Extracting Privileged Information from Untagged Corpora for Classifier Learning

Yazhou Yao[†], Jian Zhang[†], Fumin Shen[§*], Wankou Yang[♮], Xian-Sheng Hua[♯], Zhenmin Tang[‡]
[†]University of Technology Sydney, NSW, Australia
[§]University of Electronic Science and Technology of China, Chengdu, China
[♮]Southeast University, Nanjing, China
[♯]DAMO Academy, Alibaba Group, Hangzhou, China
[‡]Nanjing University of Science and Technology, Nanjing, China
{yaoyazhou, jzhang2211, fumin.shen, huaxiansheng, wankou.yang, tangzhenmin}@gmail.com

Abstract

The performance of data-driven learning approaches is often unsatisfactory when the training data is inadequate either in quantity or quality. Manually labeled privileged information (PI), *e.g.,* attributes, tags or properties, is usually incorporated to improve classifier learning. However, the process of manually labeling is time-consuming and labor-intensive. To address this issue, we propose to enhance classifier learning by extracting PI from untagged corpora, which can effectively eliminate the dependency on manually labeled data. In detail, we treat each selected PI as a subcategory and learn one classifier for per subcategory independently. The classifiers for all subcategories are then integrated together to form a more powerful category classifier. Particularly, we propose a new instance-level multi-instance learning (MIL) model to simultaneously select a subset of training images from each subcategory and learn the optimal classifiers based on the selected images. Extensive experiments demonstrate the superiority of our approach.

1 Introduction

Over the past decades, classifier learning approaches have been mostly data-driven [Coates *et al.*, 2011; Liu *et al.*, 2013; 2015; Yao *et al.*, 2016; Shen *et al.*, 2017]. The classifier is purely learned from a set of training samples (x_1,y_1),...,(x_n,y_n). Despite the success achieved, data-driven approaches become very brittle and prone to overfitting when the training data is inadequate either in quantity or quality.

A natural solution to alleviate this limitation is incorporating additional PI [Wang and Ji, 2015; Li *et al.*, 2014; Niu *et al.*, 2017]. For example, in object recognition, in addition to the image features and labels (*e.g.,* "horse"), the learner may also leverage object attributes (*e.g.,* "walking" and "jumping") in the training process. In human action recognition, besides the RGB features and human action labels, human joint positions can be incorporated into the classifier training. In practice, the PI can be tags, properties, attributes, positions or the context of the web images.

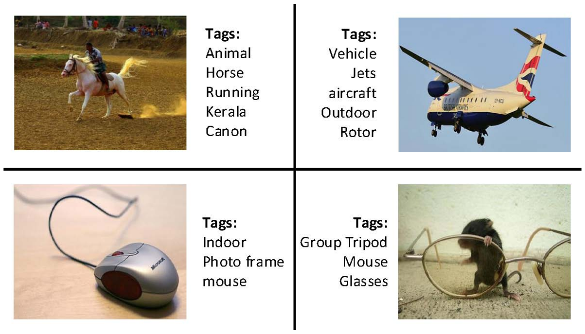


Figure 1: Examples of textual tags (privileged information) for images on image sharing website "Flickr". Both of useful and noisy tags are included.

However, learning classifier with PI is a challenging problem. The difficulty lies in three aspects. Firstly, the process of manually labeling PI is very expensive. Secondly, it is only available during training and unseen during testing. Thirdly, learning classifiers with PI overly depends on the quality of the collected PI. As shown in Fig 1, PI is often associated with noise in practice. If we failed to remove noise, the robustness of the learned classifier would be greatly reduced, and, in extreme cases, may become even worse.

Motivated by that, we seek to extract and leverage useful PI to enhance classifier learning. Different from previous works which discover PI from manually labeled descriptions, our approach extracts the PI from untagged corpora. The motivation is to eliminate the dependency on labeled data. In addition, different from previous works which usually encode PI into the parameters of the classifier, we focus on encoding PI into the structure of the classifier during training.

In our work, we mainly consider two critical issues. The

[*]Corresponding author: Fumin Shen

first is the PI derived from untagged corpora are usually noisy, how to remove noise and select the useful PI. The second is the retrieved web images are often associated with inaccurate labels, so the learned classifiers may be less robust and the classification performance may be significantly degraded, how one can purify noise and select useful images for learning robust classifiers. Specifically, we formulate noisy PI and noisy web images removing as a multi-view and multi-instance learning problem respectively. We propose a new instance-level MIL model to select images from each subcategory and simultaneously learn the optimal classifiers based on the selected images. To verify the effectiveness of our proposed approach, we conducted experiments on the tasks of image categorization and sub-categorization. Experimental results demonstrated the superiority of our proposed method.

The main contributions of this work consist of three aspects. Firstly, we eliminate the dependency on manually labeled PI and propose to extract that from untagged corpora. Secondly, we propose a novel instance-level MIL model to jointly learn the classifiers for categories and subcategories. Thirdly, from the experimental results, our proposed approach shows substantial improvement over existing weakly supervised methods.

2 Framework and Methods

Our proposed approach mainly consists of three major steps. Namely, discovering privileged information, purifying privileged information, and learning integrated classifier.

