UNSUPERVISED MULTIPLE-INSTANCE LEARNING FOR INSTANCE SEARCH

Zhenzhen Wang *

Electrical and Electronic Engineering,
Nanyang Technological University
zwang033@e.ntu.edu.sg

Junsong Yuan

Computer Science and Engineering,
University at Buffalo
jsyuan@buffalo.edu

ABSTRACT

Traditional supervised Multiple-Instance Learning (MIL) has served as an important tool for a wide range of vision applications, for instance, image classification, object detection, and visual tracking. In this paper, we move forward one step further to tackle unsupervised computer vision problems by proposing an unsupervised multiple-instance learning algorithm, termed *UnMIL*. Different from classical MIL, our proposed unsupervised MIL does not require any manual annotations on neither bags nor instances. Given a collection of bags without any labels, our goal is to jointly optimize the bag label and instance label in a unified framework under the constraint of Noisy-OR model. The proposed *UnMIL* can be easily applied to object discovery in wild images by treating the object proposals extracted from images as instances and the according images as bags. Extensive experiments on MUSK1 & MUSK2, which is popularly used in MIL literature, on Oxford5k dataset for instance search, and on Object Discovery dataset for object co-localization, demonstrate the effectiveness of the proposed *UnMIL*.

Index Terms— Unsupervised Learning, Multiple-Instance Learning, Object discovery, Image search

1. INTRODUCTION

Exploring and analyzing large scale visual data has received a sustained attention in computer vision, especially in this big data era where there are millions of GB visual data are uploaded to websites such as Flickr and Facebook. The concrete tasks for exploring visual data include: image/video classification, object detection, image search, object co-localization, etc. To get satisfactory results for these tasks, a large body of fully/strongly annotated data is required during training phase. However, manually labeling the presence of objects and even their locations in visual data is time-consuming, expensive and laborious. Therefore, designing algorithms with weak supervision or even no supervision has been of great interest in recent years.

Multiple-Instance Learning (MIL), which was firstly proposed by [1] for classifying molecules in the context of drug design, are popularly applied to computer vision tasks. In MIL, training samples are usually given in the form of bags, and each bag consists of multiple instances (see Fig.1). In contrast to traditional supervised learning, in classical MIL [2, 3, 4, 5, 6] the labels are only provided for bags to indicate the positive or negative attribute. A bag is labeled positive if there exists at least one positive instance in it, and is negative if all of the instances contained in it are negative. This setting can be easily transferred to weakly-supervised learning in computer vision, e.g., weakly-supervised object detection, where each image can be seen as a bag and object proposals extracted from images as instances. Although enjoying popular applications in weakly-supervised computer vision, traditional MIL methods haven't been extended to unsupervised problems such as instance search and object co-localization due to the requirement of bag labels. Thus, it comes to the question: if there is no label of the bags and instances, can we still differentiate positive and negative bags from them?

To answer this problem, in this work we study unsupervised MIL. The difference between traditional MIL and the unsupervised MIL is illustrated in Fig.1. The supervised MIL is usually formulated as a classification problem with the labeled training bags, however the scenario is much more different and difficult for unsupervised MIL where no supervision could be utilized directly. Unsupervised MIL is also more complex than traditional single-instance unsupervised learning due to the inherent ambiguity in MIL. The advantages of unsupervised MIL are obvious: (1) the unlabeled data are much easier and cheaper to obtain; (2) unsupervised learning could help find the inherent structure of a data set.

After fully understanding the characteristics of unsupervised MIL, we propose a novel algorithm, termed *UnMIL*. Given unlabeled input data, the proposed *UnMIL* aims to predict both bag and instance labels. Specifically, we present a novel way to transform the general unsupervised MIL problem into a constrained sub-graph mining problem based on the assumption that the positive instances are more common since the negative ones are usually diverse in their own (see Fig. 1). The proposed *UnMIL* is a general method

*This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, and the Infocomm Media Development Authority, Singapore.

