisy captions of visually similar images in the proposed ways can indeed help in improving the performance of the object detection.

2 Related Work

In this section, we briefly review some of the relevant literature related to our problem.

2.1 Employing tags and captions to improve image parsing

Image parsing refers to the process of densely assigning a class label to each pixel in an image, which traditionally requires a large set of training images with pixel-level annotations. Similar to our goals here, some recent studies have focused on exploiting image tags (Xu et al., 2014) or sentences (Fidler et al., 2013; Cheng et al., 2014) associated with images to improve the performance by using objects or attributes exist in the images.

2.2 Weakly-supervised object localization

Another line of research close to ours is weakly-supervised object localization where the training set involves image-level labels which indicate the object classes present in the images. In addition to generic object detection approaches (e.g. (Pandey and Lazebnik, 2011; Siva and Xiang, 2011; Cinbis et al., 2016)), related studies also include face recognition with supervision from captions and script (Berg et al., 2004; Everingham et al., 2009).

2.3 Text-to-image co-referencing

Motivated from co-reference resolution tasks in NLP, a number of studies have investigated matching free-form phrases with images where the task is to locate each visual entity mentioned in a caption by predicting a bounding box in the corresponding image (Hodosh et al., 2010; Kong et al., 2014; Plummer et al., 2015; Rohrbach et al., 2015).

2.4 Automatic image captioning

Image captioning aims at automatically generating a description of a query image (Raffaella et al., 2016). As opposed to recent neural models, early image captioning methods mostly follow a grounded approach and generate descriptions by first detecting objects present in the images (Ordonez et al., 2015). The main drawback with this approach, however, is that object detectors may

Figure 2: Motivation. Given an image, Faster R-CNN detects the dog successfully however also produces many false positives (Left). A naive way to incorporate the caption would be to run detectors *only* mentioned in the caption of the image (Middle). This would also lead to false detections as the photographer did not mention the dog. In our approach, we leverage several captions to estimate the candidate objects in the image, in this case, the dog (Right).

produce many false positives and moreover, not all objects are important to be mentioned in the descriptions (Berg et al., 2012).

2.5 Detecting visual text

Lastly, a few works aim at detecting visual text, i.e., understanding whether an image caption contains visually relevant phrases or not (Dodge et al., 2012; Chen et al., 2013). Here, the approach in (Dodge et al., 2012) is especially quite related to our work because it involves the subtask of running several object detectors to infer what is present in the image using information from the captions.

3 Dataset

Recent datasets for language and vision research include natural images with natural language sentences. These sentences are either the photo captions generated by the users (aka. captions in the wild) (Ordonez et al., 2015; Chen et al., 2015; Thomee et al., 2016) or the descriptions collected via crowd-sourcing (Farhadi et al., 2010; Rashtchian et al., 2010b; Young et al., 2014b; Keller et al., 2014; Lin et al., 2014; Yatskar et al., 2014; Plummer et al., 2015). Although the datasets containing the crowd-sourced descriptions, namely Pascal Sentences (Farhadi et al., 2010), Visual and Linguistic Treebank (Keller et al., 2014), Flickr30K Entities (Plummer et al., 2015), Microsoft Research Dense Visual Annotation Corpus (Yatskar et al., 2014) and MS-COCO (Lin et al., 2014) datasets have extra object-level annotations, none of the datasets that consist of user-generated captions have these kind of information. Hence, in our work, we collected a new dataset of object-level annotated images that includes captions in the wild.

We built on our dataset named SBU-Objects from (Ordonez et al., 2015) which includes 1 million Flickr images and associated captions provided by the corresponding users. Although much effort has been made to eliminate noisy, non-visual captions, an important portion of these images have sentences that do not directly describe the visual content of these images. Figure 1 demonstrates such examples. The first example includes a caption mentioning an aeroplane, but it is mentioned only because the image is captured from the window of the airplane. The second example associates an image to a figurative caption that does not describe the visual content.

