

Cap2Det: Learning to Amplify Weak Caption Supervision for Object Detection

Keren Ye^{*1}, Mingda Zhang¹, Adriana Kovashka¹, Wei Li^{†2}, Danfeng Qin³, and Jesse Berent³

¹Department of Computer Science, University of Pittsburgh, Pittsburgh PA, USA

²Facebook, Inc. Seattle WA, USA

³Google Research, Zurich, Switzerland

{yekeren, mzhang, kovashka}@cs.pitt.edu lwthuucs@gmail.com {qind, jberent}@google.com

Abstract

Learning to localize and name object instances is a fundamental problem in vision, but state-of-the-art approaches rely on expensive bounding box supervision. While weakly supervised detection (WSOD) methods relax the need for boxes to that of image-level annotations, even cheaper supervision is naturally available in the form of unstructured textual descriptions that users may freely provide when uploading image content. However, straightforward approaches to using such data for WSOD wastefully discard captions that do not exactly match object names. Instead, we show how to squeeze the most information out of these captions by training a text-only classifier that generalizes beyond dataset boundaries. Our discovery provides an opportunity for learning detection models from noisy but more abundant and freely-available caption data. We also validate our model on three classic object detection benchmarks and achieve state-of-the-art WSOD performance. Our code is available at <https://github.com/yekeren/Cap2Det>.

1. Introduction

Learning to localize and classify visual is a fundamental problem in computer vision. It has a wide range of applications, including robotics, autonomous vehicles, intelligent video surveillance, and augmented reality. Since the renaissance of deep neural networks, object detection has been revolutionized by a series of groundbreaking works, including Faster-RCNN [26], Mask-RCNN [13] and YOLO [25]. Modern detectors can run in real-time on mobile devices, and have become the driving force for future technologies.

Despite these achievements, most modern detectors suf-

^{*}Work partially done during an internship at Google

[†]Work done at Google Research, Zurich, Switzerland

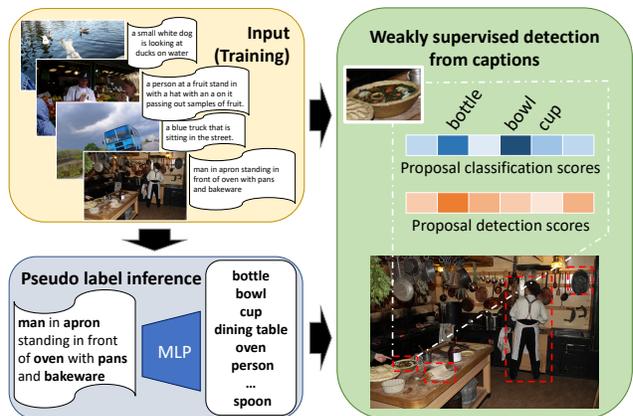


Figure 1: **An overview of our approach.** We leverage a novel source of weak supervision for object detection, namely natural language captions, by amplifying the signal these captions provide using a textual classifier. Even though “bottle”, “bowl”, “cup” and “spoon” are not mentioned in the caption, they are actually in the image, and our method correctly infers these image-level labels. We then train a model that scores object proposals in iterative fashion, to localize objects as its final output.

fer from an important limitation: they are trained with expensive supervision in the form of large quantities of bounding boxes meticulously drawn by a large pool of human annotators. Due to the well-known domain shift problem [8, 33, 19, 29, 12] and imperfect domain adaptation techniques, this means when detection is to be performed in a novel domain, the expensive annotation procedure needs to be repeated. Weakly supervised object detection (WSOD) techniques aim to alleviate the burden of collecting such expensive box annotations. The classic WSOD problem formulation [4, 37, 31, 30] treats an image as a bag of proposals, and learns to assign instance-level semantics to these proposals using multiple instance learning (MIL). WSOD

has shown great potential for object detection, and the state-of-the-art model has reached 40% mAP [30] on Pascal VOC 2012. However, one critical assumption in WSOD is that the image-level label should be *precise*, indicating at least one proposal instance in the image needs to associate with the label. This assumption does not always hold especially for real-world problems and real-world supervision.

Weakly supervised detection methods need large-scale image-level object category labels. These labels require human effort that is provided in an unnatural, crowdsourced environment. However, more natural supervision for objects exists—for example, in the form of natural language descriptions that web users provide when uploading their photos to social media sites such as YouTube or Instagram. There are tens of millions of photos uploaded to Instagram every day, and a majority of them have titles, tags, or descriptions. Abundant videos with subtitles are similarly available on YouTube. These annotations are “free” in that no user was *paid* to provide them; they arise out of innate needs of users to make their content available to others.