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

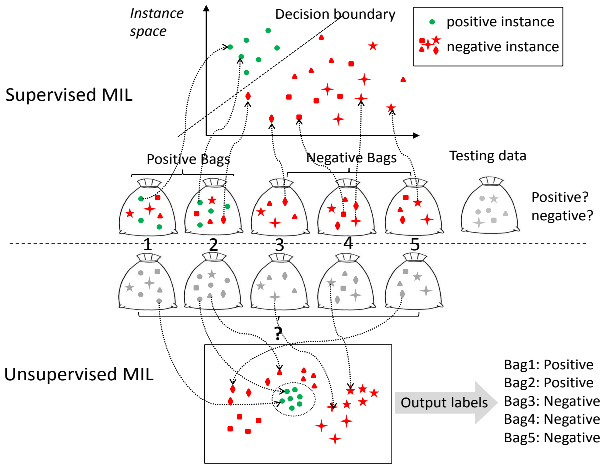


Fig. 1. Comparison of supervised MIL with unsupervised MIL. With labels in training data, the supervised MIL methods are usually formulated as classification problems. The unsupervised MIL problems could be solved based on the assumption that the positive instances appear more frequently than negative ones. (Better viewed in color.)

without strict constraint on the ratio of positive to negative bags/instances. In addition, different from previous MIL methods which focus on constructing learners either for classifying bags or for instances, the proposed *UnMIL* is a joint model that can simultaneously optimize the bag labels and instance labels.

To validate the effectiveness of the proposed *UnMIL*, we conduct experiments from two aspects: (1) comparing with classical MIL methods on two popularly used benchmarks in the MIL literature, MUSK1 and MUSK2 [1]; (2) comparing with state-of-the-art methods on the application of instance search and object co-localization on Oxford5k building dataset [7] and Object Discovery dataset [8], respectively. Essentially, the fundamental problems of the two applications are both object discovery. The difference is that the desired object is given as a query in instance search while it needs to be discovered automatically in object co-localization. Common object discovery is also the fundamental problem of computer vision, thus our proposed *UnMIL* has more potentials in computer vision applications.

2. METHOD

In this section, we will firstly introduce our unsupervised MIL algorithm, then give an example to show how to apply the proposed method to common object discovery.

2.1. MI Assumption

We use the following notations throughout this paper. Let $\mathcal{B} = \{B_1, B_2, \dots, B_N\}$ be a set of N bags, each of which contains several instances: $B_i = \{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \dots, \boldsymbol{x}_{i,n_i}\}$, $\boldsymbol{x}_{i,k} \in \mathcal{R}^d$. Each bag B_i is associated with a bag label $g_i \in \{0, 1\}$ and each instance is associated with an instance label $f_{i,k} \in \{0, 1\}$, where $0, 1$ represent negative and positive label, respectively. The relation between bag label and instance labels follows the Noisy-OR model:

$$g_i = \begin{cases} 1 & \text{if } \exists\, f_{i,k} = 1, \\ 0 & \text{if } \forall\, f_{i,k} = 0. \end{cases} \tag{1}$$

For traditional MIL algorithms, bag labels $\{g_i\}_{i=1}^N$ are usually given and the goal is to train a classifier based on these labeled bags. Since the majority of methods in the literature are evaluated by the bag-level classification accuracy, the key of designing classifiers is to achieve high bag-level performance. However, in many practical applications, predicting instance labels is as important as predicting bag labels. In this paper, we tackle a much more challenging scenario where neither the bag labels nor instance labels are known, and we will predict both instance labels and bag labels simultaneously.