2.1 Discovering Privileged Information

Inspired by recent work [Divvala *et al.*, 2014], we can use Google Books Corpora [Lin *et al.*, 2012] to discover an exhaustive vocabulary explaining all the appearance variations for the given category. Following [Lin *et al.*, 2012] (see section 4.3), we specifically treat the dependency gram data with parts-of-speech (POS) as the PI. For example, given a category (*e.g.,* "horse") and its corresponding POS tag (*e.g.,* 'jumping, VERB'), we find all its occurrences annotated with POS tag within the dependency gram data. Of all the n-gram dependencies retrieved for the given category, we choose those whose modifiers are NOUN, VERB, ADJECTIVE, and ADVERB as the discovered candidate PI.

2.2 Purifying Privileged Information

Not all the candidate PI is useful, some noise may also be included. Using the noisy PI to enhance classifier learning will hurt both of the accuracy and robustness. To this end, we need to separate useful PI from noise before learning classifiers.

Our basic idea is to filter out the noisy PI from the perspective of relevance. Specifically, we denote the semantic distance of all discovered PI by a graph in which the given category (*e.g.,* "dog") is center y. Other candidate PI has a score S_{xy} corresponds to the Normalized Google Distance (NGD) [Cilibrasi and Vitanyi, 2007] between term x and y. Semantically relevant PI tend to have a smaller semantic distance than less-relevant PI. For example, PI "yawning dog", "Eskimo dog" and "police dog" which has a score 0.388, 0.286 and 0.372 respectively is much smaller than "down dog" which has a score 0.703.

However, this assumption is not always true from the perspective of visual relevance. For example, "hot dog" has a relatively smaller semantic distance 0.213, but is not relevant to the given category "dog" from the visual perspective. Thus, we need to identify both semantic and visual relevant PI. We retrieve the top K images from image search engine for each candidate PI. We denote each image as x_i and the compound feature of K images as $\phi_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ to represent its visual distribution. We calculate the Euclidean Distance between each candidate PI and the given category by using the compound feature to represent the visual distance. Similarly, visual relevant PI usually has a relatively smaller visual distance to the given category.

By treating semantic and visual distance as features from two different views, we formulate less relevant PI pruning as a multi-view learning problem. Our motivation is to find both semantically and visually relevant PI. During training, we model each view with one classifier and jointly learn two classifiers with a regularization term that penalizes the differences between two different classifiers.

Two views are reproducing kernel Hilbert spaces $\mathcal{H}_{K^{(1)}}$ and $\mathcal{H}_{K^{(2)}}$. Given l labeled data $(x_1, y_1), ...(x_l, y_l) \in \mathcal{X} \times \{\pm 1\}$ and u unlabeled data $x_{l+1}, ...x_{l+u} \in \mathcal{X}$, we seek to find predictors $f^{(1)*} \in \mathcal{H}_{K^{(1)}}$ and $f^{(2)*} \in \mathcal{H}_{K^{(2)}}$ that minimize the following objective function:

$$(f^{(1)*}, f^{(2)*}) = \text{argmin}_{\substack{f^{(1)}\in\mathcal{H}_{K^{(1)}}\\ f^{(2)}\in\mathcal{H}_{K^{(2)}}}} \text{Loss}(f^{(1)}, f^{(2)}) + \gamma_1 \left\| f^{(1)} \right\|^2_{\mathcal{H}_{K^{(1)}}} + \gamma_2 \left\| f^{(2)} \right\|^2_{\mathcal{H}_{K^{(2)}}} + \lambda \sum_{i=l+1}^{l+u} (f^{(1)}(x_i) - f^{(2)}(x_i))^2. \tag{1}$$

The first term is loss function and the next two are the regularization terms. The last term is called "co-regularization" which encourages the selection of a pair predictors $(f^{(1)*}, f^{(2)*})$ that agree on the unlabeled data. During testing, we make predictions by averaging the classification results from both of two views and the prediction rule is:

$$\mathcal{J} = \frac{1}{2}(f^{(1)}(x) + f^{(2)}(x)). \tag{2}$$

Following [Sindhwani *et al.*, 2005], we adopt the form of loss function as: $\text{Loss}(f^{(1)}, f^{(2)}) = \frac{1}{2l}\sum_{i=1}^{l}\left((f^{(1)}(x_i) - y_i)^2 + (f^{(2)}(x_i) - y_i)^2\right)$. We take the co-regularised least squares regression algorithm proposed in [Brefeld *et al.*, 2006] to solve (1). After we obtain the models for two views, we use (2) to prune noise and select useful PI.

2.3 Learning Integrated Classifier

We treat each selected PI as a subcategory for the target category. Suppose we obtain M subcategories, we collect the top few candidate images from image search engine for each subcategory. Although the image search engine has ranked the returned images, due to the error index of image search engine, some noisy images may still be included. We need to prune noise before learning integrated classifier.

By treating each subcategory as a "bag" and the retrieved images therein as "instances", we formulate noisy images removing and robust classifiers learning as an instance-level

MIL problem. For ease of presentation, we denote each instance as x_i with its label y_i and each bag G_m with the label Y_m. A matrix/vector is denoted by an uppercase/lowercase letter in boldface. The transpose of a vector or matrix is represented by $^\top$.