However, existing WSOD methods cannot use such supervision. First, these natural language descriptions are unstructured; they need to be parsed and words relevant for object recognition need to be extracted, while non-object words are removed. Second, these descriptions are both imprecise and non-exhaustive—they might mention content that is not in the image (e.g. what event the user was attending or who they met after the photo was taken), and also omit content that actually is in the image but is not interesting. Consider the image in the bottom-right of Fig. 1. It contains numerous objects, many of which fairly large—e.g. dining table and bowls—yet the human providing the description did not mention these objects. Thus, directly feeding web data to the state-of-the-art WSOD system contains numerous limitations—at the very least, it under-utilizes the rich supervision that captions can provide.

To address this issue, we build an object detector from images paired with accompanying captions (sentences). Our model bridges human-written free-form texts and visual objects, and generates accurate bounding boxes over objects in an image. Our method relies on two key components. First, we train a textual classifier to map captions to discrete object labels. This classifier is not dataset-dependent, requires only a small set of labels, and generalizes beyond dataset boundaries. It enables us to bridge the gap between what humans mention in a caption, and what truly is in an image. Second, we use the pseudo ground truth labels predicted by this textual classifier, to train a weakly supervised object detection method. The method we propose extracts region proposals off-the-shelf, then for each proposal and each class, learns both a class score and a detection score. These scores are then refined using an iterative approach, to produce final detection results.

The contributions of our paper are three-fold. First, we propose a new task of learning from noisy caption annotations, and set up a new benchmark. Rather than treating object categories as IDs only, we also leverage their semantics, as well as synonyms of those object names. Second, on this benchmark, we show that we outperform alternative uses of captions, e.g. exactly matching the captions to object category words, or retrieving hand-annotated or predicted synonyms to the object categories from the captions. We show competitive WSOD performance by learning on COCO or Flickr30K captions. We further validate the benefit of our COCO-trained text classifier by applying it on Flickr30K, and leveraging training on Flickr30K then evaluating on PASCAL. Finally, in a side-by-side comparison under the classic WSOD setting, our model demonstrates superior performance with image-level supervision and achieves state-of-the-art performance on all three WSOD benchmarks (48.5% mAP@0.5 for VOC 2007, 45.1% mAP@0.5 on VOC 2012 and 23.4% mAP@0.5 on COCO).

The remainder of the paper is organized as follows. We overview related work in Sec. 2. In Sec. 3.1, we discuss different ways to reduce the gap between free-form captions and object categories. In Sec. 3.2, we describe the backbone of our WSOD model, which combines prior work [31, 28] in a new way. In Sec. 4, we compare our method to state-of-the-art and alternative methods. We conclude in Sec. 5.

2. Related Work

We formulate a new variant of weakly supervised object detection, where the supervision is even more weak than in prior work. We leverage captions, so we also discuss work that finds alignment between image regions and text.

Weakly supervised detection via MIL. Most WSOD methods formulate the task as a multiple instance learning (MIL) problem. In this problem, proposals of an image are treated as a bag of candidate instances. If the image is labeled as containing an object, at least one of the proposals will be responsible to provide the prediction of that object. [23, 41] propose a Global Average (Max) Pooling layer to learn class activation maps. [4] propose Weakly Supervised Deep Detection Networks (WSDDN) containing classification and detection data streams, where the detection stream weighs the results of the classification predictions. [17] improve WSDDN by considering context. [31, 30] jointly train multiple refining models together with WSDDN, and show the final model benefits from the online iterative refinement. [10, 37] apply a segmentation map and [37] incorporate saliency. Finally, [36] adds a min-entropy loss to reduce the randomness of the detection results. Our work is similar to these since we also represent the proposals using a MIL weighted representation, however, we go one step further to successfully adopt a more advanced neural architecture, and a more challenging supervision scenario.

shown) mention the “bowls” or “spoons” that are present, and only some mention the “oven”.

This observation inspires us to amplify the supervision signal that captions provide, and squeeze more information out of them. Fig. 2 (bottom) shows the approach we use to amplify the signal. This text-only model takes free-form texts as input, embeds individual words to a 300D space using GloVe [24], and projects the embedded features to a 400D latent space. We then use max-pooling to aggregate the word-level representations. Then, we use this intermediate representation to predict the implied instances (e.g. 80 classes as defined in COCO, or any other categories); this prediction answers “what’s in the image” and serves as pseudo image-level labels in training object detectors.

It is worth noting that there exists a subtle balance when using pseudo labels to train object detectors. Admittedly, our strategy increases the recall rates thus more data could be utilized. However, with the increased recall, precision will drop inevitably thus the fundamental assumption in MIL is threatened. Specifically, the *precise label* assumption makes the model very sensitive to false positive cases: when inappropriate labels are given where none of the proposals have a good response, the model gets confused, resulting in non-optimal detections.

We finally adapt a two-steps procedure: first we look for an exact match of object labels from captions, following the intuition that *explicitly mentioned objects* should be significant and obvious enough in the image; second, when no object can be matched, we use our label inference model to predict labels as *unspoken intended objects* to guide the object detection. We show our method outperforms several strong alternatives that also infer pseudo labels.

