# Semi-supervised Domain Adaptation with Subspace Learning for Visual Recognition

Ting Yao[1], Yingwei Pan[2], Chong-Wah Ngo[3], Houqiang Li[2], Tao Mei[1]
[1]Microsoft Research, [2]University of Science and Technology of China, [3]City University of Hong Kong.

In many real-world applications, we are often facing the problem of cross domain learning, i.e., to borrow the labeled data or transfer the already learnt knowledge from a source domain to a target domain. However, simply applying the classifier learnt in the source domain may hurt the performance in the target domain, a phenomenon known as "domain shift." Furthermore, the labeled target data are often very few and they alone are not sufficient to construct a good classifier. Therefore, our main objective is to attain good performance on the target domain by utilizing the source data or adapting classifiers trained in the source domain. In addition, how to effectively leverage unlabeled target data also remains an important issue for domain adaptation.

To this end, we propose in this paper a novel Semi-supervised Domain Adaptation with Subspace Learning (SDASL) framework for visual recognition. It attempts to bridge the domain gap by jointly constructing good subspace feature representations to minimize domain divergence and leveraging unlabeled target data in conjunction with labeled data. The training of SDASL is performed simultaneously by minimizing the classification error, preserving the structure relationships within and across domains, and restricting similarity defined on unlabeled target instances. In particular, the objective function of SDASL is composed of three components, i.e., *structural risk*, *structure preservation* within and across domains, and *manifold regularization*. Of the three, the former two aim to explore invariant low dimensional structures across domains and meanwhile minimizing the structural risk of the learnt models on the subspace, while the last exploits the intrinsic information in the target domain. After we obtain the predictive function on the subspace, the label of a new coming target instance can be determined accordingly.

Formally, suppose there are $l_s$ labeled samples in the source domain, represented as: $\mathbf{X}_S = \{\mathbf{x}_1^S, \mathbf{x}_2^S, \ldots, \mathbf{x}_{l_s}^S\}^\top \in R^{l_s \times d_s}$, where $d_s$ is the dimensionality of source data. Similarly, assume there are $l_t$ ($l_t \ll l_s$) labeled instances and $u_t$ unlabeled examples in the target domain, denoted as: $\mathbf{X}_T = \{\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_{l_t}^T\}^\top \in R^{l_t \times d_t}$ and $\mathbf{X}_T^U = \{\mathbf{x}_1^U, \mathbf{x}_2^U, \ldots, \mathbf{x}_{u_t}^U\}^\top \in R^{u_t \times d_t}$, respectively. The corresponding labels of $\mathbf{X}_S$ and $\mathbf{X}_T$ are given as column vectors $\mathbf{Y}_S \in \{-1,+1\}^{l_s}$ and $\mathbf{Y}_T \in \{-1,+1\}^{l_t}$, respectively. We project the original features into the low-dimensional subspace to explore the invariant structures across domains and minimize domain divergence. Accordingly, the linear predictive functions are defined as

$$\begin{cases} f_S(\mathbf{x}^S) = \mathbf{x}^S \mathbf{m}_S \mathbf{w}_S + b_S \\ f_T(\mathbf{x}^T) = \mathbf{x}^T \mathbf{m}_T \mathbf{w}_T + b_T \end{cases}, \tag{1}$$

where $\mathbf{w}_S, \mathbf{w_T} \in R^d$ and $b_S$, $b_T$ are the model weight and bias parameters, respectively. $\mathbf{m}_S$ and $\mathbf{m}_T$ are the feature mapping matrices, with $\mathbf{m}_S \in R^{d_S \times d}$ and $\mathbf{m}_T \in R^{d_T \times d}$, where $d$ is the dimension of the subspace.

The training objective of *structural risk* corresponds to an empirical risk minimization with a regularization penalty over the model parameters $\{\mathbf{w}_S, b_S, \mathbf{m}_S, \mathbf{w}_T, b_T, \mathbf{m}_T\}$ as

$$\begin{aligned} \min_{\substack{\mathbf{w}_S, b_S, \mathbf{m}_S \\ \mathbf{w}_T, b_T, \mathbf{m}_T}} & \|\mathbf{X}_S \mathbf{m}_S \mathbf{w}_S + b_S - \mathbf{Y}_S\|^2 + \alpha_S \|\mathbf{w}_S\|^2 \\ & + \|\mathbf{X}_T \mathbf{m}_T \mathbf{w}_T + b_T - \mathbf{Y}_T\|^2 + \alpha_T \|\mathbf{w}_T\|^2 \\ s.t. \quad & \mathbf{m}_S^\top \mathbf{m}_S = \mathbf{I}, \quad \mathbf{m}_T^\top \mathbf{m}_T = \mathbf{I} \end{aligned}, \tag{2}$$

where $\alpha_S$ and $\alpha_T$ are tradeoff parameters. The objective decomposes into the empirical risk with a least square loss of the labeled examples from both source and target domains, and the regularization penalty $\|\mathbf{w}_S\|^2$ and $\|\mathbf{w}_T\|^2$. The parameter $\alpha_S$ and $\alpha_T$ are the tradeoff parameters.