Since the retrieved images may contain noise, we need to select appropriate samples to train robust classifiers. To this end, a binary indicator $h_i \in \{0, 1\}$ is used to indicate whether or not training instance x_i is selected. To be exact, $h_i = 1$ when x_i is selected, and $h_i = 0$ otherwise. Due to the precision of images returned from the image search engine tends to have a relatively high accuracy, we define each positive bag as at least having a portion of η positive instances. The value of η can be estimated from some prior knowledge [Li *et al.*, 2011; Yao *et al.*, 2016]. We define $\mathbf{h} = [h_1, ...h_N]^\top$ as the indicator vector, and use $\mathrm{H} = \{\mathbf{h} | \sum_{i \in I_m} h_i = \eta \left| G_m \right|, \forall m\}$ to represent the feasible set of $\mathbf{h}$, where I_m represents the set of instance indices in G_m, and $|G_m|$ denotes the cardinality of G_m. We assume there are N retrieved web images coming from C categories and belonging to S subcategories. $z_{i,s} \in \{0, 1\}$ is a binary indicator variable and takes the value of 1 when x_i belongs to the *s*-th subcategory, and 0 otherwise. We denote $N_s = \sum_{i=1}^{N} z_{i,s}$ as the number of web training images from the *s*-th subcategory. $f_{c,s}(x) = (\mathbf{w}_{c,s})^\top \o(x)$ representing the classifier of the *s*-th subcategory and the *c*-th category. Based on existing MIL method [Li *et al.*, 2011], we propose our new MIL formulation as follows:

$$
\begin{aligned}
\min_{\mathbf{h}, \mathbf{w}_{c,s}, \xi_m} & \frac{1}{2} \sum_{c=1}^{C} \sum_{s=1}^{S} \|\mathbf{w}_{c,s}\|^2 + C_1 \sum_{m=1}^{M} \xi_m \\
\text{s.t. } & \frac{1}{|G_m|} \sum_{i \in I_m} h_i (\sum_{s=1}^{S} P_{i,s} (\mathbf{w}_{Y_m, s})^\top \phi(x_i) - \\
& (\mathbf{w}_{\hat{c}, \hat{s}})^\top \phi(x_i)) \geqslant \eta - \xi_m, \forall m, \hat{s}, \hat{c} \neq Y_m \\
& \xi_m \geqslant 0, \forall m
\end{aligned} \tag{3}
$$

where C_1 is a trade-off parameter, ξ_m are slack variables and $\phi(\cdot)$ is the feature mapping function. $P_{i,s}$ is the probability that the *i*-th training sample comes from the *s*-th subcategories. It can be obtained by calculating $P_{i,s} = (z_{i,s}/N_s) / \sum_{s=1}^{S} (z_{i,s}/N_s)$. In our model, we consider not only the relationship between category and its subcategories but also the relationship across different categories. Compared with classical MIL methods which mainly consider the relationship between category and its subcategories, our model has a better robustness.

Optimization

Problem (3) is a non-convex mixed integer problem and is hard to solve directly. However, the dual form of (3) can be relaxed as a multiple kernel learning (MKL) problem:

$$
\begin{aligned}
\min_{\mathbf{h}} \max_{\boldsymbol{\alpha}} & -\frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{Q}^{\mathbf{h}} \boldsymbol{\alpha} + \boldsymbol{\zeta}^\top \boldsymbol{\alpha} \\
\text{s.t. } & \sum_{c,s} \alpha_{m,c,s} = C_1, \ \forall m, \\
& \alpha_{m,c,s} \geqslant 0, \quad \forall m, c, s.
\end{aligned} \tag{4}
$$

$\boldsymbol{\alpha} \in \mathbb{R}^D$ ($D = M \cdot C \cdot S$) is a vector containing dual variables $\alpha_{m,c,s}$. $\boldsymbol{\zeta} \in \mathbb{R}^D$ is a vector, in which $\zeta_{m,c,s} = 0$ if $c = Y_m$ and $\zeta_{m,c,s} = \eta$ otherwise. Matrix $\mathbf{Q}^{\mathbf{h}} \in \mathbb{R}^{D \times D}$ can be calculated through $(1/ \left| G_m \right| \left| G_{\hat{m}} \right|)$ $\sum_{i \in I_m} \sum_{j \in I_{\hat{m}}} h_i h_j \o(\mathbf{x}_i)^\top \o(\mathbf{x}_j) \lambda(i, j, c, \hat{c}, s, \hat{s})$.