Discussion. Our text classifier relies on both captions and category labels. However, once the bridge between captions and labels is established, this classifier generalizes to other datasets, as we show in Tab. 1. Importantly, we only need a small fraction of labels to train this text classifier; as we show in Fig. 3, precision ranges between 89% and 92% when we use between only 5% and 100% of the COCO data, while recall is stable at 62%. Thus, our text model could learn from a *single source* dataset with *a few* labels, then it could transfer the knowledge to other *target* datasets, requiring only free-form text as supervision.

3.2. Detection from inferred labels

We next describe how we use the inferred pseudo labels to train an object detection model. As shown in Fig. 2, we first extract proposals with accompanying features. An image is fed into the pretrained (on ImageNet [9]) convolutional layers. Then, *ROIAlign* [13] is used for cropping the proposals (at most 500 boxes per image) generated by *Selective Search* [35], resulting in fixed-sized convolutional feature maps. Finally, a box feature extractor is ap-

plied to extract a fixed-length feature for each proposal. If $[r_1, \dots, r_m]$ are the proposals of a given image x , this process results in proposal feature vectors $[\phi(r_1), \dots, \phi(r_m)]$ where each $\phi(r_i) \in \mathbb{R}^d$. Note that while our model is pretrained on ImageNet, it *does not leverage* any image labels at all on the datasets on which we train and evaluate our detection models (PASCAL and COCO).

3.2.1 Weakly supervised detection

We next introduce the prediction of image-level labels \hat{p}_c ($c \in \{1, \dots, C\}$, where C is the number of classes) and of detection scores as a by-product. The proposal features $\phi(r_i)$ are fed into two parallel fully-connected layers to compute the detection scores $o_{i,c}^{\text{det}} \in \mathbb{R}^1$ (top branch in the green MIL module in Fig. 2) and classification scores $o_{i,c}^{\text{cls}} \in \mathbb{R}^1$ (bottom branch), in which both scores are related to a specific class c and the particular proposal r_i :

$$o_{i,c}^{\text{cls}} = w_c^{\text{cls}\top} \phi(r_i) + b_c^{\text{cls}}, \quad o_{i,c}^{\text{det}} = w_c^{\text{det}\top} \phi(r_i) + b_c^{\text{det}} \quad (1)$$

We convert these scores into: (1) $p_{i,c}^{\text{cls}}$, the probability that object c presents in proposal r_i ; and (2) $p_{i,c}^{\text{det}}$, the probability that r_i is important for predicting image-level label y_c :

$$p_{i,c}^{\text{cls}} = \sigma(o_{i,c}^{\text{cls}}), \quad p_{i,c}^{\text{det}} = \frac{\exp(o_{i,c}^{\text{det}})}{\sum_{j=1}^m \exp(o_{j,c}^{\text{det}})} \quad (2)$$

Finally, the aggregated image-level prediction is computed as follows, where greater values of $\hat{p}_c \in [0, 1]$ mean higher likelihood that c is present in the image:

$$\hat{p}_c = \sigma \left(\sum_{i=1}^m p_{i,c}^{\text{det}} o_{i,c}^{\text{cls}} \right) \quad (3)$$

Assuming the label $y_c = 1$ if and only if class c is present, the **multiple instance detection** loss used for training the model is defined as:

$$L_{\text{mid}} = - \sum_{c=1}^C \left[y_c \log \hat{p}_c + (1 - y_c) \log(1 - \hat{p}_c) \right] \quad (4)$$

Preliminary detection scores. The weakly supervised detection score given both proposal r_i and class c is the product of $p_{i,c}^{\text{cls}}$ and $p_{i,c}^{\text{det}}$ which is further refined as described in Sec. 3.2.2.

3.2.2 Online instance classifier refinement

The third component of our WSOD model is Online Instance Classifier Refinement (OICR), as proposed by Tang *et al.* [31]. The main idea behind OICR is simple: Given a ground-truth class label, the top-scoring proposal, as well as proposals highly overlapping with it, are selected as references. These proposals are treated as positive examples for training the box classifier of this class while others are treated as negatives. The initial top-scoring proposal may only partially cover the object, so allowing highly-overlapped proposals to be treated as positives gives them a

second chance to be considered as containing an object, in the subsequent model refinement. This reduces the chance of propagating incorrect predictions. In addition, sharing the convolutional features between the original and refining models makes the training more robust.

