TASK-DRIVEN DEEP TRANSFER LEARNING FOR IMAGE CLASSIFICATION

Zhengming Ding[†], *Nasser M Nasrabadi*[‡] *and Yun Fu*[†♯]

[†] Department of Electrical and Computer Engineering, Northeastern University, USA
[‡]Lance Computer Science and Electrical Engineering, West Virginia University, USA
[♯] College of Computer and Information Science, Northeastern University, USA

ABSTRACT

Transfer learning tends to be a powerful tool that can mitigate the divergence across different domains through knowledge transfer. Recent research efforts on transfer learning have exploited deep neural network (NN) structures for discriminative feature representation to better tackle cross-domain disparity. However, few of these techniques are able to jointly learn deep features and train a classifier in a unified transfer learning framework. To this end, we design a task-driven deep transfer learning framework for image classification, where the deep feature and classifier are obtained simultaneously for optimal classification performance. Therefore, the proposed deep structure can generate more discriminative features by using the classifier performance as a guide. Furthermore, the classifier performance is increased since it is optimized on a more discriminative deep feature. The developed supervised formulation is a task-driven scheme, which will provide better learned features for the classification task. By giving pseudo labels for target data, we can facilitate the knowledge transfer from source to target through the deep structures. Experimental results witness the superiority of our proposed algorithm by comparing with other ones.

Index Terms— task-driven, deep transfer learning

1. INTRODUCTION

In typical pattern recognition problems, there is always a situation that we have plenty of unlabeled data while there are limited or even no labeled data for training from the test (target) domain. Transfer learning [1] has been demonstrated as a promising technique to address such difficulty by borrowing knowledge from other well-learned source domains, which might lie in a different distribution than the target domain. Recent research on transfer learning have witnessed appealing performance by seeking a common feature space where knowledge from source can be well transferred to assist the recognition task in target domain [2, 3, 4, 5, 6, 7].

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, NPS award N00244-15-1-0041, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

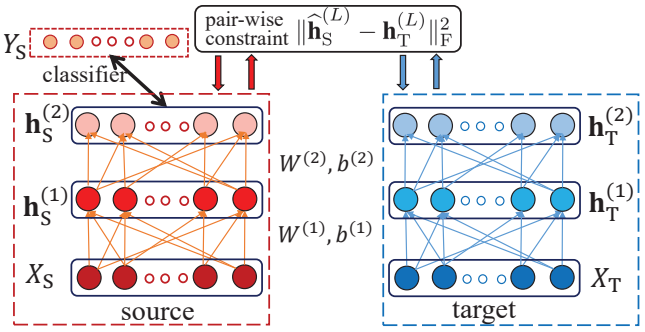

Fig. 1. Illustration of our proposed $(L+1)$-layer coupled deep neural network (here $L = 2$). Deep structures are built to learn deep features for source X_{S} and target domains X_{T}, which share the same networks weights $\{W^{(l)}, b^{(l)}\}(1 \leq l \leq L)$. A pair-wise constraint is developed to couple the similar pair of source and target in the $(L+1)$-th layer to transfer knowledge. Moreover, a classifier is jointly trained on source data, where Y_{S} is the label matrix of source data.

Recent research activities on deep structure learning have attracted increasing interests in capturing a better feature representation, because discriminative knowledge can be embedded in multiple layers of the feature hierarchy [5, 6, 8]. Most recently, the concept of deep learning has been incorporated into transfer learning scenarios, which aims to align different domains and learn deep structural features simultaneously [7]. In this way, deep transfer learning technique obtains a set of highly discriminative features across the two domains in order to alleviate the recognition task in the target domain.

Moreover, recent task-driven formulations have achieved very promising performance in various classification tasks through learning the features and classifier parameters in a unified framework [9, 10]. Therefore, supervised information can be utilized to minimize a misclassification cost and at the same time capture optimal features. Along this line, Zhuang et al. [7] proposed a supervised transfer learning with deep autoconders, where label information of the source domain is encoded using a softmax regression model. However, the feature learning structure is shallow only one-layer so that it cannot exploit the rich information behind the data.

To this end, we develop a Task-driven Deep Transfer

Learning algorithm (TDTL) in order to jointly learn a deep structure of feature representation and simultaneously train a classifier in a unified framework (Fig. 1). Our major contributions are summarized as:

- Deep transfer structures are designed to capture the rich information across source and target domains. Through refining features for source and target in a layer-wise fashion, our proposed algorithm can preserve more essential information to the target domain.
- A task-driven scheme is developed to train a classifier jointly with a deep structure feature learning procedure, which can feedback back its classification error to refine the parameters of the deep structures. In this way, the classifier and deep feature learning parameters are jointly trained in a unified framework. Moreover, the discriminative deep features could generate a more accurate classifier.
- By providing pseudo labels for the target, a novel constraint is incorporated to couple the pair of source and target in the deep structure, therefore, the coupled deep NN would capture more intrinsic information from source to target.