2.2. Unsupervised MIL

In this part, we elaborate the proposed unsupervised MIL algorithm, *UnMIL*. Given a collection of bags without any supervision, our goal is to simultaneously predict the labels for bags and instances. We formulate the task as a constrained sub-graph mining problem, where the graph is built based on all the given instances with nodes and edges respectively corresponding to instances and their similarities. To simultaneously predict the bag and instance labels, we incorporate the regularization term on bags so that the negative bags can significantly avoid being selected. We begin by introducing the terms in our objective function, *i.e.*, the nodes and edges of the graph, which enable us to jointly optimize the bag labels and instance labels. Since the graph is built based on instances from all bags, we concatenate all instances into $\mathcal{B} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_M\}$ to reduce notation confusion. Then the instance labels can be denoted by $\boldsymbol{f} \in \mathbb{R}^M$, $M = \sum_{i=1}^N n_i$ is the total number of instances from the given dataset.

Node Confidences. Node confidence is defined as the probability of being positive instance. Generally, the node confidences represent some pieces of prior knowledge of the instances that can be observed, *e.g.*, the frequency of key words in the text recognition task, the saliency of the foreground in the object discovery task, and the similarity to the given query. We use $\{c_{i,k}\}_{i,k=1}^{N,n_i}$ to indicate the instance confidence of being positive, $\boldsymbol{c} \in \mathbb{R}^M$ is concatenated by instances confidences from all bags.

Edges of Graph. We encourage instances with similar appearance to have the same label through a similarity matrix. It is noted that the measurement for different features or different tasks varies a lot. For example, χ^2 distance is more suitable for Histogram-based features, while l_2 distance is more popular for measuring features extracted from DCNN. Therefore, the edges of the graph, *i.e.*, the similarities between nodes, are also determined data-dependently. Here we use $W \in \mathbb{R}^{M\times M}$, where $W_{k,l} \sim \text{similarity}(\boldsymbol{x}_k, \boldsymbol{x}_l)$, to denote the similarity matrix.

Index Matrix. We denote a binary matrix $A \in \mathbb{R}^{M\times N}$ to record the source bags of instances when building a graph, where $A_{k,i} = 1$ if the kth instance is from the ith bag. Then the following equation must hold: $A^\top \boldsymbol{f} = \boldsymbol{g}'$, where $\{g'_i\}_{i=1}^N$ is the number of positive instances in each bag. Following the *MIL constraints*, we have that $\boldsymbol{g} = \text{sign}(\boldsymbol{g}')$. In some practical applications, such as object discovery and image retrieval, there is usually only one positive instance in each positive bag. In such a case $\boldsymbol{g}'$ is also binary and follows $\boldsymbol{g}' = \boldsymbol{g}$.

Joint Formulation. Based on the terms presented above, we build the following sub-graph mining problem:

$$\begin{aligned} \min_{\boldsymbol{f},\boldsymbol{g}'} \quad & -2\boldsymbol{c}^\top \boldsymbol{f} + \lambda \boldsymbol{f}^\top L \boldsymbol{f} + \|\boldsymbol{g}'\|_1 \\ \text{s.t.} \quad & \boldsymbol{f} \in \{0,1\}^M, \\ & A^\top \boldsymbol{f} = \boldsymbol{g}'. \end{aligned} \tag{2}$$

where the Laplacian matrix L is defined as $L = D - W$, in which W is similarity matrix defined above, D is the diagonal matrix in which $D_{j,j} = \sum_{k=1}^M W_{j,k}$. The linear term $\boldsymbol{c}^\top \boldsymbol{f}$ aims at maximizing the cumulative score of the selected subgraph. The quadratic term $\boldsymbol{f}^\top L \boldsymbol{f}$ is to minimize the negative instances by emphasizing more on the edge connections with higher weights, and the parameter λ controls the influence of the connectivity. The regularization term $\|\boldsymbol{g}'\|_1$ attempts to avoid negative bags being selected.

Compared to previous methods which separately train models based on bags or instances, the proposed *UnMIL* is able to eliminate the false positive instances by regularizing on the bags via the joint formulation of bags and instances. In addition, different from previous methods which usually put efforts on predicting bag labels, we attach importance to both bags and instances. The necessity of prediction both is a crucial issue in computer vision tasks. For example, in the field of object discovery from images or videos, the images or frames are usually regarded as bags, and the patches extracted from them are regarded as instances, predicting the patch labels is as important as predicting the image/video labels. Thus, in the experiments we also evaluate the performance of the proposed method in terms of the instances.