Following [31], we stack multiple refining classifiers and use the output of the previous one to generate instance-level supervision to train the successor. The detection score at the 0-th iteration is computed using $s_{i,c}^{(0)} = p_{i,c}^{cls, det}$, $s_{i,C+1}^{(0)} = 0$ (where $C + 1$ is the background class). Given the detection score $s_{i,c}^{(k)}$ at the k -th iteration, we use the image-level label to get the *instance-level* supervision $y_{i,c}^{(k+1)}$ at the $(k + 1)$ -th iteration. Assume that c' is a label attached to image x , we first look for the top-scoring box r_j ($j = \arg \max_i s_{i,c'}^{(k)}$).

We then let $y_{i,c'}^{(k+1)} = 1, \forall i \in \{l | IoU(r_l, r_j) > threshold\}$. When $k > 0$, $s_{i,c}^{(k)}$ is inferred using a $(C + 1)$ -way FC layer, as in Eq. 1. The OICR training loss is defined in Eq. 5.

$$L_{\text{oicr}}^k = -\frac{1}{m} \sum_{i=1}^m \sum_{c=1}^{C+1} \hat{y}_{i,c}^{(k)} \log s_{i,c}^{(k)}, \quad k = 1, \dots, K \quad (5)$$

Unlike the original OICR, our WSOD module aggregates logits instead of probability scores, which in our experience stabilizes training. We also removed the reweighing of untrustworthy signals emphasized in [31] since we found it did not contribute significantly.

The final loss we optimize is Eq. 6. We refine our model for 3 times ($K = 3$) if not mentioned otherwise.

$$L = L_{\text{mid}} + \sum_{k=1}^K L_{\text{oicr}}^k \quad (6)$$

4. Experiments

We evaluate all components of our method: the text classifier that learns to map captions to object labels, the weakly supervised detection module, and the refinement. We show that compared to alternative strategies, our approach extracts the most accurate and expansive information from the captions (Sec. 4.2). By training on COCO captions, we achieve close to state-of-the-art results on weakly supervised detection on PASCAL, even though the supervision we leverage is weaker than competitor methods. Importantly, our text classifier allows us to excel at the task of training on Flickr30K to detect on PASCAL, even though that classifier was trained on a different dataset (COCO). Finally, we show our approach outperforms prior methods on the task of learning from image-level labels (Sec. 4.3).

4.1. Implementation details

Before training the detector, we use Selective Search [35] from OpenCV [5] to extract at most 500 proposals for each image. We follow the ‘‘Selective search quality’’ parameter settings in [35]. We prefer Selective Search because

it is a generic, dataset-independent proposal generation procedure, as opposed to other CNN-based alternatives which are trained end-to-end from a specific dataset in a supervised fashion. We also experimented with Edge Boxes [42] but got inferior performance. We use TensorFlow [1] as our training framework. To compute the proposal feature vectors, we use the layers (‘‘Conv2d_1a_7x7’’ to ‘‘Mixed_4e’’) from Inception-V2 [28] to get the conv feature map, and the layers (‘‘Mixed_5a’’ to ‘‘Mixed_5c’’) from the same model to extract the proposal feature vectors after the ROIAlign [13] operation. The Inception-V2 model is pretrained on ImageNet [9]; the supervised detector counterpart of our model, using this architecture, was explored by [15]. To augment the training data, we resize the image randomly to one of the four scales $\{400, 600, 800, 1200\}$. We also randomly flip the image left to right at training time. At test time, we average the proposal scores from the different resolution inputs. We set the number of refinements to 3 for the OICR since it gives the best performance. For post-processing, we use non-maximum-suppression with IoU threshold of 0.4. We use the AdaGrad optimizer, a learning rate of 0.01, and a batch size of 2 as commonly used in WSOD methods [31, 30]. The models are usually trained for 100K iterations on Pascal VOC (roughly 40 epochs on VOC2007 and 17 epochs on VOC2012) and 500K on COCO (8.5 epochs), using a validation set to pick the best model. Our implementation is available at <https://github.com/yekeren/Cap2Det>.

4.2. Using captions as supervision

In this section, we evaluate our full method, including our proposal for how to squeeze the most information out of the weak supervision that captions provide (Sec. 3.1). We also experiment with alternative strategies of generating pseudo labels, and evaluate the performance in terms of precision and recall by comparing with ground-truth labels.

Alternative strategies. We compared with multiple pseudo-label generation baselines when lexical matching (EXACTMATCH) fails to find a match. As previous examples show, considering synonyms can effectively reduce off-target matching rates. Thus our first baseline adopts a *manually constructed, hence expensive* COCO synonym vocabulary list (EXTENDVOCAB) which maps 413 words to 80 categories [21]. Another variant, GLOVEPSEUDO, takes advantage of GloVe word embeddings [24], assigning pseudo-labels for a sentence by looking for the category that has the smallest embedding distance to any word in the sentence. We also follow a similar strategy with [38] to finetune the GloVe word embeddings on COCO using a visual-text ranking loss, and use the pseudo labels retrieved by the resultant LEARNEDGLOVE as a stronger baseline. The final reference model of using ground-truth image-level labels GT-LABEL is an up-