Problem (4) is a mixed integer programming problem and is hard to directly optimize the indicator vector $\mathbf{h}$. However, we can find the coefficients of $\mathbf{h}_t \mathbf{h}_t^\top$. We denote $\mathbf{d} = [d_1, ...d_T]^\top$, $T = |\mathrm{H}|$, and the feasible set of $\boldsymbol{\alpha}$, $\mathbf{d}$ as ν and $D = \{\mathbf{d} | \mathbf{d}^\top \mathbf{1} = 1, \mathbf{d} \geqslant 0\}$, respectively. Then we can get the following optimization problem:

$$
\min_{\mathbf{d} \in D} \max_{\boldsymbol{\alpha} \in \nu} \quad -\frac{1}{2} \sum_{t=1}^{T} d_t \boldsymbol{\alpha}^\top \mathbf{Q}^{\mathbf{h}_t} \boldsymbol{\alpha} + \boldsymbol{\zeta}^\top \boldsymbol{\alpha}. \tag{5}
$$

When we set the base kernel as $\mathbf{Q}^{\mathbf{h}_t}$, the above problem is similar to the MKL dual form and we are able to solve it on its primal form, which is a convex optimization problem:

$$
\begin{aligned}
\min_{\mathbf{d} \in \mathbf{D}, \mathbf{w_t}, \xi_\mathbf{m}} & \frac{1}{2} \sum_{t=1}^{T} \frac{\|\mathrm{w}_t\|^2}{d_t} + C_1 \sum_{m=1}^{M} \xi_m \\
\text{s.t. } & \sum_{t=1}^{T} \mathbf{w}_t^\top \varphi(\mathbf{h}_t, G_m, c, s) \geqslant \zeta_{m,c,s} - \xi_m, \ \forall m, c, s
\end{aligned} \tag{6}
$$

where $\varphi(\mathbf{h}_t, G_m, c, s)$ is the feature mapping function induced by $\mathbf{Q}^{\mathbf{h}_t}$. We solve the convex problem in (6) by updating $\mathbf{d}$ and $\{\mathbf{w_t}, \xi_m\}$ in an alternative way.

Update $\mathbf{d}$: We firstly fix $\{\mathbf{w_t}, \xi_m\}$ to solve $\mathbf{d}$. By introducing a dual variable β for constraint $\mathbf{d}^\top \mathbf{1} = 1$, the Lagrangian form of (6) can be derived as:

$$
\begin{aligned}
\pounds = & \frac{1}{2} \sum_{t=1}^{T} \frac{\|\mathrm{w}_t\|^2}{d_t} + C_1 \sum_{m=1}^{M} \xi_m - \sum_{m,c,s} \alpha_{m,c,s} \\
& (\sum_{t=1}^{T} \mathbf{w}_t^\top \varphi(\mathbf{h}_t, G_m, c, s) - \zeta_{m,c,s} + \xi_m) + \beta (\sum_{t=1}^{T} d_t - 1).
\end{aligned} \tag{7}
$$

Through set the derivative of (7) w.r.t d_t as zero, we get:

$$
d_t = \frac{\|\mathbf{w}_t\|}{\sqrt{2\beta}}, \ \forall t = 1, ..., T. \tag{8}
$$

For parameter β, $\|\mathbf{w}_t\| / \sqrt{2\beta}$ is monotonically decreasing. Parameter d_t satisfy $\sum_{t=1}^{T} d_t = 1$. We can use binary search method to solve β and recover d_t according to (8).

Update $\mathbf{w}_t$: When $\mathbf{d}$ is fixed, $\mathbf{w}_t$ can be obtained by solving $\boldsymbol{\alpha}$ in (5). Problem (5) is a quadratic programming problem w.r.t $\boldsymbol{\alpha}$. We employ the cutting-plane algorithm [Kelley, 1960] to solve this problem. By setting the derivatives of (7) w.r.t $\{\mathbf{w}_t, \xi_t, d_t\}$ as zeros, (5) can be rewritten as:

$$
\begin{aligned}
\max_{\beta, \boldsymbol{\alpha} \in \nu} & -\beta + \boldsymbol{\zeta}^\top \boldsymbol{\alpha} \\
\text{s.t. } & \frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{Q}^{\mathbf{h}_t} \boldsymbol{\alpha} \leqslant \beta, \ \forall t.
\end{aligned} \tag{9}
$$

We solve (9) by solving α with only one constraint at the first, then add a new violating constraint iteratively. We obtain the most violated constraint by optimizing:

$$
\max_{\mathbf{h}} \ \frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{Q}^{\mathbf{h}} \boldsymbol{\alpha} \tag{10}
$$

After a simple derivation, we can rewrite (10) as:

$$\max_{\mathbf{h}} \mathbf{h}^\top (\frac{1}{2}\hat{\mathbf{Q}} \odot (\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}^\top))\mathbf{h} \tag{11}$$

where $\hat{\alpha}_i = 1/|G_m| \sum_{c,s} \alpha_{m,c,s}$ for $i \in I_m$ and $\hat{\mathbf{Q}} = \sum_{c,\hat{c},s,\hat{s}} \phi(x_i)^\top \phi(x_j) \lambda(i,j,c,\hat{c},s,\hat{s})$. Problem (11) can be solved approximately through enumerate the binary indicator vector $\mathbf{h}$ in a bag by bag fashion iteratively to maximize (11) until there is no change in $\mathbf{h}$.

We train one classifier for each category and each subcategory. In general, a total of $C \times S$ classifiers $f_{c,s}(x)|c = 1,...C, s = 1,...S$ will be learned. The decision function for category C is obtained by integrating the learned classifiers from multiple subcategories: $f_c(x_i) = \sum_{s=1}^{S} P_{i,s} f_{c,s}(x_i)$.

