WEBLY-SUPERVISED VISUAL CONCEPT LEARNING WITH CARDINALITY GUIDED INSTANCE MINING AND CLUSTERED MULTITASK REFINEMENT

Saijie Ni, Xiaopeng Zhang, Botao Wang, Hongkai Xiong

Department of Electronic Engineering, Shanghai Jiao Tong University, China
{nsj161, zxphistory, botaowang, xionghongkai}@sjtu.edu.cn

ABSTRACT

Conventional image classification and object detection methods depend on manual annotations, such as image-level labels and bounding boxes. However, the acquisition of such annotations for millions of images is trivial. This paper addresses the problem of webly-supervised visual concept learning, and develops an automatic algorithm using parallel text and visual corpora to discover informative visual patterns from the web images. Based on the mined patterns, a cardinality-guided multiple instance learning algorithm is designed to establish the link between the image patterns and the literal concepts. Furthermore, due to the diversity of visual concepts, we perform clustered multitask refinement on the learned concept classifiers to enhance their generalization capability via a clustered regularization. Experiments demonstrate the superiority of the proposed method over traditional approaches.

Index Terms— Visual concept learning, multiple instance learning, multitask learning

1. INTRODUCTION

The past few decades have witnessed the explosive growth of data on the Internet, and we have achieved remarkable advances in data acquisition, storage and computation. Facing the dawn of Big Data, researchers are elevating the task of image classification [1] and object detection [2] to the next level — large scale visual concept learning. Unfortunately, although methods may be scalable and images can be enriched, the acquisition of manual annotations for millions of images is trivial. Therefore, it would be desirable to develop automatic approaches to learn visual concepts from large-scale image datasets with minimal human supervision.

Most existing visual concept learning methods discover the visual detectors in two ways. One way is to exploit image search engines to mine related object examples from image query results. For example, NEIL [3] starts with a few exemplar images per concept, and iteratively discovers common sense relationships and refines its concept detectors using image search results. The other is to discover visual patterns from weakly labeled data, which do not heavily depend on the search engines. For example, ConceptLearner [4] uses noisily tagged images to train concept detectors without considering the semantic similarity among different tags.

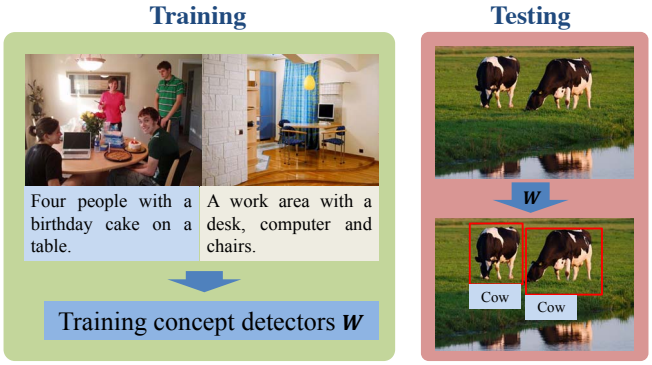


Fig. 1. Webly-supervised visual concept learning problem.

In this paper, we take the advantage of web images from the news sites and social media to learn visual concepts automatically. Fig. 1 shows the problem setting of the proposed webly-supervised visual concept learning method. Given a collection of web images with caption annotations, our goal is to mine visual concepts and train reliable object detectors. Although this paradigm is promising, we need to solve three challenges. First, how to determine the visual concepts from the textual descriptions. Second, how to efficiently extract relevant regions of the visual concepts from millions of possible regions in images. Third, how to establish the link between the visual concepts and image regions, which is also the most difficult task in a webly-supervised learning system.

This paper proposes an effective method to learn visual concepts from webly-supervised images with the following two contributions. First, we propose a cardinality-guided multiple instance learning algorithm for object instance mining. Conventional approaches [5] usually extract one positive instance from each positive bag, which is a waste of image information, since there might be multiple object instances in an image. In particular, the proposed method makes use of

The work was supported in part by the NSFC grants 61501294, 61622112, 61529101, 61472234, 61425011, China Postdoctoral Science Foundation 2015M581617 and Program of Shanghai Academic Research Leader 17XD1401900.