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
Training on different datasets using ground-truth labels:																					
GT-LABEL VOC	68.7	49.7	53.3	27.6	14.1	64.3	58.1	76.0	23.6	59.8	50.7	57.4	48.1	63.0	15.5	18.4	49.7	55.0	48.4	67.8	48.5
GT-LABEL COCO	65.3	50.3	53.2	25.3	16.2	68.0	54.8	65.5	20.7	62.5	51.6	45.6	48.6	62.3	7.2	24.6	49.6	34.6	51.1	69.3	46.3
Training on COCO dataset using captions:																					
EXACTMATCH (EM)	63.0	50.3	50.7	25.9	14.1	64.5	50.8	33.4	17.2	49.0	48.2	46.7	44.2	59.2	10.4	14.3	49.8	37.7	21.5	47.6	39.9
EM + GLOVEPSEUDO	66.6	43.7	53.3	29.4	13.6	65.3	51.6	33.7	15.6	50.7	46.6	45.4	47.6	62.1	8.0	15.7	48.6	46.3	30.6	36.4	40.5
EM + LEARNEDGLOVE	64.1	49.9	58.6	24.9	13.2	66.9	49.2	26.9	13.1	57.7	52.8	42.6	53.2	58.6	14.3	15.0	45.2	50.3	34.1	43.5	41.7
EM + EXTENDVOCAB	65.0	44.9	49.2	30.6	13.6	64.1	50.8	28.0	17.8	59.8	45.5	56.1	49.4	59.1	16.8	15.2	51.1	57.8	14.0	61.8	42.5
EM + TEXTCLSF	63.8	42.6	50.4	29.9	12.1	61.2	46.1	41.6	16.6	61.2	48.3	55.1	51.5	59.7	16.9	15.2	50.5	53.2	38.2	48.2	43.1
Training on Flickr30K dataset using captions:																					
EXACTMATCH (EM)	46.6	42.9	42.0	9.6	7.7	31.6	44.8	53.2	13.1	28.0	39.1	43.2	31.9	52.5	4.0	5.1	38.0	28.7	15.8	41.1	31.0
EM + EXTENDVOCAB	37.8	37.6	35.5	11.0	10.3	18.0	47.9	51.3	17.7	25.5	37.0	47.9	35.2	46.1	15.2	0.8	27.8	35.6	5.8	42.0	29.3
EM + TEXTCLSF	24.1	38.8	44.5	13.3	6.2	38.9	49.9	60.4	12.4	47.4	39.2	59.3	34.8	48.1	10.7	0.3	42.4	39.4	14.1	47.3	33.6

Table 1: Average precision (in %) on the VOC 2007 test set (learning from COCO and Flickr30K captions). We learn the detection model from the COCO captions describing the 80 objects, but evaluate on only the overlapping 20 VOC objects.

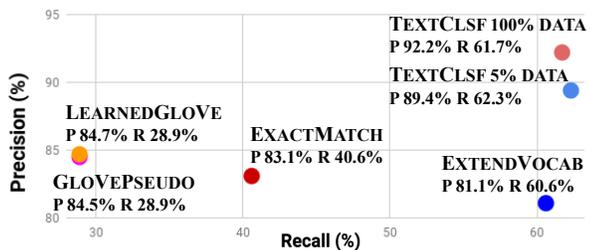


Figure 3: Analysis of different text supervision. We compare the pseudo labels (Sec. 3.1) to COCO *val* ground-truth.

per bound. Note that apart from the strategy used to mine image-level labels, these strategies all use the same architecture and WSOD approach as our method (Sec. 3.2). In later sections, we show combinations of the exact match strategy with these methods (when exact match fails), resulting in EM+GLOVEPSEUDO, EM+LEARNEDGLOVE, EM+EXTENDVOCAB and EM+TEXTCLSF. We examine how well these and other strategies leverage captions from COCO and Flickr30K [39] to produce accurate detection.

Analysis of textual supervision. In Fig. 3 we show the *precision* and *recall* of these label inference methods evaluated directly on the COCO image-level labels (5,000 examples of the *val2017* set). We observe that EXTENDVOCAB, which uses the hand-crafted word-synonyms dictionary, provides the best recall (60.6%) among all methods but provides the worst precision of 81.1%. The word-embedding-based top-scoring matching methods of GLOVEPSEUDO and LEARNEDGLOVE provide precise predictions (84.5% and 84.7% respectively, which are the highest). However, our TEXTCLSF achieves significantly improved precision compared to these. We would like to point out that while in Tab. 1 and 2, our method uses the full COCO training set (118,287 concatenated captions), it achieves very similar performance with even a small fraction of the data. With 5% of the data, the method achieves 89% precision (vs 92%

precision with 100% of the data), both of which are much higher than any other baselines; recall is about 62% for both 5% and 100% training data. In other words, it is sufficient to use a small portion of precise text labels to train a generalizable label inference classifier, and the knowledge can transfer to other datasets as we show in Tab. 1.