During testing, we want to find the labels of the most matched subcategory and category, whose classifier achieves the largest decision value from all the subcategories and categories respectively. Thus, the subcategory label of image $\mathbf{x}$ can be predicted by:

$$\arg\max_{s} \mathbf{w}_{c,s}^\top \phi(\mathbf{x}) \tag{12}$$

and the category label by:

$$\arg\max_{c}(\max_{s} \mathbf{w}_{c,s}^\top \phi(\mathbf{x})). \tag{13}$$

3 Experiments

In this section, we first conduct experiments on both image categorization and sub-categorization to demonstrate the effectiveness of our proposed approach. We then quantitatively analyze the role of different steps contributing to the final results. In addition, we also analyze the time complexity of our proposed model.

3.1 Image Categorization

Experimental setting: We follow the setting in [Yao *et al.*, 2017; 2018] and exploit web images as the training set, human-labeled images as the testing set. Particularly, we evaluate the performance of our approach and other baseline methods on dataset PASCAL VOC 2007 [Everingham *et al.*, 2010], CIFAR-10 [Krizhevsky and Hinton, 2009] and STL-10 [Coates *et al.*, 2011].

For each category, we first discover the PI by searching in the Google Books Corpora. We calculate the NGD between the discovered PI and its target category as the semantic distance. The top $K = 100$ images from image search engine were retrieved for each discovered PI. We obtain the compound feature of each PI and calculate the Euclidean Distance between PI and its target category as the visual distance. We model each view with one classifier and jointly learn two classifiers with a regularization term that penalizes the differences between two different classifiers. We label a set of 500 relevant PI and 500 irrelevant PI to learn the prediction rule (2) for removing noise and selecting useful PI.

After we obtain the selected PI, the top 100 images were chosen for constructing the positive bags which corresponding to the selected PI. Negative bags can be obtained by randomly sampling a few irrelevant images. We use the proposed MIL model to select a subset of training images from each bag and simultaneously learn the optimal classifiers based on the selected images. We define each positive bag as having at least a portion of $\eta = 0.7$ positive instances and set the trade-off parameter $C_1 = 10^{-1}$. We evenly select 500 images from positive bags for each category to learn the integrated classifier. The features are 4096 dimensional deep features based on AlexNet [Krizhevsky *et al.*, 2012].

Method	Dataset		
	PASCAL	STL-10	CIFAR-10
sMIL	0.383	0.351	0.254
mi-SVM	0.414	0.381	0.278
RN-CMF	0.499	0.394	0.313
Sub-Cate	0.432	0.426	0.336
sMIL-PI	0.437	0.454	0.355
LIR	0.482	0.472	0.376
WSDG-PI	0.522	0.485	0.432
VCL	0.545	0.513	0.429
Ours	**0.582**	**0.557**	**0.464**

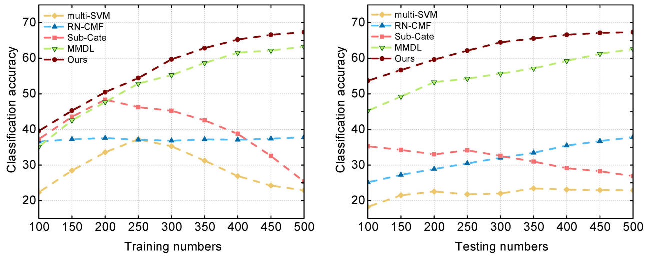
Table 1: The average performance comparison on the PASCAL VOC 2007, STL-10 and CIFAR-10 dataset.

Baselines: We compare our approach with three sets of weakly supervised baseline methods, the sub-categorization methods, the MIL methods, and the PI methods. The sub-categorization methods include Sub-Cate [Hoai and Zisserman, 2013] and RN-CMF [Ristin *et al.*, 2015]. The MIL methods contain instance-level method mi-SVM [Andrews *et al.*, 2003] and bag-level method sMIL [Bunescu and Mooney, 2007]. The PI methods consist of sMIL-PI [Li *et al.*, 2014], LIR [Wang and Ji, 2015], WSDG-PI [Niu *et al.*, 2017] and VCL [Divvala *et al.*, 2014].

Experimental results: The average performance comparison results are summarized in Table 1. From Table 1, we have the following observations:

PI methods VCL and our method performed better than three other PI methods sMIL-PI, LIR and WSDG-PI on the task of image categorization. One possible explanation is that the PI extracted from untagged corpora in both of our method and VCL is much richer and more accurate than three other methods in which the PI is obtained from the surrounding textual descriptions. Due to the complexity of the Internet, it is difficult to extract PI correctly from the descriptions.

Our proposed approach achieved the best average performance on all three datasets. Compared to MIL and sub-categorization methods, the classifiers learned by our approach not only using the visual features, but also the textual PI. Privileged information is usually more discriminative than the visual features in practical applications. Compared to PI methods which extract PI from the surrounding textual descriptions, the PI extracted by our method from untagged corpora is much more accurate and general. So the learned classifiers are more robust. Compared to VCL which takes multiple PI to learn a single classifier, our method exploits multiple PI to learn integrated classifier is more robust.