2.3. Optimization

To solve the sub-graph mining problems defined in Eq. (2), we resort to the maximal flow algorithm proposed by Boykov and Kolmogorov [9]. Although the worst case complexity is in $\mathcal{O}(M^2en)$, where e represents the number of edges in the graph and n is the size of the minimum cut, it performs efficiently in practice since the graph is rather sparse.

2.4. Application to Object Discovery

In this part, we show how to apply the proposed *UnMIL* to the two popular tasks in computer vision: instance search and object co-localization. The two problems can be easily wrapped as MIL by treating each image as a bag, and object proposals cropped from images as instances.

Instance Confidence. To utilize our proposed *UnMIL* for object discovery, we first extract object proposals as instances. It is well acknowledged that the foreground objects usually appear in high saliency regions. Thus we use saliency as the indicator of being positive instances for object co-localization. For instance search, the node confidence can be achieved by measuring the similarities of reference patches to the query.

Similrity Matrix. Given instance representations which are extracted by a pretrained CNN model, we can calculate the similarity matrix W based on l_2 distance: $W_{k,l} = \exp(-\|\boldsymbol{x}_k - \boldsymbol{x}_l\|^2)$. For object co-localization, the similarity matrix is calculated on all instances as edges of the graph, while for instance search, it is calculated on all instances from references. Since the number of positive instances accounts for only a small proportion of the total instances, theoretically the similarity matrix should be sparse. Based on such an observation, the similarity $W_{k,l}$ is set to be 0 when $\boldsymbol{x}_k \in \mathcal{N}_K(\boldsymbol{x}_l)$ and $\boldsymbol{x}_l \in \mathcal{N}_K(\boldsymbol{x}_k)$ are not satisfied simultaneously, where $\mathcal{N}_K(\cdot)$ means the nearest K instances. In this paper, we always fix K as the number of bags.

3. EXPERIMENTS

To evaluate the performance of our proposed method, we compare with several representative supervised MIL algorithms, such as DD [2], EM-DD [3], citation(k)-NN [4], mi-SVM and MI-SVM [5] and RMI-SVM [10], on two MIL benchmarks, MUSK1&MUSK2 [1]. We then apply the proposed *UnMIL* to instance search and object co-localization, on Oxford5k building dataset [7] and Object Discovery dataset [8], respectively. In the following, we use "*UnMIL*-c" to denote the setting where the results are only determined by node confidences.

3.1. Ablation Studies

Effectiveness of each term. To visualize the effectiveness of our proposed *UnMIL* algorithm, we run a test experiment on some simulated 2D data points (see Fig.2(a)). In this simulation experiment, the node confidence is set as its density, and the similarity matrix is calculated by $W_{k,l} = \exp(-\|\boldsymbol{x}_k - \boldsymbol{x}_l\|^2)$. The affinity graph is constructed in the same way as described in Section 2.2, and the results are shown in Fig. 2. Each mark represents an instance and the instances with the same mark are from the same bag. There are 10 bags in total. The first column shows the ground-truth positive and negative instances (Top) and bags (Bottom) in green and black, respectively. In order to show the relative importance of each part in the affinity graph formulated in Eq. (2), we display the results, which are optimized based on the nodes only ($\boldsymbol{c}$) and on the nodes and edges (linear and Laplacian term in Eq. (2)), in the second and third columns. The results using the joint objective (Eq. (2)) are visualized in the last column. Comparing

(a) Detailed characteristics for MUSK1 & MUSK2.

Dataset	# attr.	# bag			# instance
		# pos.	# neg.	# total	
MUSK1	166	47	45	92	476
MUSK2	166	39	63	102	6,598

(b) Comparison to classical MIL algorithms (Acc(%)).