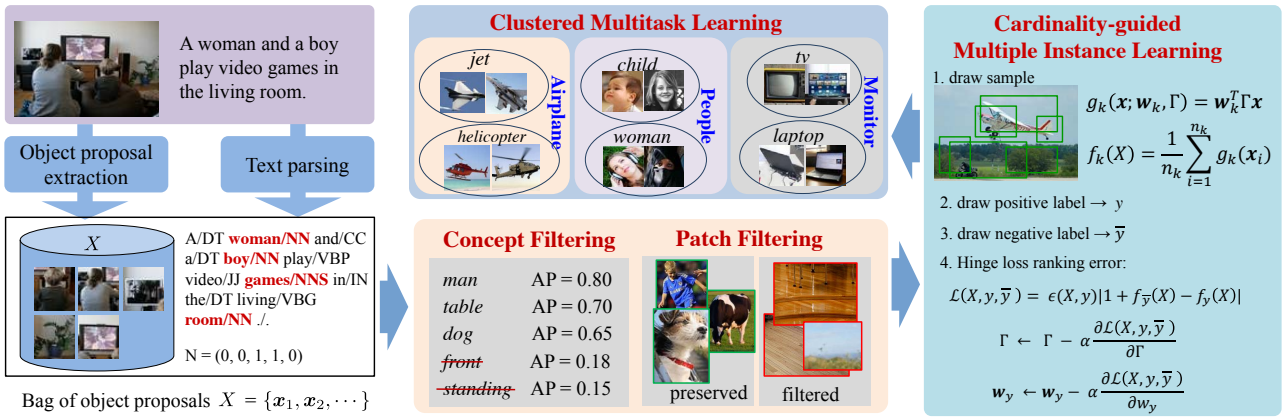


Fig. 2. The framework of the proposed visual concept learning algorithm.

the cardinality (i.e., the number of instances) inferred from the text to collect object instances more effectively. With this cardinality prior, we can obtain more object instances from the images for robust training of detectors. Second, we propose a multitask refinement approach to group semantically related concepts to super-category clusters. In practice, due to the diversity of visual concepts, many of them are semantically similar and refer to the same object actually. Hence, we further adopt clustered multitask refinement to enhance the learned concept classifiers by automatically grouping the relevant visual concepts together and refining their parameters with cluster related regularizations.

2. WEBLY-SUPERVISED VISUAL CONCEPT LEARNING

The flow chart of the proposed webly-supervised visual concept learning algorithm is illustrated in Fig. 2. Given a parallel corpus of images and the associated text descriptions, we first extract the names and cardinalities of the visual concepts from the captions. Subsequently, visually non-salient concepts and irrelevant patches are pruned with image search engine and submodular formulation, respectively. Then a cardinality-guided multiple instance learning algorithm is performed to identify the most relevant patches for each class of visual concept. Finally, in order to learn robust object detectors from the subcategory level visual concepts, a clustered multitask learning is applied to learn multiple object classifiers simultaneously. In the following sections, we elaborate the details of each module.

2.1. Extraction of Visual Concepts from Captions

First, given a parallel corpus of images and captions, we first extract the names of the visual concepts and their cardinalities from the captions, which is illustrated in Fig. 3. Specifically, the visual concepts of an image are indicated by the nouns in the associated caption. To retrieve the nouns, we utilize the Stanford Parser [6] to label the part-of-speech (POS) tag for each word in the caption, and the words marked as NN and NNS are the singular nouns and the plural nouns, respectively.