We restrict ourselves to the images containing captions where the object classes from the PASCAL challenge (Everingham et al., 2012) are mentioned such as dog, aeroplane, car, etc. To that end, we queried the dataset considering these PASCAL classes as well as their synonyms (e.g., motorbike, motorcycle). We also favoured image-caption pairs that include place prepositions such as *in, on, under, front* and *behind* coupled with the query noun (e.g., dog *under* the tree) if exist. This ensures the image-caption pairs to be used for exploring the effect of spatial information in captions and images as well. We observed that captions that are short (e.g., max 4 words) or in the form

Table 1: Corpus statistics. For each object class, we provide the number of instances in the dataset and their visibility rates $p(visible|mentioned)$.

Class	dog	bottle	chair	horse	cat	d. table	bird	cow	bike	sofa
# Instances	289	79	119	289	135	69	308	255	294	289
$p(visible\|mentioned)$	0.77	0.65	0.62	0.61	0.60	0.59	0.58	0.58	0.58	0.77
Class	sheep	boat	p. plant	m. bike	car	plane	monitor	bus	train	
# Instances	79	119	289	135	69	308	255	294	321	
$p(visible\|mentioned)$	0.65	0.62	0.61	0.60	0.59	0.58	0.58	0.58	0.18	

of phrases tend to be cleaner than longer captions. However, as our main aim is to leverage captions in the wild for object detection, we uniformly sampled captions that have different lengths between $[3 - 19]$ tokens, preventing the bias against caption lengths. We sampled 3.2k of such images for annotation and collected object-level bounding boxes for each and every PASCAL object available in the image. Table 1 shows the distribution of the number of object instances along with their visibility rates which is measured as the conditional probability given that a class name is mentioned in a caption, how frequent it actually exists in the image. As can be seen, animate objects like dogs, horses and cats appear frequently when mentioned while vehicles like aeroplane, bus and train have low visibilities.

4 Improving object detection with captions

In its simplest form, our aim is to determine candidate objects that can be detected from the image. Formally, given an image I_i, our aim is to estimate candidate object classes $C_i \in C$ visually present in the image, so that to eliminate false positives, only detectors of C_i are applied to the image. For simplicity, we assume that the set of possible object classes C is fixed, and the list of mentioned objects M_i is simply obtained from the captions via text-based search.

We begin with a simple, constrained scenario that assumes access to test image captions. Then, we proceed to explore more general setting where the captions are observed only at training.

4.1 Using pure captions

As stated previously, this simple model determines candidate objects directly from image's caption and hence, assumes that the caption of the image is given (at test time). This idea has previously been evaluated by (Ordonez et al., 2015) with a limited set of images for motivational purposes. Formally, given an image I_i, its caption T_i and the list of mentioned objects within that caption M_i, the candidate object classes is simply the list of mentioned objects, *ie.* $C_i = \{c_j, c_j \in M_i\}$.

This simple idea works surprisingly well, however, it restricts the search space for candidate objects C_i to the list of mentioned objects in the caption. The captions may be noisy, thus this procedure may suffer from typical issues stated previously; not all objects may be mentioned in the caption, and not all of the mentioned objects may be visible in the image.

4.2 Data-driven estimation of candidate objects

A more general setting is the case where we do not have access to the captions of newly seen images. Here, we describe three alternative data-driven methods for candidate object estimation.

4.2.1 Nearest-neighbor based estimation

For a given image, the captions of the visually similar images can be retrieved and utilized to identify potential object candidates. Our first method explores this approach, by directly measuring the similarity between images in visual feature space. With this setting, our aim is to see how well we can estimate candidate objects of a test image using uni-modal similarity.