Deriving from the idea of seeking for a joint latent space that corresponding views are mapped to nearby locations in multi-view learning [2], we incorporate a discriminative regularization term in the objective function to take into account of the structure within and across domains. That is, the distance between the mappings in the latent subspace of the same category from source and target domains should be as small as possible. Technically, positives from both domains are represented as: $\mathbf{A} = \begin{bmatrix} \mathbf{X}_S^+ \mathbf{m}_S \\ \mathbf{X}_T^+ \mathbf{m}_T \end{bmatrix}$, where $\mathbf{X}_S^+$ and $\mathbf{X}_T^+$ denote the positives in the source and target domain, respectively. The distance between positives from source and target domains is measured by $tr(\mathbf{A}^\top \mathbf{L}_1 \mathbf{A})$, where $\mathbf{L}_1 = \mathbf{D}_1 - \mathbf{1}\mathbf{1}^\top$, $\mathbf{1}$ denotes a column vector with all 1 entries, and $\mathbf{D}_1$ is the diagonal matrix that contains the row sums of $\mathbf{1}\mathbf{1}^\top$. To learn a shared latent space across different domains, we integrate the *structure preservation* within and across domains as a regularization for domain adaptation.

*Manifold regularization* has been shown effective for semi-supervised learning [3]. This regularizer is to measure the smoothness of the predicted class labels along the inherent structure of unlabeled target data. In other words, the outputs of the predictive function are restricted to have similar values for similar examples. The estimation of the *manifold regularization* can be measured by the appropriate pairwise similarity $\mathbf{S}$ between the unlabeled target samples. By defining the graph Laplacian $\mathbf{L}_2 = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a diagonal matrix with its elements defined as $\mathbf{D}_{ii} = \sum_j \mathbf{S}_{ij}$, the regularization can be computed as $(\mathbf{X}_T^U \mathbf{m}_T \mathbf{w}_T)^\top \mathbf{L}_2 (\mathbf{X}_T^U \mathbf{m}_T \mathbf{w}_T)$.

The overall objective function integrates the above three components as

$$\begin{aligned} \min_{\substack{\mathbf{w}_S, b_S, \mathbf{m}_S \\ \mathbf{w}_T, b_T, \mathbf{m}_T}} & \|\mathbf{X}_S \mathbf{m}_S \mathbf{w}_S + b_S - \mathbf{Y}_S\|^2 + \alpha_S \|\mathbf{w}_S\|^2 \\ & + \|\mathbf{X}_T \mathbf{m}_T \mathbf{w}_T + b_T - \mathbf{Y}_T\|^2 + \alpha_T \|\mathbf{w}_T\|^2 \\ & + \gamma tr(\mathbf{A}^\top \mathbf{L}_1 \mathbf{A}) + \eta (\mathbf{X}_T^U \mathbf{m}_T \mathbf{w}_T)^\top \mathbf{L}_2 (\mathbf{X}_T^U \mathbf{m}_T \mathbf{w}_T) \\ s.t. \quad & \mathbf{m}_S^\top \mathbf{m}_S = \mathbf{I}, \quad \mathbf{m}_T^\top \mathbf{m}_T = \mathbf{I} \end{aligned}, \tag{3}$$

where $\gamma$ and $\eta$ are tradeoff parameters. To address the difficult non-convex problem (3) due to the orthogonal constrains, a gradient descent optimization procedure with curvilinear search [6] is used for a local optimal solution.

After the optimization, we can obtain the linear predictive functions defined in Eq.(1). Next, given a target test visual instance, $\hat{\mathbf{x}} \in R^{d_t}$, we compute the prediction values using the linear function as

$$f(\hat{\mathbf{x}}) = \hat{\mathbf{x}} \mathbf{m}_T \mathbf{w}_T + b_T. \tag{4}$$

The label of instance $\hat{\mathbf{x}}$ is $\text{sign}(f(\hat{\mathbf{x}}))$, where $\text{sign}(\bullet)$ is the signum function.

We empirically verify the merit of SDASL from both image-to-image and image-to-video transfer tasks, i.e., object recognition on the image dataset studied in [5], and video concept detection on the challenge TRECVID 2011 Semantic Indexing task [4] with the assistance of images from ImageNet [1]. Encouraging results validate our proposal and analysis.

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[2] Z. Fang and Z. Zhang. Discriminative feature selection for multi-view cross-domain learning. In *CIKM*, 2013.

[3] S. Melacci and M. Belkin. Laplacian support vector machines trained in the primal. *Journal of Machine Learning Research*, 12:1149–1184, 2011.

[4] Paul Over, George Awad, Jon Fiscus, Brian Antonishek, et al. Trecvid 2011-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *NIST TRECVID workshop*, 2011.

[5] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[6] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constrains. *Mathematical Programming*, 142:397–434, 2013.

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