In addition to the nouns themselves, the cardinalities of the nouns will be determined by text parsing in the meantime, which would facilitate the discovery of object instances in MIL step. To be concrete, there are two types of cardinalities: the *exact* ones and the *approximate* ones. The exact cardinalities of nouns are defined by either singular nouns or plural nouns with numerical modifiers. Obviously, the cardinality of a singular noun (e.g., *plane* in the first caption in Fig. 3) is 1, and the cardinality of a noun with numerical modifier (e.g., *gentleman* in the first caption in Fig. 3 is *two*) can be inferred from the numeral . On the other hand, the approximate cardinality of a plural noun without numerical modifier (e.g., *canoes* in the last caption in Fig. 3) is 2, since all we know is that there are at least two instances in the image. Hence, the cardinalities of the visual concepts in an image are denoted by $N = \{n_1, n_2, \cdots, n_{K_0}\}$. If the k-th noun in the vocabulary is not mentioned in the caption, $n_k = 0$, otherwise n_k equals to the cardinality of the noun.

2.2. Irrelevant Concept and Patch Pruning

Aiming at obtaining a set of discriminative and compact visual concepts, it would be necessary to prune those visually non-salient concepts, e.g., *inside*, or *stand*. We use a simple and fast image-classifier pruning method [7], based on BING image search engine. We randomly split the retrieved top images into training and validation sets $I_k = \{I_k^t, I_k^v\}$. The negative bags $\bar{I}_k = \{\bar{I}_k^t, \bar{I}_k^v\}$ are collected from random samples of images in existing labeled datasets, from those categories which do not have the same name as the category of interest. For each concept, we train a linear SVM with I_k^t as positive and $\bar{I}_k^t$ as negative training images. This classifier is then evaluated on a combined pool of validation images. The k-th concept are declared to be visually salient if the average precision (AP) of the k-classifier computed on is above a threshold. Af-

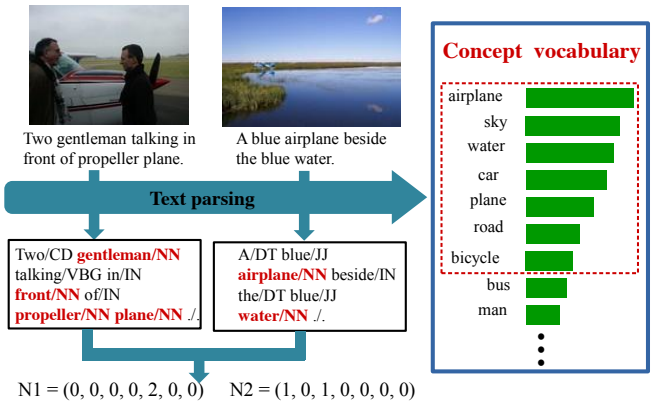


Fig. 3. Visual concept labels of the images.

ter the pruning step, we end up with $K (K \leq K_0)$ concepts.

Moreover, to remove background clutters and potential confusions from thousands of boxes per image, we adopt a flexible submodular formulation [8]. Let $\mathcal{P}$ be the set of all positively-labeled images. Each image X contains a set of candidate boxes $X = \{\boldsymbol{x}_i\}_{i=1}^{m_0}$, which is generated via object proposal methods [9]. For each box $\boldsymbol{x}$, we find its nearest neighbor box in each other image X'. The L closest of those neighbors form the neighborhood $\mathcal{N}(\boldsymbol{x})$.

We aim to define a function $F(S)$ on boxes sets S that measures how well the set S represents $\mathcal{P}$. If $\mathcal{B}(\mathcal{P})$ is the set of all patches from images in $\mathcal{P}$, then $|\mathcal{N}(\boldsymbol{x}) \cap \mathcal{B}(\mathcal{P})| \approx L$. To identify a representative set of patches among all candidate boxes like [8], we construct a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{U}, \varepsilon)$ where $\mathcal{U}$ and $\mathcal{V}$ are all boxes occurring in $\mathcal{P}$. The most representative patches maximize the covering function

$$F(S) = |\Gamma(S)| \tag{1}$$

where $\Gamma(S) = \{u \in \mathcal{U} | (u, v) \in \varepsilon \text{ with } v \in S\} \subseteq \mathcal{U}$ is the neighborhood of $S \subseteq \mathcal{V}$ in the bipartite graph. $\mathcal{S}$ denotes a set of selected boxes. $F(S)$ measures the number of boxes in $\mathcal{P}$ that are neighborhoods of S. The function F is monotone and submodular.