To better understand the generated labels, we show two qualitative examples in Fig. 4. The image on the right shows that our model infers “tie” from the observation of “presenter”, “conference” and “suit”, while all other methods fail to extract this object category for visual detection. We argue the capability of inferring reasonable labels from captions is critical for learning detection model from noisy captions.

Training with COCO captions. We next train our detection model using the COCO captions [7]. We use the 591,435 annotated captions paired to the 118,287 *train2017* images. For evaluation, we use the COCO *test-dev2017* and PASCAL VOC 2007 *test* sets. In our supplementary file, we show qualitative examples from the COCO *val* set.

Tab. 1 shows the results on PASCAL VOC 2007. At the top are two upper-bound methods that train on image-level *labels*, while the rest of the methods train on image-level *captions*. EXACTMATCH (EM) performs the worst probably due to its low data utilization rate, as evidenced by the fact that all methods incorporating pseudo labels improve performance notably. Specifically, EM+GLOVEPSEUDO uses free knowledge of the pre-trained GloVe embeddings. It alleviates the synonyms problem to a certain extent, thus it improves the mAP by 2% compared to EXACTMATCH. However, the GloVe embedding is not optimized for the specific visual-captions, resulting in noisy knowledge transformation. EM+LEARNEDGLOVE learns dataset-specific word embeddings. Its performance, as expected, is 3% better than EM+GLOVEPSEUDO in terms of mAP. The strongest baseline is EM+EXTENDVOCAB, as the manually picked vocabulary list covers most frequent occurrences.

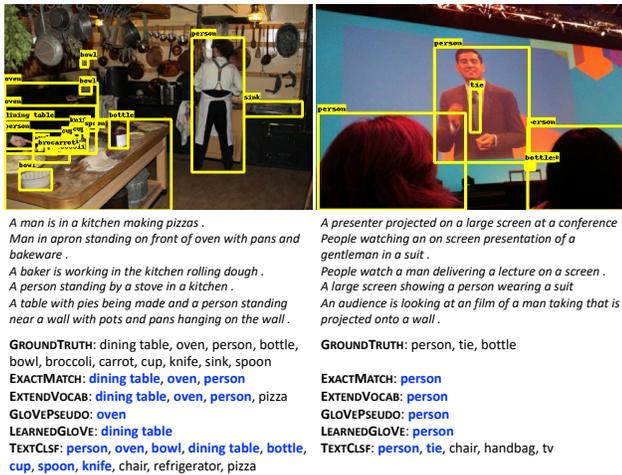


Figure 4: **Demonstration of different pseudo labels.** Our method fills the gap between what is present and what is mentioned, by making inferences on the semantic level. Matches to the ground truth are shown in blue.

Methods	Avg. Precision, IoU			Avg. Precision, Area		
	0.5:0.95	0.5	0.75	S	M	L
GT-LABEL	10.6	23.4	8.7	3.2	12.1	18.1
EXACTMATCH (EM)	8.9	19.7	7.1	2.3	10.1	16.3
EM + GLOVEPSEUDO	8.6	19.0	6.9	2.2	10.0	16.0
EM + LEARNEDGLOVE	8.9	19.7	7.2	2.5	10.4	16.6
EM + EXTENDVOCAB	8.8	19.4	7.1	2.3	10.5	16.1
EM + TEXTCLSF	9.1	20.2	7.3	2.6	10.8	16.6

Table 2: **COCO test-dev results (learning from COCO captions).** We report these numbers by submitting to the COCO evaluation server. The best method is shown in **bold**.

However, collecting such vocabulary requires human effort, and is not a scalable and transferable strategy. Our EM+TEXTCLSF outperforms this expensive baseline, especially for categories “cat”, “cow”, “horse”, and “train”.

At the top of Tab. 1 are two upper-bound methods which rely on ground-truth image-level captions. Despite the noisy supervision, our EM+TEXTCLSF almost bridges the gap to the COCO-labels upper bound.

For the results on COCO (Tab. 2), the gaps in performance between the different methods are smaller, but as before, our proposed EM+TEXTCLSF shows the best performance. We believe the smaller gaps are because many of the COCO objects are not described precisely via natural language, and the dataset itself is more challenging than PASCAL thus gain may be diluted by tough examples.