To retrieve visually similar images, we need robust descriptors V that can represent the visual content effectively. To this end, we use two alternatives; first is the *fc-7* activations of VGG-19 (Simonyan and Zisserman, 2014) and second is *fc-7* activations of Hybrid model (Zhou et al., 2014). VGG-19 is a Convolutional Neural Network (CNN) model trained on ImageNet dataset which consists of 1000 different image classes (Russakovsky et al., 2015), while Hybrid is an CNN architecture that is trained on a combina-

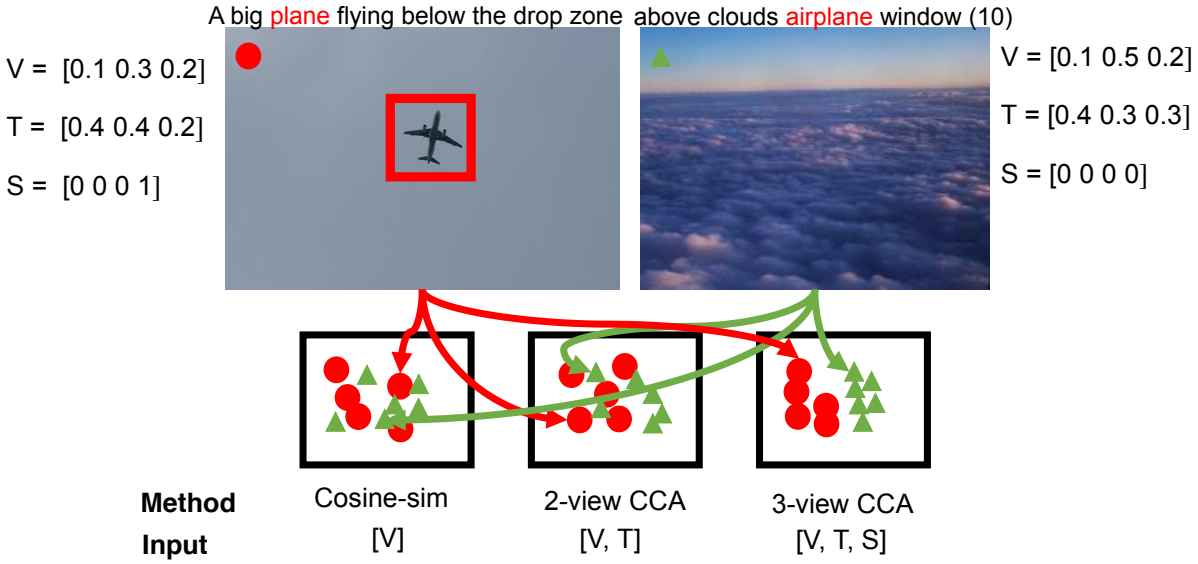


Figure 3: Here, three different embedding spaces are shown. Suppose red circle denotes the image on the left (and all images with aeroplanes visible) and green triangle denotes the image on the right (and all images with aeroplane missing). Nearest neighbor approach takes only visual representation of images V as input, thus these images may be considered similar. Projection gets better for 2-view CCA using $[V, T]$, however since they have similar textual representations, they still lie close in space. For 3-view CCA, with the inclusion of semantic category S, the embedding becomes distinguishable.

tion of Places (Zhou et al., 2014) (a large-scale scene recognition dataset) and ImageNet. Both architectures yield a $4096d$ representation per image. We use cosine-similarity between visual descriptors of each image and retrieve N nearest neighbors (images and their captions) per query. When measuring similarity, we also experimented with Euclidean distance, but found cosine distance to perform better for our purposes. After retrieving N neighbors, denoted as $NN(I_i)$, the candidate object classes for image I_i is the list of all objects in the captions of the neighbors $M_{NN(I_i)}$ that occur more than the mean frequency of the class occurrence counts. Formally, $C_i = \{c_j, c_j \in M_{NN(I_i)}, |c_j| \geq \tau\}$, where $\tau = \frac{1}{N}\sum_{c_j \in M_{NN(I_i)}} |c_j|$.

4.2.2 2-view CCA based estimation

Canonical Correlation Analysis (CCA) embedding (Hardoon et al., 2004) is an excellent tool for modelling data of different modalities, such as images I and their captions T (Hodosh et al., 2013). By using CCA, one can measure similarities (or differences) between different modalities in a common embedding space. Formally, CCA aims to minimize the following objective function:

$$\underset{W_1, W_2}{\text{minimize}} \quad \|(V_{train})W_1 - (L_{train})W_2\|_2^F \quad (1)$$

where W_1 and W_2 are visual and textual projection vectors and V_{train} and L_{train} are visual and textual representations of the training data, respectively. Here, for textual representation of captions, we use Fisher-encoded word2vec features (Klein et al., 2014; Plummer et al., 2015; Mikolov et al., 2013). Each word in a caption is first represented with a 300-D word2vec feature, then encoded within a Fisher Vector framework using 30 clusters. This results in a 18.000-D textual representation of each caption. Before projection, we reduce each modality's dimension to 1000-D for computational efficiency. Then, we learn the projection vectors using training data.