2.3. Instance Mining by Multiple Instance Learning

To establish the relation between the image-level visual concept labels and the object proposals, we represent each sample (i.e., an image-caption pair) as (X, N), where $X = \{\boldsymbol{x}_i\}_{i=1}^{m}$ is the bag of object proposals after patch filtering, $\boldsymbol{x} \in \mathbb{R}^d$ is the feature vector of an object proposal, d is the feature dimension, m is the number of object proposals, and N is the cardinalities of the visual concepts. First, we define the classification score of a patch $\boldsymbol{x}$ by the k-th classifier as

$$g_k(\boldsymbol{x}; \boldsymbol{w}_k, \Gamma) = \boldsymbol{w}_k^\top \Gamma \boldsymbol{x}, \quad k = 1, \cdots, K \tag{2}$$

In Eq. (2), $\Gamma \in \mathbb{R}^{h \times d}$ maps the original d-dimensional feature to an h-dimensional one ($h < d$), which is shared by all visual concepts. $\boldsymbol{w}_k \in \mathbb{R}^h$ is the coefficients of the k-th specific concept classifier, and K is the number of visual concepts. Both Γ and $\{\boldsymbol{w}_k\}_{k=1}^K$ will be estimated during training.

Based on the classification scores of the patches, we further define the classification score of a bag X. Conventional multiple instance learning approaches [5] assume that there is only one positive instance in each positive bag. However, as illustrated by the first image in Fig. 2, there are usually multiple object instances of the same class in an image. Therefore, to take full advantage of the object proposals and the textual descriptions, we develop a cardinality-guided multiple instance learning algorithm. Concretely, the classification score of a bag X is defined as

$$f_k(X) = \begin{cases} \dfrac{1}{n_k} \displaystyle\sum_{i=1}^{n_k} g_k(\boldsymbol{x}_i^*), & \text{if } n_k > 0 \\ g_k(\boldsymbol{x}_1^*), & \text{if } n_k = 0 \end{cases} \tag{3}$$

where $\{\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_{n_k}^*\}$ are the *key instances* in X, i.e., the n_k instances of the largest classification scores in X_k satisfying $f_k(\boldsymbol{x}_1^*) > f_k(\boldsymbol{x}_2^*) > \cdots > f_k(\boldsymbol{x}_{n_k}^*)$. Eq. (3) indicates that the classification score of a negative bag (i.e., $n = 0$) is the maximum classification score of the instances. On the other hand, the classification score of a positive bag (i.e., $n > 0$) is the mean classification score of the key instances. In this way, the proposed method is capable of mining more positive instances from the images, and achieve stronger generalization capability than methods that capture only one instance from each positive bag.

In order to optimize concept classifiers, i.e. Γ and $\{\boldsymbol{w}_k\}_{k=1}^K$, we apply a similar scheme as [10] using stochastic gradient descent (SGD). Specifically, given a sample (X, N), let $\bar{Y} = \{i | n_i = 0\}$ be the set of negative classes, and $Y = \{i | n_i > 0\}$ be the set of positive classes. The *ranking error* of this sample with respect to a positive class $y \in Y$ is defined as

$$\epsilon(X, y) = \sum_{i=1}^{R(X,y)} \frac{1}{i}, \tag{4}$$

where $R(X, y)$ is the number of negative classes that have larger score than $f_k(X)$. In each iteration of SGD, we first randomly select one sample (X, N), and then randomly pick a positive class $y \in Y$ and a negative class $\bar{y} \in \bar{Y}$ based on the concept labels of the chosen sample. The hinge loss of the ranking error of the sample is

$$\mathcal{L}(X, y, \bar{y}) = \epsilon(X, y) |1 + f_{\bar{y}}(X) - f_y(X)|_+ \tag{5}$$

Subsequently, Γ, $\boldsymbol{w}_y$ and $\boldsymbol{w}_{\bar{y}}$ can be updated accordingly by the corresponding derivatives of $\mathcal{L}(X, y, \bar{y})$.

