Deep Transfer Metric Learning

Junlin Hu[1], Jiwen Lu[2], Yap-Peng Tan[1]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. [2]Advanced Digital Sciences Center, Singapore.

How to design a good similarity function plays an important role in many visual recognition tasks. Recent advances have shown that learning a distance metric directly from a set of training examples can usually achieve proposing performance than hand-crafted distance metrics [2, 3]. While many metric learning algorithms have been presented in recent years, there are still two shortcomings: 1) most of them usually seek a single linear distance to transform sample into a linear feature space, so that the nonlinear relationship of samples cannot be well exploited. Even if the kernel trick can be employed to addressed the nonlinearity issue, these methods still suffer from the scalability problem because they cannot obtain the explicit nonlinear mapping functions; 2) most of them assume that the training and test samples are captured in similar scenarios so that their distributions are assumed to be the same. This assumption doesn't hold in many real visual recognition applications, when samples are captured across datasets.

We propose a deep transfer metric learning (DTML) method for cross-dataset visual recognition. Our method learns a set of hierarchical nonlinear transformations by transferring discriminative knowledge from the labeled source domain to the unlabeled target domain, under which the inter-class variations are maximized and the intra-class variations are minimized, and the distribution divergence between the source domain and the target domain at the top layer of the network is minimized, simultaneously. Figure 1 illustrates the basic idea of the proposed method.

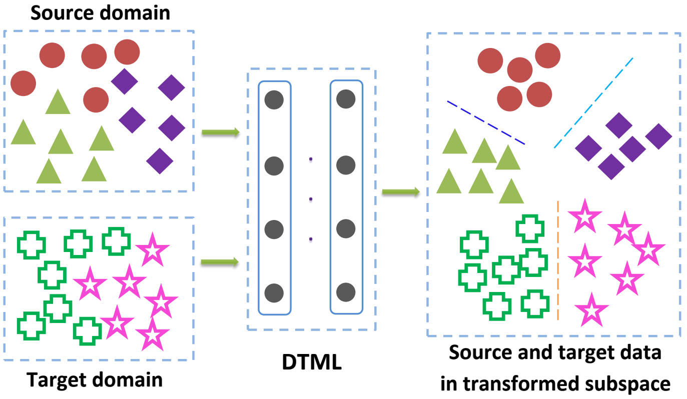

Figure 1: The basic idea of the proposed DTML method. For each sample in the training sets from the source domain and the target domain, we pass it to the developed deep neural network. We enforce two constraints on the outputs of all training samples at the top of the network: 1) the inter-class variations are maximized and the intra-class variations are minimized, and 2) the distribution divergence between the source domain and the target domain at the top layer of the network is minimized.

Deep Metric Learning. We construct a deep neural network to compute the representations of each sample $\mathbf{x}$. Assume there are $M+1$ layers of the network and $p^{(m)}$ units in the mth layer, where $m = 1, 2, \cdots, M$. The output of $\mathbf{x}$ at the mth layer is computed as:

$$f^{(m)}(\mathbf{x}) = \mathbf{h}^{(m)} = \varphi\left(\mathbf{W}^{(m)}\mathbf{h}^{(m-1)} + \mathbf{b}^{(m)}\right) \in \mathbb{R}^{p^{(m)}}, \tag{1}$$

where $\mathbf{W}^{(m)} \in \mathbb{R}^{p^{(m)} \times p^{(m-1)}}$ and $\mathbf{b}^{(m)} \in \mathbb{R}^{p^{(m)}}$ are the weight matrix and bias of the parameters in this layer; and φ is a nonlinear activation function which operates component-wisely, *e.g.*, *tanh* or *sigmoid* functions. The nonlinear mapping $f^{(m)}: \mathbb{R}^d \mapsto \mathbb{R}^{p^{(m)}}$ is a function parameterized by $\{\mathbf{W}^{(i)}\}_{i=1}^m$ and $\{\mathbf{b}^{(i)}\}_{i=1}^m$. For the first layer, we assume $\mathbf{h}^{(0)} = \mathbf{x}$.