At test stage, we project the visual representation of a test image V_{test} to the common embedding space as $V_{projected} = V_{test}.W_1$ and measure similarity between projections of training images and the test image. Here, we again use the cosine similarity metric between projections. In our experiments, we use normalized-CCA as it yields better performance (Gong et al., 2014b) and normalize projections using corresponding eigenvectors. Similar to nearest-neighbour based candidate object estimation, we again retrieve N training images (and captions $M_{NN_{2CCA}(I_i)}$) on the common embedding space, and use the list of all object classes frequently occurring in the re-

trieved captions as C_i, *ie.* $C_i = \{c_j, c_j \in M_{NN_{2CCA}(I_i)}, |c_i| \geq \tau\}$.

4.2.3 3-view CCA based estimation

Our final retrieval strategy utilizes 3-view CCA embeddings. 3-view CCA, firstly proposed by (Gong et al., 2014a) is a generalized form of 2-view CCA by including a third view that correlates with the other views. In (Gong et al., 2014a), the authors propose 3-view CCA to achieve multi-modal retrieval between images and tags/keywords associated with images on the web. Third view can be seen as an additional supervision that guides visual and textual projections W_1-W_2 such that semantically related data are more accurately grouped. Formally, 3-view CCA solves the following minimization problem:

$$\begin{aligned}\underset{W_1,W_2,W_3}{\text{minimize}} &\|(V_{train})W_1 - (L_{train})W_2\|_2^F + \\ &\|(V_{train})W_1 - (S_{train})W_3\|_2^F + \\ &\|(L_{train})W_2 - (S_{train})W_3\|_2^F +\end{aligned}$$

where the first term is equal to the 2-view formulation and third view is induced by second and third terms, using S_{train} and W_3. S_{train} represents our third-view representation for the training set and W_3 is the corresponding projection matrix into embeddding space. Semantically, similar visual and textual representations should be projected to nearby locations and the semantic view S should be aligned with both V and L. For V and L, we use the same setting as in 2-view CCA.

In (Gong et al., 2014a), the authors use keyword or tag-derived textual representations for the third view. In our case, we use two alternatives:

- Class view from captions (denoted as S_T): Each class name is assigned a unique index $i \in [1, 19]$ and then convert it to a 16-bit binary $\in (0, 1)$. For each training image, we assign corresponding binary vector to annotated object's class(es). If more than one object is available, we apply bitwise OR operation to account for each object in the image.

- Visual view from annotated image regions (denoted as S_R): For each annotated object region in an image, we extract visual descriptors. Note that, the first view is extracted from the whole image, whereas this third view alternative uses visual information from individual regions. If there is more than one image annotation, we apply mean pooling.

Both alternatives try to assign images and captions with similar (candidate) objects to lie on close regions in the embedding space. Similar to nearest-neighbor and 2-view CCA, we retrieve N most similar images and corresponding captions for each test image to form the set of candidate object classes.

Figure 3 illustrates an example for the intuition behind using the third view. Suppose there are two images where each caption includes the *aeroplane* class. Although one of the images really shows an image of an aeroplane, the other is captured from an aeroplane window, so no aeroplane is seen. Both their textual representations T include aeroplane, whereas their visual representations V and semantic representations S differ significantly. Using both of these views, these images project into farther points in the embedding space compared to the naive cosine-similarity space and 2-view CCA embeddings, thus can easily be distinguished.