2.4. Clustered Multitask Refinement

After the cardinality-guided multiple instance learning, we have collected key instances from bags of object proposals

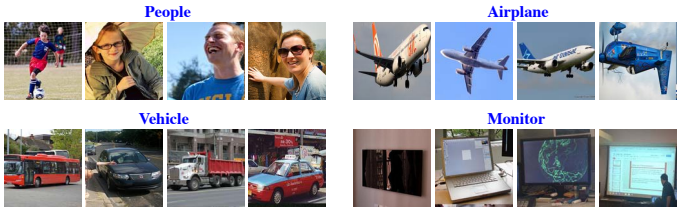


Fig. 4. Visual concepts discovered by the proposed method.

and the coefficients of the concept classifiers, i.e., $\{\boldsymbol{w}_k\}_{k=1}^K$. Concretely, the key instances and the their labels are denoted by $\{(\boldsymbol{x}_i', y_i^k)\}_{i=1}^M$ for $k = 1, \cdots, K$, where M is the number of instances, and y_i^k indicates whether the i-th instance belongs to the k-th concept.

In practice, due to the diversity of nouns, the visual concepts are cluster-structured. Namely, many visual concepts actually refer to the same object class. For example, visual concepts *person*, *girl* and *policeman* all stand for the object *person*. Hence, in order to strengthen the concept classifiers, we perform clustered multitask refinement on $\{\boldsymbol{w}_k\}_{k=1}^K$.

The weights of all visual classifiers are denoted by $W = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K) \in \mathbb{R}^{d \times K}$. Let $A \in \{0,1\}^{K \times T}$ be the cluster assignment of the visual concepts, i.e., $A(k,t) = 1$ if the k-th visual concept belongs to the t-th cluster. We further define $V = A(A^\top A)^{-1}A^\top$, which is only determined by the assignment of the visual concepts. The objective of the clustered multitask refinement is

$$\min_{W,V} \ell(W) + \lambda \Omega(W, V), \tag{6}$$

where

$$\ell(W) = \frac{1}{MK} \sum_{k=1}^{K} \sum_{i=1}^{M} |1 - \boldsymbol{w}_k^\top \boldsymbol{x}_i'|_+ \tag{7}$$

is the mean hinge loss of classification, and

$$\Omega(W, V) = \Omega_{\text{mag}}(W) + \alpha \Omega_{\text{inter}}(W, V) + \beta \Omega_{\text{intra}}(W, V) \tag{8}$$

is the regularization term with $\alpha, \beta \in \mathbb{R}$ being the weights.

In Eq. (8), $\Omega_{\text{mag}}(W)$, $\Omega_{\text{inter}}(W, V)$ and $\Omega_{\text{intra}}(W, V)$ penalize the magnitude of W, the inter-cluster variance and the compactness of the clusters, respectively. To solve Eq. (6), [11] provides a convex relaxation solution to the non-convex problem by optimizing over a convex set of positive semidefinite matrices.

3. EXPERIMENTS

We first experiment on MSCOCO [12] to validate the effectiveness of the learned concept detectors. Then we perform ablation study to evaluate the performance of each module. Finally, we adapt our detectors to PASCAL VOC 2007 [13] for comparison with previous object detection methods.