Training with Flickr30K captions. We also train our model on the Flickr30K [39] dataset, which contains 31,783 images and 158,915 descriptive captions. Training on Flickr30K is more challenging: on one hand, it includes less data compared to COCO; on the other hand, we observe that the recall rate of the captions is only 48.9% with EXACTMATCH which means only half of the data can be



Figure 5: **Data vs. Performance.** Our text classifier learned on COCO generalized well on Flickr30K and the noisier Flickr200K data formed by user-generated content tags.

matched to some class names. The results are shown in the bottom of Tab. 1. We observe that due to the limited training size, the detection models trained on Flickr30K captions achieve weaker performance than those trained on COCO captions. However, given the “free” supervision, the 33.6% mAP is still very encouraging. Importantly, we observe that even though our text classifier is trained on COCO captions and labels, it generalizes well to Flickr30K captions, as evidenced by the gap between EM+TEXTCLSF and EM+EXTENDVOCAB.

Data v.s. performance. We show the potential of our model using Flickr30K and MIRFlickr1M [16]. For the latter, we concatenate the title and all user-generated content tags to form caption annotation. We then use our text classifier learned on COCO to rule out examples unlikely to mention our target classes. This filtering results in a dataset with around 20% of the original data, and we refer to it as Flickr200K. We use 10%, 20%, 50%, 100% data from both datasets, and report average precision on VOC 2007. We see from Fig. 5 that as training data increases, mAP increases accordingly. To estimate model potential, we fit a square root function to the rightmost four points in the figure and use it to estimate 54.4 mAP at 1 million samples.

4.3. Using image labels as supervision

We finally show the performance of our method in the classic WSOD setting where *image-level supervision* is available. These results validate the method component described in Sec. 3.2. They also serve as an approximate *upper bound* for the more challenging task in Sec. 4.2.

Results on PASCAL VOC. For each image, we extract object categories from all the ground-truth bounding boxes, and only keep these *image-level* labels for training, discarding box information. For VOC 2007 and 2012, we train on 5,011 and 11,540 *trainval* images respectively and evaluate on 4,952 and 10,991 *test* images.¹ We report the standard mean Average Precision (mAP) at IoU > 0.5. We compare against multiple strong WSOD baselines. The results are shown in Tab. 3, and our single model outperforms the baseline methods (sometimes even ensemble methods) by a large margin. On VOC 2007, our model improves the

¹VOC 2012 result: <http://host.robots.ox.ac.uk:8080/anonymous/NOR9IV.html>

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
VOC 2007 results:																					
OICR VGG16 [31]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
PCL-OB-G VGG16 [30]	54.4	69.0	39.3	19.2	15.7	62.9	64.4	30.0	25.1	52.5	44.4	19.6	39.3	67.7	17.8	22.9	46.6	57.5	58.6	63.0	43.5
TS ² C [37]	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
OICR Ens.+FRCNN [31]	65.5	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	<i>63.1</i>	49.5	30.3	<i>64.7</i>	66.1	13.0	<i>25.6</i>	<i>50.0</i>	57.1	60.2	59.0	47.0
PCL-OB-G Ens.+FRCNN [30]	63.2	<i>69.9</i>	47.9	22.6	<i>27.3</i>	<i>71.0</i>	<i>69.1</i>	49.6	12.0	60.1	<i>51.5</i>	37.3	63.3	63.9	15.8	23.6	48.8	55.3	61.2	62.1	48.8
Ours	68.7	49.7	53.3	27.6	14.1	64.3	58.1	76.0	23.6	59.8	50.7	57.4	48.1	63.0	15.5	18.4	49.7	55.0	48.4	67.8	48.5
VOC 2012 results:																					
OICR VGG16 [31]	67.7	61.2	41.5	25.6	22.2	54.6	49.7	25.4	19.9	47.0	18.1	26.0	38.9	67.7	2.0	22.6	41.1	34.3	37.9	55.3	37.9
PCL-OB-G VGG16 [30]	58.2	66.0	41.8	24.8	27.2	55.7	55.2	28.5	16.6	51.0	17.5	28.6	49.7	70.5	7.1	25.7	47.5	36.6	44.1	59.2	40.6
TS ² C [37]	67.4	57.0	37.7	23.7	15.2	56.9	49.1	64.8	15.1	39.4	19.3	48.4	44.5	67.2	2.1	23.3	35.1	40.2	46.6	45.8	40.0
OICR Ens.+FRCNN [31]	71.4	69.4	55.1	29.8	28.1	55.0	57.9	24.4	17.2	<i>59.1</i>	21.8	26.6	57.8	71.3	1.0	23.1	52.7	37.5	33.5	56.6	42.5
PCL-OB-G Ens.+FRCNN [30]	69.0	<i>71.3</i>	<i>56.1</i>	30.3	27.3	55.2	57.6	30.1	8.6	56.6	18.4	43.9	<i>64.6</i>	<i>71.8</i>	7.5	23.0	46.0	44.1	42.6	58.8	44.2
Ours	74.2	49.8	56.0	32.5	22.0	55.1	49.8	73.4	20.4	47.8	32.0	39.7	48.0	62.6	8.6	23.7	52.1	52.5	42.9	59.1	45.1

Table 3: Average precision (in %) on the Pascal VOC test set using image-level labels. The top shows VOC 2007 and the bottom shows VOC 2012 results. The best single model is in **bold**, and best ensemble in *italics*.