(a) (b)

Figure 2: Sub-categorization accuracy (%) of different methods: (a) using a varying number of training images for per subcategory, (b) using a varying number of testing images for per subcategory.

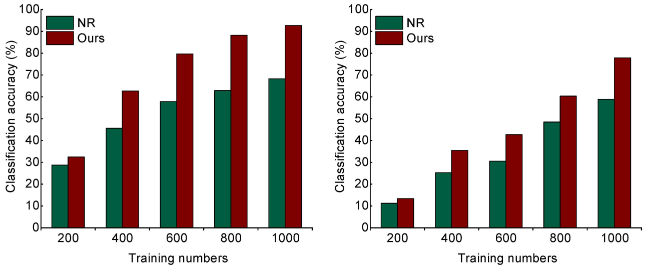
3.2 Image Sub-categorization

Experimental setting: For image sub-categorization, we choose a subset of ImageNet as the testing dataset. The reason is that ImageNet has a hierarchy structure. In particular, we select five categories including "airplane", "bird", "cat", "dog" and "horse" as the target categories and all their leaf synsets as the subcategories. Therefore, we can obtain 5 categories and 97 subcategories. The top 1000 images for each subcategory were retrieved from image search engine (Bing Image Search API-v7). We perform a cleanup step for broken links, webpages and obtain top ranked 700 images for each subcategory. We leverage the proposed MIL model to remove noise and learn classifiers. Specifically, we exploit the learned classifiers to re-rank the images in each subcategory according to the probability to be a positive sample. We sequentially select the top-ranked [100, 150, 200, 250, 300, 350, 400, 450, 500] images from each subcategory as the positive training samples to learn classifiers. 500 images per subcategory from ImageNet were selected as the testing data. In addition, we leveraged the top-ranked 500 images per subcategory as the positive training samples to learn classifiers and sequentially select [100, 150, 200, 250, 300, 350, 400, 450, 500] images per subcategory from ImageNet as the testing data. For this experiment, we also use the 4096 dimensional deep features based on AlexNet.

Baselines: We compare the image sub-categorization ability of our method with four weakly supervised baseline methods, multi-SVM [Weston and Watkins, 1998], Sub-Cate [Hoai and Zisserman, 2013], RN-CMF [Ristin *et al.*, 2015], and MMDL [Wang *et al.*, 2013].

Experimental results: The experimental results were provided in Fig 2 (a) and Fig 2 (b) respectively. The accuracy is measured by the average classification rate per subcategory.

By observing Fig 2 (a), we can see the best performance is achieved by our method, which produces significant improvements over other methods, particularly the number of positive training images over 250 for each subcategory. The reason is that our method considers the noisy images during the process of classifier learning. Due to the error index of image search engine, some noise may be included. We need to remove noise and select useful images from the retrieved web images to learn robust classifiers for each subcategory.



(a) (b)

Figure 3: Image categorization ability of NR and ours on PASCAL dataset: (a) "airplane", (b) "dog".

It is interesting to note in Fig 2 (a), while method RN-CMF implements a form of noisy images removing, the accuracy did not improve with the number of positive training images increase. The reason is that the noise is not the only factor that affects the classification accuracy. The visual distribution of selected images is another important factor.

By observing Fig 2 (a) and Fig 2 (b), our approach compares very favorably with competing algorithms, in terms of different numbers of training and testing images. Compared to method multi-SVM, Sub-Cate, RN-CMF, and MMDL, our approach achieves significant improvements in the subcategorization accuracy. The reason is our proposed MIL model not only considers the possible presence of noise in the web training data, but also tries to ensure the diversity of the selected images for classifier learning.

3.3 Performances of Methods with/without PI

To compare the classification performance with/without PI, we construct a new framework NR. We directly retrieve the web images as the candidate training data. We then take the MIL model to prune noisy images and train classifiers.

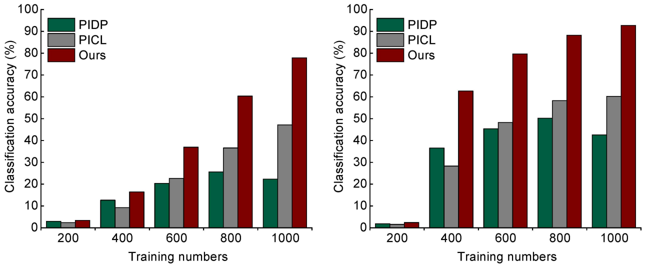
Specifically, "airplane" and "dog" are selected as two target categories to compare the image categorization ability. We sequentially collect [200,400,600,800,1000] images for each category to learn the classifiers. We test the categorization ability of NR and ours on the PASCAL VOC 2007 dataset. The results were shown in Fig. 3.

From Fig. 3, we can observe that PI-based approach performs better than NR, especially when the training number is greater than 200. The explanation is that with the increase of image numbers for each category, the retrieved images contain more and more noise. The noisy images caused by the image search engine have a worse effect than those induced by noisy PI. In this condition, our approach obviously outperforms method NR.

3.4 Different Steps Analysis

Our proposed framework involves three major steps. To quantify the role of different steps contributing to the final classifiers, we construct two new frameworks.