Method	MUSK1	MUSK2
DD [2]	88.9	82.5
EM-DD [3]	84.8	84.9
citation(k)-NN [4]	92.4	86.3
mi-SVM [5]	78.0	70.2
MI-SVM [5]	80.4	77.5
RMI-SVM [10]	80.8	82.4
UnMIL-c	76.1	73.5
UnMIL	89.1	83.3

Table 1. Introduction and Results of MUSK1 & MUSK2.

the second and third column we can find that with the Laplacian term, some negative instances with higher density can be removed (highlighted in red circle). From the third and last column we can see that with the regularization term, the outlier instances near the positive ones can be further eliminated.

Influence of paramter λ. We evaluate the effects of the parameter λ using MUSK1 and MUSK2, which consist of molecule (bag) and its various conformations (instances), only bag-level labels are provided. The goal is to predict whether a new drug molecule can bind well to a target protein. The detailed characteristics of the two datasets are listed in Tabel 1(a). For both datasets, we use the instance density as node confidence $\boldsymbol{c}$, and the similarity between any two conformations is calculated using $W_{k,l} = \exp(-\|\boldsymbol{x}_k - \boldsymbol{x}_l\|^2)$. We tune λ by uniformly sampling 10 values between $[\frac{\min \text{Eq.}(4)}{\max \text{Eq.}(5)}, \frac{\max \text{Eq.}(4)}{\min \text{Eq.}(5)}]$, which leads to $\lambda \in [0, 16]$ for MUSK1 & MUSK2. From Fig.2(b), we can see that the classification accuracy will increase when appropriate edge information is incorporated. Table 1(b) shows the average prediction accuracy in terms of bag-level, the results from literature are reported in their papers. Although all compared methods are supervised, our unsupervised algorithm, *UnMIL*, outperforms all methods except citation(k)-NN on MUSK1, and is also inferior to EM-DD on MUSK2. The huge gap between "*UnMIL*-c" and "*UnMIL*" ($\sim$ 13% on MUSK1 and $\sim$ 10% on MUSK2) demonstrates the effectiveness of the proposed joint algorithm.

3.2. Instance Search

Instance search is one of the most popular problems in computer vision. The straightforward solution is to compute the similarity between the query object and all reference images, then the results are achieved by ranking and thresholding the reference images according to the similarities. Although good performance has been observed, the results can be further improved by incorporating the interdependency between reference images. The motivation is obvious.If a reference $\mathcal{I}$ is associated with a query $\mathcal{Q}$, then it is possible that the references which enjoy high similarity with $\mathcal{I}$ may also be associated with $\mathcal{Q}$. On the contrary, this can help to remove false positive references which enjoy high similarity to query but do not affiliate the query, if most of the neighbors have large distance to the query.

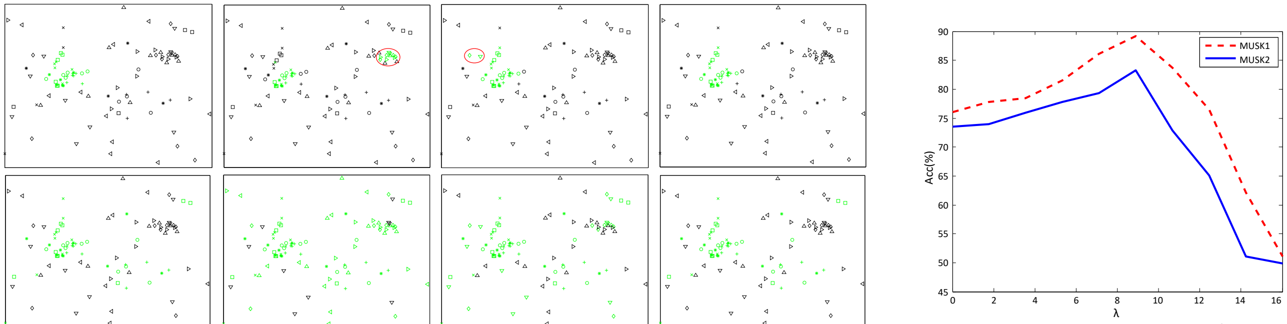
Method	Dimension	mAP
Babenko *et al.* [11]	256	65.7
Tolias *et al.* [12]	512	77.3
Arandjelovic *et al.* [13]	256	63.5
Radenovic *et al.* [14]	512	77.0/80.1
Rezende *et al.* [15]	512	64.1
Yu *et al.* [16]	512	73.9
R-MAC: UnMIL-c	512	77.6
R-MAC: UnMIL	512	80.2
MAC: UnMIL-c	512	73.9
MAC: UnMIL	512	76.3