3.1. Experimental Settings

To evaluate the proposed method, we select the Microsoft COCO dataset (MSCOCO) , which consists of 82,783 training images and 40,504 validation images. For this dataset, each image is accompanied with five captions. For concept discovery, each word is labeled with the Stanford English Parser [6]. We kept nouns with at least 50 occurrences in the training sentences, which gives us an initial list of 1550 ($K_0 = 1550$) concepts. Bing image search API is then exploited to retrieve 100 images per concept, of which 80 for training and 20 for validation. Subsequently, we use linear SVM to prune irrelevant concepts. During the pruning process, concepts with AP lower than 10% is filtered, which results in 650 ($K = 650$) visual concepts. We generate object proposals with Multiscale Combinational Grouping (MCG) [9]. For feature representation, the fc7 layer [2] features are used for each region proposal.

3.2. Performance in Concept Discovery and Detection

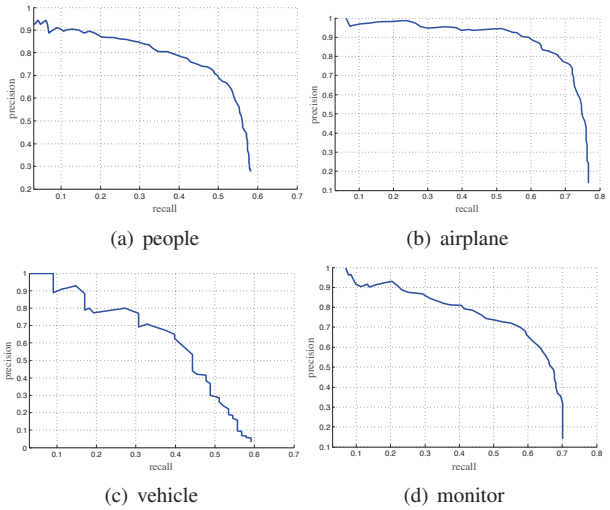
It should be noted that an extensive and comprehensive evaluation for the whole system is an extremely difficult task, because it is impractical to evaluate every labeled instance. Hence, we display the detailed results of four representative categories, namely *people*, *airplane*, *vehicle* and *monitor*. Fig. 4 shows the extracted visual concepts along with a few labeled instances belonging to the same category. It can be seen from the figure that the proposed method effectively handles the intra-class variation and polysemy via the clustering process. The purity and diversity of the clusters for different concepts indicate that contextual relationships help make our system robust to semantic drift and ensure diversity.

The precision-recall curves of the proposed method are shown in Fig. 5. In addition to the quantitative evaluation, we illustrate some detection samples in Fig. 8. It can be shown that our method is able to detect out the overlapped objects (e.g., *tv*), which is a challenging task without abundant instance-level annotations.

3.3. Ablation Study

In this section, we conduct ablation study to investigate the behaviors of different configurations. First, we evaluate the performance of the proposed cardinality-guided multiple instance learning algorithm. In particular, we analyze the recall of the positive instances labeled by the concept classifiers, which is demonstrated in Fig. 6. It can be observed from Fig. 6 that the algorithm basically converges in eight iterations for the four object classes. In general, more than half of the instances can be successfully discovered by the proposed method *in a fully webly-supervised fashion*.

In addition, we evaluate the influence of clustered multitask learning. The proposed method is is compared against two variants:



(a) people (b) airplane

(c) vehicle (d) monitor

Fig. 5. Precision-recall curves of the proposed methods.