One is based on PI discovering and PI purifying (which we refer to PIDP). Another one is based on PI discovering and integrated classifier learning (which we refer to PICL).



(a)

(b)

Figure 4: Image categorization ability of PIDP, PICL and ours on PASCAL dataset: (a) "bird", (b) "horse".

For framework PIDP, we first obtain the PI through searching in the Google Books Corpora. Then we apply the PI purifying procedure to get the selected PI. We directly retrieve the top images from image search engine for selected PI to train image classifiers (without noisy images removing). For framework PICL, we also first obtain the candidate PI. Then we retrieve the top images from image search engine for all the candidate PI (without noisy PI purifying). We apply the MIL model to select images and train image classifiers.

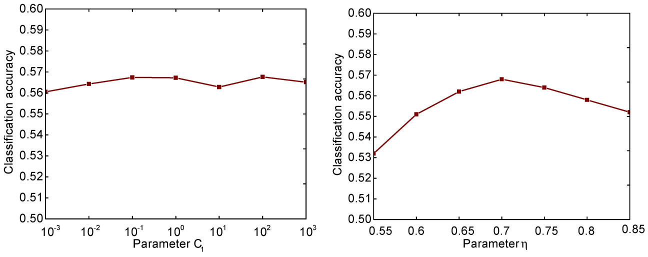
We compare the image categorization ability of these two new frameworks with our proposed framework. Specifically, "horse" and "bird" are selected as two target categories to compare the image categorization ability. We sequentially collect [200,400,600,800,1000] images for each category as the positive training samples and use 1000 fixed irrelevant negative samples to learn image classifiers. We test the image classification ability of these three frameworks on the PASCAL VOC 2007 dataset. The results are shown in Fig. 4.

From Fig. 4, we can observe that Framework PIDP usually performs better than PICL when the training number for each category is below 600. The explanation is that the first few retrieved images tend to have a relatively high accuracy. When the number of training images is below 600, the noisy images induced by noisy PI are more serious than those caused by the image search engine. With the increase of image numbers for each category, the images retrieved from the image search engine contain more and more noise. In this condition, the noisy images caused by the image search engine have a worse effect than those induced by noisy PI.

Our proposed framework outperforms both PIDP and PICL. This is because our approach, which takes a combination of noisy PI removing and noisy images filtering, can effectively remove the noisy images induced by both noisy PI and the error index of image search engine.

3.5 Parameter Sensitivity Analysis

For parameter sensitivity analysis, we mainly analyse two parameters C_1 and η in our MIL model. PASCAL VOC 2007 was selected as the benchmark testing dataset to evaluate the performance variation of our proposed approach. In particular, we vary one parameter by fixing another parameter as the default value. Fig 5 presents the parameter sensitiveness of C_1 and η in terms of image categorization accuracy.



(a)

(b)

Figure 5: The parameter sensitiveness of C_1 and η in terms of image categorization accuracy.

By observing Fig 5 (a), we found our method is robust to the parameter C_1 when it is varied in a certain range [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]. From Fig 5 (b), we noticed that the performance of our method is growing when η increases but less than 0.7. The reason is perhaps that our training data was derived from image search engine. Due to the error index of image search engine, there may be too much noise in each bag which will result in decreasing the classification accuracy when $\eta \leqslant 0.7$. When η increases over 0.7, the performance of our method decreases. One possible explanation is that the training set is less diverse. With the increase of η, the number of subcategories is decreasing, which may lead to the degradation of domain robustness of the classifier.

3.6 Time Complexity Analysis

During the process of solving our proposed MIL model, we solve the convex problem by using the cutting-plane algorithm. Through finding the most violating candidate $\mathbf{h}_t$ and solve the MKL subproblem at each iteration, the time complexity of our model can be approximately computed as $T \cdot O(\text{MKL})$, where the T is the number of iterations and the $O(\text{MKL})$ is the time complexity of the MKL sub-problem. According to [Platt *et al.*, 1999], the time complexity of MKL is between $t \cdot O(LCM)$ and $t \cdot O((LCM)^{2.3})$, where M, L, C are the numbers of latent domains, bags and categories respectively. t is the number of iterations in MKL.

4 Conclusion

In this paper, we presented a new approach for enhancing classifier learning by using privileged information. Different from previous works, our approach, while improving the accuracy and robustness of the classifier, greatly reduces the time and labor dependence. In our work, three successive modules were employed including PI discovering, PI purifying and integrated classifier learning. Specifically, we proposed a new instance-level MIL model to select a subset of training images from each selected PI and simultaneously learn the optimal classifiers based on the selected images. To verify the effectiveness of our proposed approach, we conducted experiments on both image categorization and subcategorization tasks. The experimental results demonstrated the superiority of our proposed approach over existing weakly supervised methods.

Acknowledgments

This research is supported by the National Science Foundation of China (Grant No. 61473154, 61773117, 61473086).

References

[Cilibrasi and Vitanyi, 2007] R Cilibrasi and P Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3): 370–383, 2007.

[Liu *et al.*, 2015] X Liu, L Wang, GB Huang, J Zhang, J Yin Multiple kernel extreme learning machine. *Neurocomputing*, 149: 253-264, 2015.

