Multi-Manifold Deep Metric Learning for Image Set Classification

Jiwen Lu[1], Gang Wang[1,2], Weihong Deng[3], Pierre Moulin[1,4], and Jie Zhou[5]
[1]ADSC, Singapore, [2]NTU, Singapore, [3]BUPT, China, [4]UIUC, USA, [5]Tsinghua University, China

Image set classification [3, 4] aims to recognize an object of interest from a set of image instances captured from varying viewpoints or under varying illuminations, which is different from the conventional image classification where each training and testing example is a single still image. Compared to a single image, an image set offers us more useful information to describe objects of interest. However, it is also more challenging to exploit discriminative information from image sets because there are usually larger intra-class variations within a set, which makes the classification task more difficult.

In this paper, we propose a new multi-manifold deep metric learning (MMDML) approach for image set classification, where the key idea of the proposed approach is shown in Figure 1. Given each image set, we first model it as a nonlinear manifold because manifolds can effectively describe the geometrical and structural information of image instances within image sets. Motivated by the fact that deep learning has demonstrated superb capability to model the nonlinearity of samples, we propose a MMDML method to learn multiple sets of nonlinear transformations, one set for each object class, to nonlinearly map multiple sets of image instances into a shared feature subspace, under which the manifold margin of different class is maximized, so that both discriminative and class-specific information can be exploited, simultaneously.

Let $X = [X_1, \cdots, X_c, \cdots, X_C]$ be the training set of C different classes, where $X_c = [x_{c1}, x_{c2}, \cdots, x_{ci}, \cdots, x_{cN_c}] \in \mathbb{R}^{d \times N_c}$ denotes the cth image set, $1 \le c \le C$, N_c is the number of samples in this image set, x_{ci} is the ith image in this image set, and d is the feature dimension of each image. As shown in Figure 1, we construct a deep neural network for each class, and pass the image set X_c into the cth network. Assume there are $L+1$ layer in the work, and d_c^l denote the number of nodes in lth layer of the cth network, where $1 \le l \le L$. For the image x_{ci}, its output of the first layer in the cth network is computed as: $h_{ci}^1 = s(W_c^1 x_{ci} + b_c^1)$, where W_c^1 is the projection matrix and b_c^1 is the bias vector to be learned in the first layer of the cth network, s is a nonlinear active function which applies component-wisely, which is widely used in previous deep learning algorithms [1, 2]. Then, the output of the first layer of this network is used as the input of the second layer. Therefore, the output of the second layer is $h_{ci}^2 = s(W_c^2 h_{ci}^1 + b_c^2)$, where W_c^2 is the projection matrix and b_c^2 is the bias vector to be learned in the second layer of the cth network, respectively. Similarly, the output for the lth layer is $h_{ci}^l = s(W_c^l h_{ci}^{l-1} + b_c^l)$, and for the top layer is: $h_{ci}^L = s(W_c^L h_{ci}^{L-1} + b_c^L)$, where W_c^L is the projection matrix and b_c^L is the bias vector to be learned for the top layer of the cth network, respectively.

For each sample h_{ci}^L from the cth manifold, we compute two squared distances $D_1(h_{ci}^L)$ and $D_2(h_{ci}^L)$, which measure the dissimilarity between this sample and its intra-class and inter-class neighbors as follows:

$$D_1(h_{ci}^L) = \frac{1}{K_1} \sum_{p=1}^{K_1} \|h_{ci}^L - h_{cip}^L\|_2^2 \quad (1)$$

$$D_2(h_{ci}^L) = \frac{1}{K_2} \sum_{q=1}^{K_2} \|h_{ci}^L - h_{ciq}^L\|_2^2 \quad (2)$$

where h_{cip}^L and h_{ciq}^L are the feature representations at the top layer of the pth intra-manifold and qth inter-manifold neighbors, K_1 and K_2 are two parameters to define the neighborhood size, respectively.

Let $f_c = \{W_c^1, W_c^2, \cdots, W_c^L, b_c^1, b_c^2, \cdots, b_c^L\}$ be the parameters of the cth network, we formulate the following optimization problem to maximize the margin between the cth manifold and other manifolds:

$$\min_{f_c} \sum_{i=1}^{N_c} \left(D_1(h_{ci}^L) - D_2(h_{ci}^L) \right) \quad (3)$$

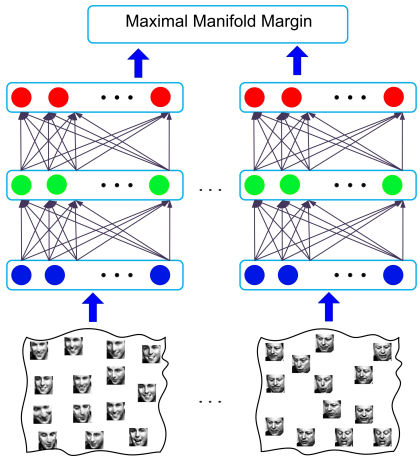

Figure 1: The basic idea of our proposed image set classification approach.

By applying the criterion in (3) on each sample from all image sets in the training set, we formulate the following optimization problem for our MMDL modal:

$$\min_{f_1, f_2, \cdots, f_C} \sum_{c=1}^{C} \sum_{i=1}^{N_c} g\left(D_1(h_{ci}^L) - D_2(h_{ci}^L) \right) + \frac{\lambda}{2} \sum_{c=1}^{C} \sum_{l=1}^{L} \left(\|W_c^l\|_F^2 + \|b_c^l\|_2^2 \right) \quad (4)$$

where the first term maximizes the manifold margins to exploit the discriminative information for classification, and the second term regularizes the parameters of these networks, λ is a parameter to balance the contributions of different terms, and $g(a)$ is a generalized logistic loss function to smoothly approximate the hinge loss function $a = \max(a, 0)$.

Since h_{cip}^L and h_{ciq}^L depend on the network parameters $W_c^1, W_c^2, \cdots, W_c^L$, and $b_c^1, b_c^2, \cdots, b_c^L$, which are also to be learned in our method, the optimization function defined in (4) is an egg and chicken problem. To address this, we develop an iterative algorithm to obtain a local optimal solution. Specifically, we first initialize the network parameters with appropriate values and compute the intra-class and inter-class neighbors, then, we update these parameters by (4) until convergence.

Implementation of this method by the stochastic sub-gradient descent algorithm is described in the paper, as are the details of the procedure of the algorithm. Our conclusion is that multi-manifold deep metric learning is an effective approach to image set classification.

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[2] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[3] Jiwen Lu, Gang Wang, and Pierre Moulin. Image set classification using multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, pages 329–336, 2013.

[4] Jiwen Lu, Gang Wang, Weihong Deng, and Pierre Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *ECCV*, pages 265–280, 2014.

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

In the training set, there could be multiple image sets for some classes. For this case, we merge image sets from the same person into a large image set to learn the class-specific deep model.