5 Experiments

For experiments, we split our dataset as 50%-50% as training and test. We use Faster R-CNN (Ren et al., 2015) as our base object detector. The detector itself is trained on the PASCAL VOC 2012 data (Everingham et al., 2012). We emply PASCAL (Everingham et al., 2012) conventions while evaluating the methods and also set the set of possible object classes C to Pascal classes (excluding the person class, due to the high level of ambiguity of the captions of this class), so we have 19 classes in total. Following the regular detection experimental settings, we measure intersection-over-union (IoU) between detection and annotation windows and count the detections as positive detections if their IoU exceeds the threshold 0.50%. We evaluate the performance using average precision (AP). While selecting the similar images, the number of nearest images N is assigned to different values of $(10, 20, 50, 100, 150)$.

The first experiments evaluate the performance of Faster R-CNN by running the detector for every object class without considering any textual information, referred as *All classes*. In the second experiment, we assume that we have access to the captions of the test images and run the detector only for the objects mentioned in these captions. This experiment can be interpreted as using an unreliable oracle, since the objects mentioned

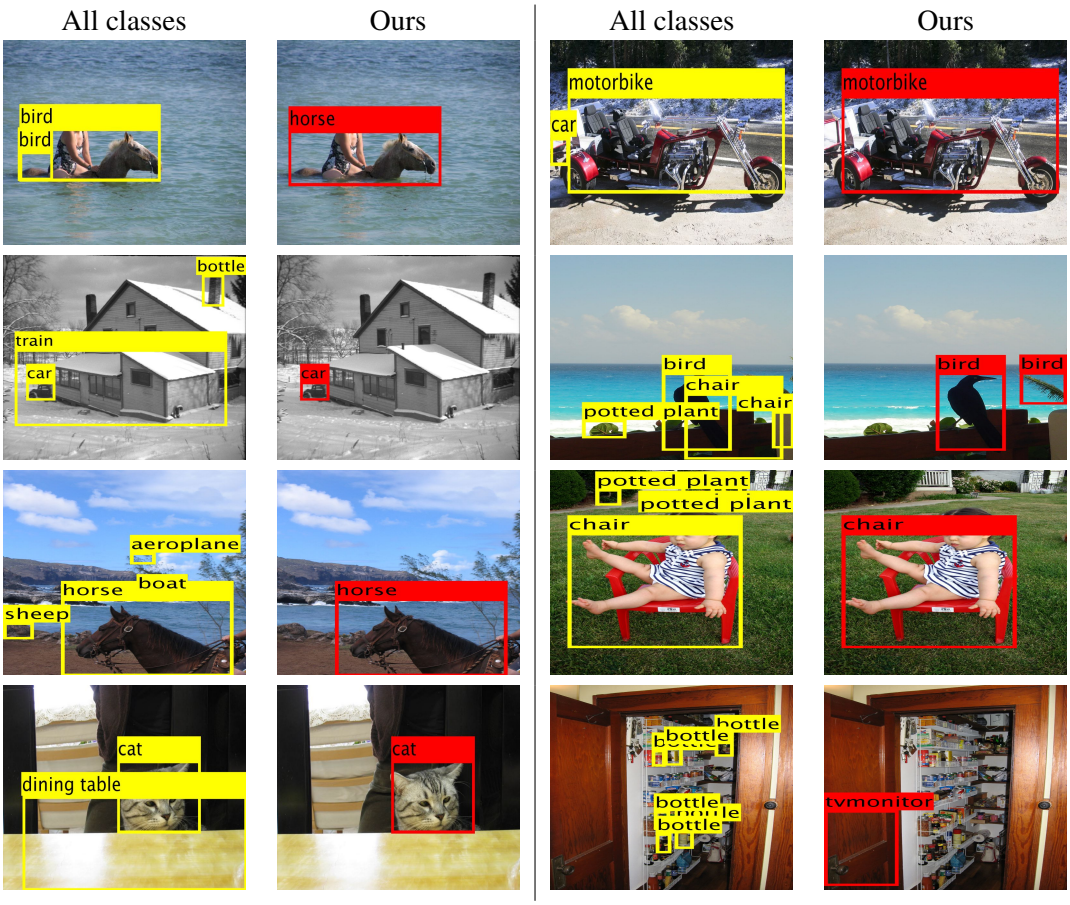


Figure 4: Example detection results illustrating the performance improvements using caption information. The last row shows two failure cases.

in the text do not need to exist in the images as discussed before. We refer to this method as *Mentioned classes*. The quantitative results of these experiments are given in Table 4. As can be seen, based detector results are quite inferior, compared to the case when the list of objects are limited to the set of objects in the given image captions.