For each pair of samples $\mathbf{x}_i$ and $\mathbf{x}_j$, they can be finally represented as $f^{(m)}(\mathbf{x}_i)$ and $f^{(m)}(\mathbf{x}_j)$ at the mth layer of our designed network, and their distance metric can be measured by computing the squared Euclidean distance between $f^{(m)}(\mathbf{x}_i)$ and $f^{(m)}(\mathbf{x}_j)$ at the mth layer:

$$d^2_{f^{(m)}}(\mathbf{x}_i, \mathbf{x}_j) = \left\| f^{(m)}(\mathbf{x}_i) - f^{(m)}(\mathbf{x}_j) \right\|_2^2. \tag{2}$$

Following the graph embedding framework, we enforce the marginal fisher analysis criterion [4] on the output of all training samples at the top layer and formulate a strongly-supervised deep metric learning method:

$$\min_{f^{(M)}} J = S_c^{(M)} - \alpha\, S_b^{(M)} + \gamma \sum_{m=1}^M \left(\|\mathbf{W}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2 \right), \tag{3}$$

where α ($\alpha > 0$) is a free parameter which balances the important between intra-class compactness and interclass separability; $\|\mathbf{Z}\|_F$ denotes the Frobenius norm of the matrix $\mathbf{Z}$; γ ($\gamma > 0$) is a tunable positive regularization parameter; $S_c^{(m)}$ and $S_b^{(m)}$ define the intra-class compactness and the interclass separability, which are defined as follows:

$$S_c^{(m)} = \frac{1}{Nk_1} \sum_{i=1}^N \sum_{j=1}^N P_{ij}\, d^2_{f^{(m)}}(\mathbf{x}_i, \mathbf{x}_j), \tag{4}$$

$$S_b^{(m)} = \frac{1}{Nk_2} \sum_{i=1}^N \sum_{j=1}^N Q_{ij}\, d^2_{f^{(m)}}(\mathbf{x}_i, \mathbf{x}_j), \tag{5}$$

where P_{ij} is set as one if $\mathbf{x}_j$ is one of k_1-*intra-class* nearest neighbors of $\mathbf{x}_i$, and zero otherwise; and Q_{ij} is set as one if $\mathbf{x}_j$ is one of k_2-*interclass* nearest neighbors of $\mathbf{x}_i$, and zero otherwise.

Deep Transfer Metric Learning. Given target domain data $\mathcal{X}_t$ and source domain data $\mathcal{X}_s$, their probability distributions are usually different in the original feature space when they are captured from different datasets. To reduce the distribution difference, we apply the Maximum Mean Discrepancy (MMD) criterion [1] to measure their distribution difference at the mth layer, which is defined as as follows:

$$D_{ts}^{(m)}(\mathcal{X}_t, \mathcal{X}_s) = \left\| \frac{1}{N_t} \sum_{i=1}^{N_t} f^{(m)}(\mathbf{x}_{ti}) - \frac{1}{N_s} \sum_{i=1}^{N_s} f^{(m)}(\mathbf{x}_{si}) \right\|_2^2. \tag{6}$$

By combining (3) and (6), we formulate DTML as the following optimization problem:

$$\begin{aligned} \min_{f^{(M)}} J = {} & S_c^{(M)} - \alpha\, S_b^{(M)} + \beta\, D_{ts}^{(M)}(\mathcal{X}_t, \mathcal{X}_s) \\ & + \gamma \sum_{m=1}^M \left(\|\mathbf{W}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2 \right), \end{aligned} \tag{7}$$

where β is a regularization parameter. We employ the stochastic gradient descent algorithm to obtain $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$.

[1] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Proc. NIPS*, pages 513–520, 2006.

[2] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009.

[3] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell. Distance metric learning with application to clustering with side-information. In *Proc. NIPS*, pages 505–512, 2002.

[4] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE T-PAMI*, 29(1):40–51, 2007.

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