Table 2. The results of instance retrieval on Oxford5k.

We evaluate our proposed *UnMIL* on this task in the Oxford5k building dataset [7], which is composed of 5063 reference images and 55 queries. We extract 300 object proposals from each reference image, then represent each proposal by the maximum activation of convolutions (MAC) feature and regional maximum activation of convolutions (RMAC) feature [12]. The final dimensions of the two kinds of features are both 512. The node confidence and similarity matrix are calculated following Section 2.4. Let $P \in \mathcal{R}^{d \times M}$ be the features of object proposals extracted from all reference images and $q \in \mathcal{R}^d$ be the feature of query object, the node confidence is expressed as $\boldsymbol{c} = q^\top P, \boldsymbol{c} \in \mathcal{R}^M$ and the similarity matrix $W = P^\top P, W \in \mathcal{R}^{M \times M}$. Note that our proposed *UnMIL* could be easily tailored to the problem of instance search, while classical supervised MIL algorithms, which require bag/image labels for training, can not be applied to such a problem. Thus, we only compare with some representative studies in the literature of instance search. From Table 2, we can see that with similar feature dimension, our *UnMIL* using R-MAC can achieve the state-of-the-art performance. Compared with the results of using only the similarity between the query and reference images ("*UnMIL-c*"), the *UnMIL* framework can increase about 3% with both MAC and R-MAC, which demonstrates the effectiveness of the proposed *UnMIL* framework. The examples of search results are shown in Fig.3.

3.3. Object Co-localization in Wild Images

The problem common object discovery and localization from real-world images is one of the most popular topics in computer vision. We perform this task in the Object Discovery dataset [8], which contains 300 images evenly separated into three categories: airplane, car, and horse. We generate 300 object proposals from each image, and then extract the 4096-dimensional feature vector using DCNN [17]. The node confidence and similarity matrix are calculated following Sec-



(a) The influence of each term w.r.t. instance-level (top) and bag-level (bottom). From left to right: groundtruth, node only, node & edge, and *UnMIL*.

(b) The influence of parameter λ on bag-level classification accuracy.

Fig. 2. The ablation studies on each part of the proposed *UnMIL*.

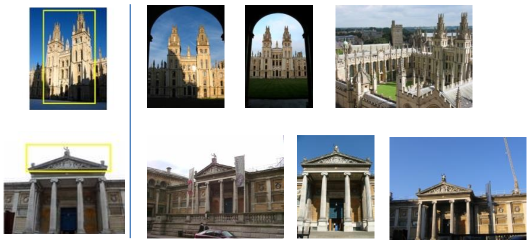


Fig. 3. Examples of instance search results.

(a) Comparison to classical MIL algorithms (Acc(%)).

Method	bag			instance		
	Airplane	Car	Horse	Airplane	Car	Horse
DD [2]	68.3	83.4	71.4	65.9	75.3	72.6
EM-DD [3]	69.5	83.1	72.1	70.2	77.6	70.5
citation(k)-NN [4]	81.3	76.1	75.8	79.2	74.7	73.6
mi-SVM [5]	81.1	88.5	75.7	76.5	77.2	73.1
MI-SVM [5]	82.3	88.0	78.6	-	-	-
UnMIL-c	69.8	76.0	68.2	68.4	80.0	67.9
UnMIL	82.1	87.3	84.9	79.9	87.2	79.6

(b) Comparison to state-of-the-arts (CorLoc(%)).