The third set of experiments consider a more general setup, where we do not have access to captions of newly seen images, and assess the performance of data-driven estimation of object classes from similar images. In particular, we run the detector for only those candidate object classes that are gathered by retrieving the N closest images and using the frequent object classes mentioned in the retrieved captions. Here, we consider three different approaches. Firstly, we consider only visual similarities of VGG (Simonyan and Zisserman, 2014) and Hybrid (Zhou et al., 2014) activations of the test and training images as described in Sec.4.2.1. In the second and third approaches, we use the embedding spaces learned via the 2-view and 3-view CCA as introduced in Sec.4.2.2 and Sec.4.2.3, respectively.

Table 2 shows the results of our object detection schemes which consider data-driven approaches to limit the object detectors. In general, we observe that VGG activations as deep features yield better results than HYBRID activations. As the number of closest images increase, we are able to predict the candidate object classes more accurately, and obtain better performances for all retrieval scenarios. In general, 3-view CCA gives the best results over the other alternatives.

In Table 3, we show the object detection results for different choices of the third view for 3-way CCA. As demonstrated, the region-based deep activations result in a better embedding space than the binary class vectors, providing more accurate object detection results.

Finally, we compare the results of all of our experiments. As can be seen in Table 4 and Figure 4, Faster R-CNN produces many false positive when run with all the object classes. When it is run with the classes mentioned in the given caption, the accuracy improves as expected. Interestingly,

Table 2: Mean Average precision (mAP) values for detection through data-driven estimation of object classes. Each approach is tested by retrieving $N = (10, 20, 50, 100, 150)$ similar images. For 3-view CCA, the binary class (S_T) is used as the third view.

Deep image feature	Method	10	20	50	100	150
Hybrid	Single view	0.385	0.428	0.473	0.495	0.504
	2-view CCA	0.396	0.421	0.480	0.492	0.504
	3-view CCA	0.399	0.425	**0.487**	0.501	0.499
VGG	Single view	0.403	0.432	0.479	0.492	0.499
	2-view CCA	0.413	0.443	0.484	**0.511**	0.512
	3-view CCA	**0.416**	**0.451**	0.486	0.508	**0.515**

Table 3: Mean Average precision (mAP) values for detection using the embedding spaces learned through 3-view CCA using binary class vectors (S_T) or deep visual feature averaged over annotated object regions (S_R) as the third views. V and S represents our first and third view choices respectively.

V	S	Method	10	20	50	100	150
Hybrid	Binary class (S_T)	3-view CCA	0.399	0.425	0.487	0.501	0.499
	Region features (S_R)	3-view CCA	0.418	**0.455**	0.496	**0.511**	0.517
VGG	Binary class (S_T)	3-view CCA	0.416	0.451	0.486	0.508	0.515
	Region features (S_R)	3-view CCA	**0.419**	0.448	**0.503**	0.508	**0.518**

Table 4: Mean Average precision (mAP) values of the Faster R-CNN run with all classes, classes mentioned in the captions and the predicted candidate object classes.

			Predicted classes		
Method	All classes	Mentioned classes	Single view	2-view CCA	3-view CCA
AP	0.304	0.508	0.504	0.512	**0.518**

our multi-view prediction approaches give highly competitive and even better results than using the captions of the test images.

6 Conclusion

In this paper, we develop methods to improve performance of object detection using captions in the wild. Captions are freely available textual image descriptions written by the users, exhibiting a high range of challenges due to excessive noise. To overcome these limitations, we develop data-driven methods that can achieve better performance than the current state-of-the-art object detector Faster R-CNN by means of estimating likely objects in the images. We compare different strategies that use different levels of supervision. We show that superior results can be obtained even without access to image's *own* caption, by leveraging (somewhat noisy) captions of similar images. The results clearly indicate that captions are beneficial supervisory signals for object detection problem, when used in a data-driven manner.