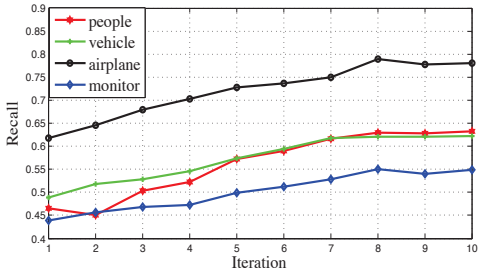
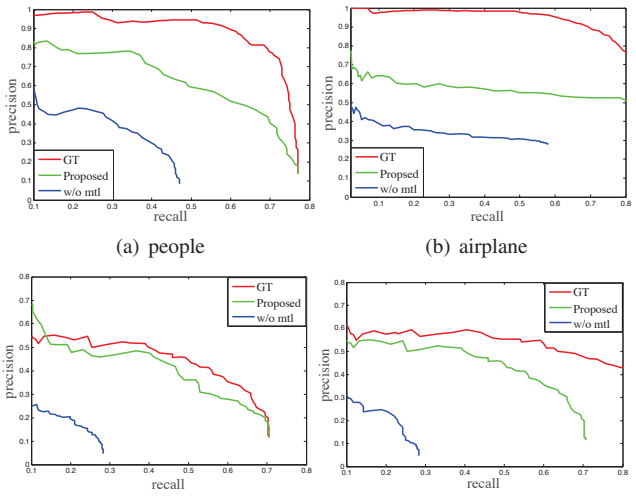


Fig. 6. Recall increasement as the number of MIL iterations.

• The ground truth object instances serve as the inputs to the clustered multitask learning.

• Clustered multitask learning is not performed. In other words, the object classifiers are trained directly using the visual concepts of the same name, *e.g.*, object *people* are learned with the positive patches of noun *people*.

The precision-recall curves of the two variants and the proposed method are illustrated in Fig. 7. We can observe from Fig. 7 that the best performance is achieved by multitask learning with ground truth object instances, which can be regarded as the upper bound of the proposed method. It should be noted that for object such as *vehicle*, the performance of the proposed method is close to the performance of the ground truth, which demonstrates that the proposed method is particularly effective in identifying instances like vehicles. On the other hand, simply using the exact words of the object class to learn the classifier achieves the worse performance among the three approaches, because an object class exhibits long-tailed distribution in textual description, so that a single visual concept does not offer sufficient instances to learn robust models. Remarkably, by grouping subcategory-level visual concepts using clustered multitask learning, the proposed method improves the generalization capability of the object detectors by utilizing extra semantically relevant instances.



(a) people (b) airplane

(c) vehicle (d) monitor

Fig. 7. Comparisons of precision-recall for different methods.

Fig. 8. Sample detections by the proposed method.

3.4. Detection on PASCAL VOC Dataset

We further evaluate the concept detectors on PASCAL VOC 2007 dataset. We follow the pipeline of region proposal and deep feature extraction in [2] as trainval and test. This dataset consists of about 5k trainval images and 5k test images over 20 object categories. Here, we need to apply a subset of the detectors from the pool of detectors learned from MSCOCO dataset to PASCAL VOC 2007 dataset. We simply use a winner-take-all selection protocol for detector selection. To be specific, we first define PASCAL VOC 2007 as the selection set and select relevant learned concept detectors with the highest precision. Then we evaluate the 20 best concept detectors for all 20 objects in PASCAL VOC 2007 respectively.

Table 1 compares the results of our concept discovery algorithm with other state-of-the-art baselines with various kinds of supervision. According to the source of supervision, we categorize these results into four different aspects.

• **Fully supervised method**. R-CNN [2] is a fully supervised state-of-the-art method on PASCAL VOC 2007. It is trained with deep features of the ground truth bounding boxes, and search for objects in pool of object proposals. In contrast to supervised methods, the proposed algorithm needs no instance-level annotations.

• **Weakly supervised method**. This type of methods uses image-level labels without bounding boxes to train the object

Table 1. Average precision of object detection.