Method	Airplane	Car	Horse	Av.
Joulin *et al.* [19]	32.9	66.3	54.8	51.4
Kim *et al.* [20]	22.0	0.00	16.1	12.7
Joulin *et al.* [21]	57.3	64.0	52.7	58.0
Rubinstein *et al.* [8]	74.4	87.6	63.4	75.2
Cho *et al.* [22]	82.9	94.4	75.3	84.2
UnMIL-c	69.8	76.0	68.2	71.3
UnMIL	82.1	87.3	84.9	84.8

Table 3. Results on Object Discovery dataset.

tion 3.4. We manually label the object proposals/instances as positive if the criteria $\frac{area(\mathcal{P}_i \cap \mathcal{P}_{gt})}{area(\mathcal{P}_i \cup \mathcal{P}_{gt})} > 0.5$ is satisfied, and if $\frac{area(\mathcal{P}_i \cap \mathcal{P}_{gt})}{area(\mathcal{P}_i \cup \mathcal{P}_{gt})} < 0.3$, then the instance is labeled negative.

Comparison to classical MIL algorithms. Table 3(a) shows the classification accuracy (%) in terms of instance and bag level. The experiments for DD, EM-DD, citation(k)-NN, mi-SVM and MI-SVM are implemented based on the MIL library [18]. It can be seen that the learning ability of our *UnMIL* outperforms the majority classical supervised MIL algorithms by a great margin, especially in the instance level.

Comparison to state-of-the-arts. Table 3(b) shows the comparison of state-of-the-art cosegmentation [20, 19, 21] and colocalization [8, 23, 22] methods and the proposed *UnMIL*. The extensively used correct localization (CorLoc) metric is adopted for a fair comparison. CorLoc (%) is defined as the percentage of images correctly localized according to the PASCAL criteria: $\frac{area(\mathcal{P}_i \cap \mathcal{P}_{gt})}{area(\mathcal{P}_i \cup \mathcal{P}_{gt})} > 0.5$, where $\mathcal{P}_i$ is the predicted box (instance) and $\mathcal{P}_{gt}$ is the ground-truth box. From table 3(b) we can see that our method is comparable to the state-of-the-art [22], and outperforms other methods by a large margin. Fig.4 shows some examples achieved by our *UnMIL*.

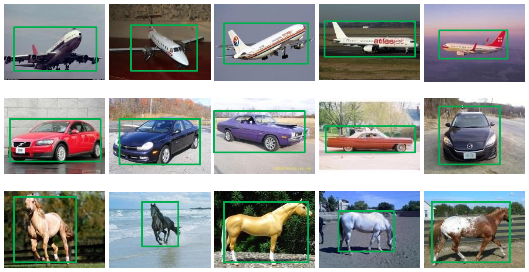


Fig. 4. Examples of successful co-localization results.

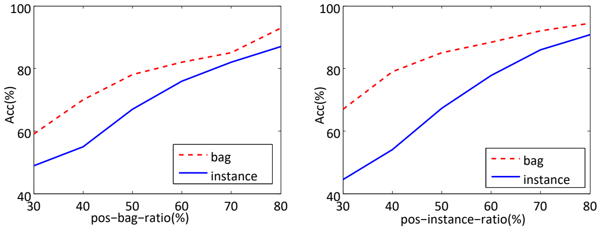


Fig. 5. Influence of various pos.-to-neg. ratios in terms of bags (Left) and instance (Right).

