Visual Recognition by Learning from Web Data: A Weakly Supervised Domain Generalization Approach

Li Niu, Wen Li, Dong Xu,
School of Computer Engineering, Nanyang Technology University (NTU), Singapore

Recently, there is an increasing research interest in exploiting web images/videos crawled from Internet as training data to learn robust classifiers for recognizing new images/videos. However, the visual feature distributions of training and testing samples may differ considerably in terms of statistical properties, which is known as the dataset bias problem [6].

When target domain data is unavailable during the training process, the domain adaptation problem becomes another related task called domain generalization, which aims to learn robust classifiers that can generalize well to any unseen target domain [5]. Domain generalization is also an important research problem for the real-world visual recognition applications. For example, the datasets containing the photos/videos from each user can be considered as one target domain because different users may use different cameras to capture the photos/videos in their own ways. So we have a large number of target domains from various users and meanwhile some users may not be willing to share their photos/videos to others as target domain data due to the privacy issue. In this case, it is more desirable to develop new domain generalization approaches without using target domain data in the training process.

In this work, we study the domain generalization problem by exploiting web images/videos as source domain data. In our approach, we consider two important issues when exploiting web images/videos as source domain data: 1) the training web images and videos are often associated with inaccurate labels, so the learnt classifiers may be less robust, and the recognition performance may be significantly degraded as well; 2) the test data in the target domain usually has a different distribution from the training images/videos, and the target domain data is often unavailable in the training stage.

To cope with the label noise of web training images and videos, inspired by the multi-instance learning (MIL) methods, we partition the training samples from each class into a set of clusters, and then treat each cluster as a "bag" and the samples in each cluster as "instances". Inspired by multi-class SVM [1] and the MIL method KI-SVM [4], we formulate our task based on multi-class MIL. Specifically, we partition our training samples into L training bags, *i.e.*, $\{(\mathcal{B}_l, Y_l)|l = 1,\ldots,L\}$, and select the training samples from each training bag. The bag label Y_l is determined by using the corresponding class label while the labels of instances remain unknown. We define a ratio η to represent the proportion of training instances in each training bag, in which their instance labels are consistent with the bag-level label Y_l. We use a binary indicator $h_i \in \{0,1\}$ to indicate whether the training sample $\mathbf{x}_i$ is selected or not and denote $\mathbf{h} = [h_1,\ldots,h_N]'$ as the indicator vector.

To enhance the generalization ability of the learnt classifiers to any unseen target domain, we assume the training web images/videos may come from multiple hidden domains with distinctive data distributions, as suggested in the recent works [2, 3]. Then, we aim to learn one classifier for each class and each latent domain. As each classifier is learnt from the training samples with a distinctive data distribution, each integrated classifier obtained by combining multiple classifiers from each class is expected to be robust to the variation of data distributions, and thus can be well generalized to predict test data from any unseen target domain. In our work, we discover latent domains by using the existing technology in [2], which aims to maximize the sum of maximum mean discrepancy (MMD) between any two latent domains. The objective function in [2] is written as follows,

$$\max_{\boldsymbol{\beta}} \quad \sum_{m\neq\tilde{m}} (\boldsymbol{\beta}_m - \boldsymbol{\beta}_{\tilde{m}})'\mathbf{K}(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{\tilde{m}}) \tag{1}$$

where $\mathbf{x}_i$ is the i-th training sample, $\mathbf{K} = [K_{i,j}]$ with $K_{i,j} = \phi(\mathbf{x}_i)'\phi(\mathbf{x}_j)$ in which $\phi(\cdot)$ is the feature mapping function, $\boldsymbol{\beta}_m = [\beta_{1,m},\ldots,\beta_{N,m}]'$ with $\beta_{i,m}$ defined as the probability that the m-th latent domain contains $\mathbf{x}_i$, and N is the number of training samples,.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