In the future, we plan to extend our dataset using larger-scale image-caption pairs datasets such as Flickr-100M (Thomee et al., 2016). We also plan to apply similar ideas to co-localization problem (Tang et al., 2014) where noisy images can also be determined by data-driven methods.

7 Acknowledgements

This research was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK), Career Development Award 113E116.

References

T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D. A. Forsyth. 2004. Names and faces in the news. In *CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–848–II–854 Vol.2.

A. C. Berg, T. L. Berg, H. Daum, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood,

K. Stratos, and K. Yamaguchi. 2012. Understanding and predicting importance in images. In *CVPR*, pages 3562–3569.

T. Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. 2013. Understanding and classifying image tweets. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 781–784. ACM.

J. Chen, Polina Kuznetsova, David Warren, and Yejin Choi. 2015. Déja image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 NAACL: Human Language Technologies*, pages 504–514.

Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J. Mitra, and Philip Torr. 2014. Imagespirit: Verbal guided image parsing. *ACM Trans. Graph.*, 34(1):3:1–3:11, December.

R. G. Cinbis, J. Verbeek, and C. Schmid. 2016. Weakly supervised object localization with multi-fold multiple instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pages 1–1.

J. Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C Berg, et al. 2012. Detecting visual text. In *Proceedings of the 2012 NAACL: Human Language Technologies*, pages 762–772. Association for Computational Linguistics.

M. Everingham, J. Sivic, and A. Zisserman. 2009. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5).

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org.

A. Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer.

S. Fidler, A. Sharma, and R. Urtasun. 2013. A sentence is worth a thousand pixels. In *CVPR, 2013 IEEE Conference on*, pages 1995–2002.

Y. Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014a. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233.

Y. Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014b. Improving image-sentence embeddings using large weakly annotated photo collections. In *Computer Vision–ECCV 2014*, pages 529–545. Springer.

D. R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

M. Hodosh, Peter Young, Cyrus Rashtchian, and Julia Hockenmaier. 2010. Cross-caption coreference resolution for automatic image understanding. In *Proceedings of the Fourteenth Conference on CoNLL 2010, Uppsala, Sweden, July 15-16, 2010*, pages 162–171.

M. Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR*, 47:853–899.

F. Keller, Desmond Elliott, et al. 2014. Visual and linguistic treebank.

B. Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*.

C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. 2014. What are you talking about? text-to-image coreference. In *2014 IEEE CVPR*, pages 3558–3565.

Y. LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

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

V. Ordonez, Xufeng Han, Polina Kuznetsova, Girish Kulkarni, Margaret Mitchell, Kota Yamaguchi, Karl Stratos, Amit Goyal, Jesse Dodge, Alyssa Mensch, et al. 2015. Large scale retrieval and generation of image descriptions. *IJCV*, pages 1–14.

M. Pandey and S. Lazebnik. 2011. Scene recognition and weakly supervised object localization with deformable part-based models. In *2011 ICCV*, pages 1307–1314.

B. A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE ICCV*, pages 2641–2649.

B. Raffaella, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *JAIR*, 55:409–442.

C. Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010a. Collecting image annotations using amazon's mechanical turk. In *NAACL: Human Language Technologies Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

C. Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010b. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

S. Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

A. Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2015. Grounding of textual phrases in images by reconstruction. *arXiv:1511.03745*.

O. Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252.

K. Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

P. Siva and Tao Xiang. 2011. Weakly supervised object detector learning with model drift detection. In *2011 ICCV*, pages 343–350.

K. Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. 2014. Co-localization in real-world images. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1464–1471. IEEE.

B. Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

J. Xu, A. G. Schwing, and R. Urtasun. 2014. Tell me what you see and i will show you where it is. In *2014 IEEE CVPR*.

M. Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Third Joint Conference on Lexical and Computation Semantics (*SEM)*.

P. Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014a. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the ACL*, 2:67–78.

P. Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014b. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the ACL*, 2:67–78.

B. Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495.