Influence of various pos.-to-neg. ratios of bags and instances. In order to demonstrate the robustness of our proposed *UnMIL* algorithm, we conduct experiments with the proportion of positive bags and instances varying from $30\% \sim 80\%$. Specifically, the evaluation size of bags and instances contained in each bag is fixed as 100. The positive bags are selected from the Object Discovery dataset and the negative bags are randomly collected from Internet, while the positive and negative instances are selected from our manually labeled object proposal pool. When evaluating the influence of bag distribution, we fix the proportion of positive instances in each positive bag as 50%. Similarly, when evaluating the influence of instance distribution, we fix the proportion of positive bags as 50%. From Fig. 5 we can see that both the bag- and instance-level predictions are rather stable over high positive ratios, and the performance is still satisfactory on bag-level evaluation even when the positive ratio decreases to 30%. It suggests that our proposed *UnMIL* can work well when we have enough positive bags and each positive bag contains enough positive instances.

4. CONCLUSION

In this paper, we propose a novel unsupervised MIL algorithm and apply it to computer vision problems. The proposed *UnMIL* is a joint model that can simultaneously inference the labels of bags and instances. It is also a general algorithm that can be easily applied to computer vision problems, such as instance search and object co-localization. To evaluate the proposed method, we conduct extensive experiments both on MIL benchmarks and also on the application of the common object discovery in real-world images. The superior performance compared with classical MIL methods and the-state-of-arts demonstrates the advantages of our proposed unsupervised method, *UnMIL*.

5. REFERENCES

[1] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez, “Solving the multiple instance problem with axis-parallel rectangles,” *Artificial intelligence*, 1997.

[2] Oded Maron and Tomás Lozano-Pérez, “A framework for multiple-instance learning,” *Advances in neural information processing systems*, 1998.

[3] Qi Zhang and Sally A Goldman, “Em-dd: An improved multiple-instance learning technique,” in *Advances in neural information processing systems*, 2001.

[4] Jun Wang and Jean-Daniel Zucker, “Solving multiple-instance problem: A lazy learning approach,” 2000.

[5] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann, “Support vector machines for multiple-instance learning,” *Advances in neural information processing systems*, 2003.

[6] Yixin Chen and James Z Wang, “Image categorization by learning and reasoning with regions,” *JMLR*, 2004.

[7] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *CVPR*, 2007.

[8] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu, “Unsupervised joint object discovery and segmentation in internet images,” in *CVPR*, 2013.

[9] Yuri Boykov and Vladimir Kolmogorov, “An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision,” *PAMI*, 2004.

[10] Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai, “Relaxed multiple-instance svm with application to object discovery,” in *ICCV*, 2015.

[11] Artem Babenko and Victor Lempitsky, “Aggregating local deep features for image retrieval,” in *ICCV*, 2015.

[12] Giorgos Tolias, Ronan Sicre, and Hervé Jégou, “Particular object retrieval with integral max-pooling of cnn activations,” *ICLR*, 2016.

[13] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” in *CVPR*, 2016.

[14] Filip Radenović, Giorgos Tolias, and Ondřej Chum, “Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples,” in *ECCV*, 2016.

[15] Rafael Rezende, Joaquin Zepeda, Jean Ponce, Francis Bach, and Patrick Perez, “Kernel square-loss exemplar machines for image retrieval,” in *CVPR*, 2017.

[16] Tan Yu, Yuwei Wu, Sreyasee Das Bhattacharjee, and Junsong Yuan, “Efficient object instance search using fuzzy objects matching.,” in *AAAI*, 2017.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012.

[18] Jun Yang, “Mill: A multiple instance learning library,” *URL http://www. cs. cmu. edu/juny/MILL*, 2008.

[19] Armand Joulin, Francis Bach, and Jean Ponce, “Discriminative clustering for image co-segmentation,” in *CVPR*, 2010.

[20] Gunhee Kim, Eric P Xing, Li Fei-Fei, and Takeo Kanade, “Distributed cosegmentation via submodular optimization on anisotropic diffusion,” in *ICCV*, 2011.

[21] Armand Joulin, Francis Bach, and Jean Ponce, “Multi-class cosegmentation,” in *CVPR*, 2012.

[22] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce, “Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals,” in *CVPR*, 2015.

[23] Ke Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei, “Co-localization in real-world images,” in *CVPR*, 2014.