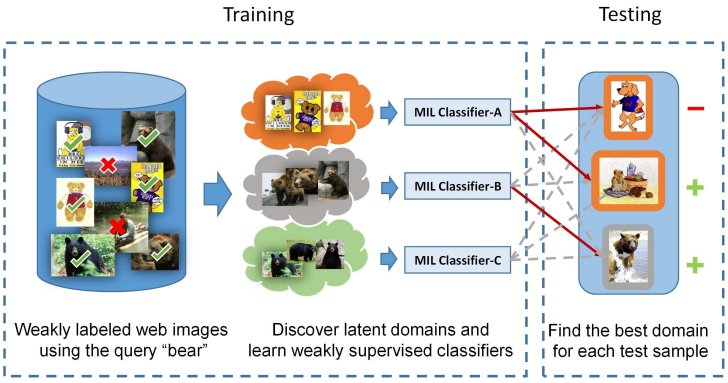

Figure 1: Illustration of how to handle label noise and enhance domain generalization ability simultaneously.

Suppose the source domain contains C classes and M latent domains, we propose to learn $C \times M$ classifiers $\{f_{c,m}(\mathbf{x})|c = 1,\ldots C,$ and $m = 1,\ldots,M\}$, where the classifier for the c-th class and the m-th latent domain $f_{c,m}(\mathbf{x}) = (\mathbf{w}_{c,m})'\phi(\mathbf{x})$. Recall that we only use a subset of training samples from each training bag for learning the classifiers, we further introduce a MMD-based regularizer to select the training samples with more distinctive data distributions in order to further enhance the domain generalization ability. Then, we arrive at our final formulation as follows,

$$\begin{aligned}\min_{\mathbf{h},\mathbf{w}_{c,m},\xi_l} \quad & \frac{1}{2}\sum_{c=1}^{C}\sum_{m=1}^{M}\|\mathbf{w}_{c,m}\|^2 + C_1\sum_{l=1}^{L}\xi_l \\ & -C_2\sum_{m\neq\tilde{m}}(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{\tilde{m}})'(\mathbf{K}\circ(\mathbf{h}\mathbf{h}'))(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{\tilde{m}}) \qquad (2)\end{aligned}$$

$$\text{s.t.} \quad \frac{1}{|\mathcal{B}_l|}\sum_{i\in I_l} h_i\left(\sum_{m=1}^{M}\hat{\beta}_{i,m}(\mathbf{w}_{Y_l,m})'\phi(\mathbf{x}_i) - (\mathbf{w}_{\tilde{c},\tilde{m}})'\phi(\mathbf{x}_i)\right) \geq \eta - \xi_l, \quad \forall l,\tilde{m},\tilde{c}\neq Y_l, \tag{3}$$

$$\xi_l \geq 0, \quad \forall l, \tag{4}$$

where C_1 and C_2 are the tradeoff parameters, ξ_l's are the slack variables, $\mathcal{I}_l$ is the set of indices of instances in the bag $\mathcal{B}_l$, $|\mathcal{B}_l|$ is the number of instances in the training bag $\mathcal{B}_l$, and $\hat{\beta}_{i,m}$ means the probability that $\mathbf{x}_i$ belongs to the m-th latent domain, which can be calculated as $\hat{\beta}_{i,m} = \frac{\beta_{i,m}}{\sum_{m=1}^{M}\beta_{i,m}}$. In the testing process, we predict the label of a given test sample $\mathbf{x}$ by using $y = arg\max_c \max_m(\mathbf{w}_{c,m})'\phi(\mathbf{x})$. Our comprehensive experiments also show that our newly proposed approach can handle label noise of training web images/videos and enhance the generalization capability to any unseen target domain.

[1] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2, 2002.

[2] Boqing Gong, Kristen Grauman, and Fei Sha. Reshaping visual datasets for domain adaptation. In *NIPS*, 2013.

[3] Judy Hoffman, Kate Saeko, Brian Kulis, and Trevor Darrell. Discovering latent domains for multisource domain adaptation. In *ECCV*. 2012.

[4] Yu-Feng Li, James T Kwok, Ivor W Tsang, and Zhi-Hua Zhou. A convex method for locating regions of interest with multi-instance learning. In *ECMLPKDD*, 2009.

[5] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

[6] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011.