Method	*aero*	*bike*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*	*table*	*dog*	*horse*	*mbike*	*person*	*plant*	*sheep*	*sofa*	*train*	*tv*	mAP
Proposed	.420	.405	.136	.098	.120	.366	.408	.159	.129	.167	.128	.172	.236	.395	.084	.154	.198	.238	.387	.216	.230
CL[4]	.345	.390	.182	.148	.084	.310	.391	.204	.155	.131	.145	.036	.206	.339	.094	.170	.147	.226	.279	.190	.209
LEVAN[7]	.140	.362	.125	.103	.092	.350	.359	.084	.100	.170	.065	.129	.306	.275	.060	.015	.188	.103	.235	.164	.172
DAV[14]	.174	-	.093	.092	-	-	.357	.094	-	.097	-	.033	.162	.273	-	-	-	-	.150	-	-
MDD[15]	.134	.440	.031	.031	.000	.312	.439	.071	.001	.093	.099	.015	.294	.383	.046	.001	.004	.038	.342	.000	.139
DCC[16]	.462	.469	.241	.164	.122	.422	.471	.352	.071	.283	.127	.215	.301	.424	.078	.200	.268	.208	.358	.296	.277
LLO[8]	.076	.419	.197	.091	.104	.358	.391	.336	.006	.209	.100	.277	.294	.392	.091	.193	.205	.171	.356	.071	.227
R-CNN[2]	.576	.579	.385	.318	.237	.512	.589	.514	.200	.505	.409	.460	.516	.559	.433	.233	.481	.353	.510	.574	.447

detectors. LLO (learning to localize objects with minimal supervision) [8] uses CNN to compute features and formulates the problem as a smoothed latent SVM optimization. DCC (object detection with convex clustering) [16] is also based on latent SVM. It introduces in the objective function an additional term to enforce similarity among the selected windows. MDD (model drift detection) [15] incorporates an initial annotation model to detect the drift of the model when training the detector. Since all these four methods (fully supervised and weakly supervised methods) use the training set and validation set of PASCAL VOC 2007 to train the detector, they are relevant to our method as "upper bound" baselines.

• **Video supervised method**. DAV (detectors from weakly annotated videos) [14] trains detectors on manually selected videos without bounding boxes and shows results on 10 classes of PASCAL VOC 2007.

• **Webly supervised method**. LEVAN (learn everything about anything) [7] uses items in Google N-grams as queries to collect images from image search engine for training the detectors. So their training set of detector could be considered as the unlimited number of images from search engines. CL (concept learner) [4] uses noisily tagged images to train concept detectors without considering the semantic similarity among different tags.

4. CONCLUSIONS

This paper presents a webly-supervised method to learn visual concepts from images with textual descriptions. Taking the advantage of natural language parsing and object proposal techniques, the proposed method designs a cardinality-guided multitask learning algorithm to establish the link between the image regions and the visual concepts. Experiments on challenging datasets demonstrate the superiority of the proposed method over traditional weakly supervised approaches.

5. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[3] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta, "Neil: Extracting visual knowledge from web data," in *ICCV*, 2013, pp. 1409–1416.

[4] Bolei Zhou, Vignesh Jagadeesh, and Robinson Piramuthu, "Conceptlearner: Discovering visual concepts from weakly labeled image collections," *Computer Science*, pp. 1492–1500, 2015.

[5] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2002, pp. 561–568.

[6] Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al., "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC*, 2006, pp. 449–454.

[7] Santosh Divvala, Ali Farhadi, and Carlos Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *CVPR*, 2014, pp. 3270–3277.

[8] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell, "On learning to localize objects with minimal supervision," *Eprint Arxiv*, pp. 1611–1619, 2014.

[9] Pablo Arbeláez, Jordi Pont-Tuset, and Barron et al., "Multiscale combinatorial grouping," in *CVPR*, 2014, pp. 328–335.

[10] Sheng-Jun Huang and Zhi-Hua Zhou, "Fast multi-instance multi-label learning," *arXiv preprint arXiv:1310.2049*, 2013.

[11] Laurent Jacob, Jean-philippe Vert, and Francis R Bach, "Clustered multi-task learning: A convex formulation," in *NIPS*, 2009, pp. 745–752.

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

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[14] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari, "Learning object class detectors from weakly annotated video," in *CVPR*, 2012, pp. 3282–3289.

[15] Parthipan Siva and Xiang Tao, "Weakly supervised object detector learning with model drift detection," in *ICCV*, 2011, pp. 343–350.

[16] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *CVPR*, 2015.