Methods	Avg. Precision, IoU	
	0.5:0.95	0.5
Faster RCNN [26]	21.9	42.7
Faster Inception-V2 [15]	28.0	-
PCL-OB-G VGG16 [30]	8.5	19.4
PCL-OB-G Ens.+FRCNN [30]	9.2	19.6
Ours	10.6	23.4

Table 4: COCO detection using image-level labels, with supervised detection models at the top, best WSOD in **bold**.

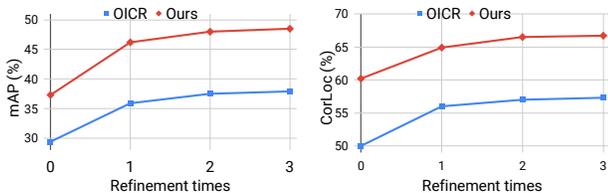


Figure 6: Analysis of our basic network and OICR components on VOC 2007. Comparison of the performance of our model and OICR VGG_M after iterative refinement.

mAP of the state-of-the-art single method TS²C method by 9%. On VOC 2012, our method outperforms the strongest single-model baseline PCL-OB-G VGG16 by 11%. Some prior work uses their WSOD detection results to further train an Fast-RCNN [11] detector (denoted as “+FRCNN” in Tab. 3) and obtain an additional 3 to 4 percents improvements on mAP. Even without such post-processing or ensemble, our model still achieves competitive performance on both VOC 2007 and 2012.

Effects of the basic network and OICR. The performance gain in our model comes from the following two aspects: (1) a more advanced detection model backbone architecture and (2) the online instance classifier refinement (OICR). Fig. 6 shows the performance of our method and that of Tang *et al.* [31] (OICR VGG_M), both refining for 0, 1, 2, 3 times. With no (0) refinement, our basic network

architecture outperforms the VGG_M backbone of Tang *et al.* by 27% in mAP. But the basic architecture improvement is not sufficient to achieve top results. If we use OICR to refine the models 1, 2, or 3 times, we gain 24%, 29%, and 30% respectively while Tang *et al.* achieve smaller improvement (22%, 28%, and 29% gains).

Results on COCO. We train our model on the 118,287 *train2017* images, using the image-level ground truth labels. We report mAP at IoU=.50:.05:.95 and mAP@0.5, on the 20,288 *test-dev2017* images. We compare to a representative fully-supervised detection model [26]; “Faster Inception-V2” [15] which is our method’s supervised detection counterpart, and a recent WSOD model [30]. As demonstrated in Tab. 4, our model outperforms the previous state-of-the-art WSOD method (PCL-OB-G Ens + FRCNN) by 15% in terms of mAP, but the gap between general WSOD methods (including ours) and the supervised methods is still large due to the disparate supervision strength.

5. Conclusion

We showed how we can successfully leverage naturally arising, weak supervision in the form of captions. We amplify the signal that captions provide by learning to bridge the gap between what human annotators mention, and what is present in the image. In the future, we will further verify the benefit of our approach by training on a larger, but even noisier dataset, namely WebVision [20]. We will also extend our method to incorporate raw supervision in the form of spoken descriptions in video. Finally, we will experiment with incorporating the text classifier training and the weakly supervised detection into the same end-to-end framework.

Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant No. 1566270. This research was also supported by Google Faculty Research Awards and an NVIDIA hardware grant.

References

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16*, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association.
- [2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [3] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.
- [4] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [5] G. Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [6] K. Chen, H. Song, C. Change Loy, and D. Lin. Discover and learn new objects from documentaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [7] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [8] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.
- [10] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [11] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [12] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, November 2011.
- [13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2017.
- [14] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2016.
- [15] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [16] M. J. Huijskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In *Proceedings of the international conference on Multimedia information retrieval*, pages 527–536. ACM, 2010.
- [17] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2016.
- [18] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [19] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [20] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.
- [21] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [22] I. Misra, C. Lawrence Zitnick, M. Mitchell, and R. Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [23] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [24] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [27] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2016.

- [28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [29] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. In *Advances in Neural Information Processing Systems*, pages 638–646, 2012.
- [30] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. L. Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [31] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [32] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [33] C. Thomas and A. Kovashka. Artistic object recognition by unsupervised style adaptation. In *Asian Conference on Computer Vision (ACCV)*, pages 460–476. Springer, 2018.
- [34] N. Turakhia and D. Parikh. Attribute dominance: What pops out? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.
- [35] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.
- [36] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [37] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [38] K. Ye and A. Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [39] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
- [40] M. Zhang, R. Hwa, and A. Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proceedings of the British Machine Vision Conference (BMVC)*, September 2018.
- [41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [42] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2014.