[Everingham *et al.*, 2010] M Everingham, L Gool, C Williams, and A Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010.

[Hoai and Zisserman, 2013] M Hoai and A Zisserman. Discriminative sub-categorization. *IEEE International Conference on Computer Vision and Pattern Recognition*, 1666–1673, 2013.

[Kelley, 1960] James E Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Mathematics*, 8(4):703–712, 1960.

[Li *et al.*, 2011] W Li, L Duan, D Xu, and I Tsang. Text-based image retrieval using progressive multi-instance learning. *IEEE International Conference on Computer Vision*, 2049–2055, 2011.

[Shen *et al.*, 2017] F Shen, Y Mu, Y Yang, W Liu, H Shen. Classification by Retrieval: Binarizing Data and Classifiers. *ACM SIGIR Conference on Research and Development in Information Retrieval*, 595-604, 2017.

[Sindhwani *et al.*, 2005] V Sindhwani, P Niyogi, and M Belkin, "A co-regularization approach to semi-supervised learning with multiple views," *International Conference on Machine Learning*, 74–79, 2005.

[Lin *et al.*, 2012] Y Lin, E Aiden, J Orwant, and S Petrov. Syntactic annotations for the google books ngram corpus. *ACL System Demonstrations*, 169–174, 2012.

[Ristin *et al.*, 2015] M Ristin, J Gall, M Guillaumin, and L Gool. From categories to subcategories: large-scale image classification with partial class label refinement. *IEEE Conference on Computer Vision and Pattern Recognition*, 231–239, 2015.

[Liu *et al.*, 2013] X Liu, L Wang, J Yin, E Zhu, J Zhang. An efficient approach to integrating radius information into multiple kernel learning. *IEEE transactions on cybernetics*, 43(2): 557-569, 2013.

[Krizhevsky and Hinton, 2009] A Krizhevsky and G Hinton. Learning multiple layers of features from tiny images. 2009.

[Wang *et al.*, 2013] X Wang, and Z Tu. Max-margin multiple-instance dictionary learning. *International Conference on Machine Learning*, 846–854, 2013.

[Weston and Watkins, 1998] J Weston and C Watkins. Multi-class support vector machines. Technical Report, 1998.

[Yao *et al.*, 2017] Y Yao, J Zhang, F Shen, X Hua, J Xu, Z Tang. Exploiting Web Images for Dataset Construction: A Domain Robust Approach. *IEEE Transactions on Multimedia*, 19(8): 1771-1784, 2017.

[Yao *et al.*, 2016] Y Yao, X Hua, F Shen, and Z Tang. A domain robust approach for image dataset construction. *ACM Conference on Multimedia*, 212–216, 2016.

[Krizhevsky *et al.*, 2012] A Krizhevsky, I Sutskever, and G Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105, 2012.

[Andrews *et al.*, 2003] S Andrews, I Tsochantaridis, and T Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 577–584, 2003.

[Yao *et al.*, 2016] Y Yao, J Zhang, F Shen, X Hua, J Xu, Z Tang. Automatic Image Dataset Construction with Multiple Textual Metadata. *IEEE Conference on Multimedia and Expo*, 1–6, 2016.

[Bunescu and Mooney, 2007] R Bunescu and R Mooney. Multiple instance learning for sparse positive bags. *ACM International conference on Machine learning*, 105–112, 2007.

[Coates *et al.*, 2011] A Coates, A Ng, and H Lee. An analysis of single-layer networks in unsupervised feature learning. *International conference on artificial intelligence*, 215–223, 2011.

[Divvala *et al.*, 2014] S Divvala, A Farhadi, and C Guestrin. Learning everything about anything: Webly-supervised visual concept learning. *IEEE Conference on Computer Vision and Pattern Recognition*, 3270–3277, 2014.

[Li *et al.*, 2014] W Li, L Niu, and D Xu. Exploiting privileged information from web data for image categorization. *European Conference on Computer Vision*, 437–452, 2014.

[Niu *et al.*, 2017] L Niu, W Li, D Xu. Visual recognition by learning from web data via weakly supervised domain generalization. *IEEE transactions on neural networks and learning systems*, 28(9):1985–1999, 2017.

[Platt *et al.*, 1999] J. Platt, "Fast training of support vector machines using sequential minimal optimization," *Advances in Kernel Methods*, 185–208, 1999.

[Wang and Ji, 2015] Z Wang and Q Ji. Classifier learning with hidden information. *IEEE Conference on Computer Vision and Pattern Recognition*, 4969–4977, 2015.

[Brefeld *et al.*, 2006] U Brefeld, T Scheffer, and S Wrobel. Efficient co-regularised least squares regression. *ACM conference on Machine learning*, 137–144, 2006.

[Shen *et al.*, 2017] F Shen, X Gao, L Liu, Y Yang, H Shen. Deep Asymmetric Pairwise Hashing. *ACM conference on Multimedia*, 1522-1530, 2017.

[Yao *et al.*, 2018] Y Yao, J Zhang, F Shen, W Yang, P Huang, Z Tang. Discovering and Distinguishing Multiple Visual Senses for Polysemous Words. *AAAI Conference on Artificial Intelligence*, 2018.