2. RELATED WORK

In this section, we talk about two lines of related work to ours.

Transfer Learning is a promising technique to handle the different distribution issue between the source domain and the target domain. In the recent years, a bunch of transfer learning techniques [2, 3, 4, 11] were designed,w which could be separated into two strategies, one is to reweight instances [11] and the other is to seek common features [3, 5]. Our algorithm follows in the second line and we aim to seek a common feature space to transfer well-learned source knowledge to the target domain. While some promising results are achieved through those transfer learning algorithms, most approaches only adopt linear mapping or non-linear mappings with kernel learning theory to build a gap across the source and target domains, therefore they are not effective enough to facilitate the knowledge transfer when there is a large distribution divergence between two domains. In our paper, we adopt the deep structure idea and propose a task-driven transfer learning to jointly learn the classifier parameters and deep features for source and target domains.

Deep Structure Learning has recently attracted lots of attention in pattern recognition, because of its appealing superiority in many real-world applications [5, 7]. In general, deep structure learning tends to build a hierarchical structure to extract discriminative features directly from original data. A lot of deep models were developed recently, which can be categorized into several classes, e.g., deep convolutional neural networks, deep denoising auto-encoder, deep belief networks. Most recently, the concept of deep structure is incorporated into transfer learning to uncover the rich information across domains [2, 5, 6]. However, most of the current deep transfer learning methods only focus on seeking better deep features across two domains. In our work, we also adopt the idea of deep transfer learning, however, our method jointly learns a classifier and transfers knowledge from source to target via a unified deep coupled NN, where we build a task-driven deep structure to capture more discriminative features across the two domains.

3. THE PROPOSED ALGORITHM

In this section, we present the details of our task-driven deep transfer learning algorithm with its training procedure.

3.1. Deep Neural Network Revisit

Deep neural network aims to compute a compact representation from each data point $x \in \mathbb{R}^{d_1}$ through putting it into multi-layer nonlinear mapping structure. The major merit of this structure is that the nonlinear transformation function is promising in explicitly achieving better feature representations. Suppose the network has $L+1$ layers with d_l units for the l-th layer, where $l \in [1, L]$. So we could have the l-th layer output for x as:

$$f^{(l)}(x) = \mathbf{h}^{(l)} = \varphi(W^{(l)}\mathbf{h}^{(l-1)} + b^{(l)}), \tag{1}$$

where $W^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ and $b^{(l)} \in \mathbb{R}^{d_l}$ are the parameters in l-th layer for transformations and bias; $\mathbf{h}^{(l)}$ is the l-th hidden layer and $\mathbf{h}^{(0)} = x$; φ denotes the nonlinear function operating in component-wise way. The overall nonlinear mapping $f^{(L)} : \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_L}$ is a bunch of nonlinear functions with those parameters $\{W^{(i)}\}_{i=1}^{L}$ and $\{b^{(i)}\}_{i=1}^{L}$.

3.2. Task-driven Deep Transfer Learning

Given a set of target domain $X_{\mathrm{T}} = \{x_{\mathrm{T},1}, \cdots, x_{\mathrm{T},n_{\mathrm{T}}}\}$ with n_{T} unlabeled data samples and a set of source domain $\{X_{\mathrm{S}}, Y_{\mathrm{S}}\} = \{(x_{\mathrm{S},1}, y_{\mathrm{S},1}), \cdots, (x_{\mathrm{S},n_{\mathrm{S}}}, y_{\mathrm{S},n_{\mathrm{S}}})\}$ with n_{S} labeled data samples and Y_{S} is the label matrix.

For transfer learning, it is very essential to transfer knowledge from source to target when learning the deep features. Current transfer learning on class-wise domain adaptation achieve appealing results [12], therefore, it is very important to incorporate class-wise adaptation into deep structure learning. However, we assume target data are totally unlabeled ahead of time. In this way, we design to predict the unlabeled data with a classifier learned from labeled source data and assign the target data with pseudo labels. Specifically, we adopt the nearest neighbor classifier (NNC).

Different from previous class-wise adaptation, we desire to find the corresponding pairs from source to target (pairwise adaptation), then minimize the distance between the deep features from each pair, see Fig. 1. To this end, we propose to couple the the output of two domains as follows:

$$\|f^{(L)}(\widehat{X}_{\mathrm{S}}) - f^{(L)}(X_{\mathrm{T}})\|_{\mathrm{F}}^2, \quad (2)$$

where $\|\cdot\|_{\mathrm{F}}^2$ is matrix Frobenius norm. $\widehat{X}_{\mathrm{S}}$ is the subset of X_{S}, which is selected as the nearest pair for the target domain. That is, each target sample is associated with one source sample. Since we don't have access to the true label of the target domain, we intend to adopt the assigned labels for target domain through some supervised classifier trained from the source data, e.g., SVM. The pair-wise constrained deep networks assume the target and source share the same networks weights [5, 7], since we aim to explore a latent domain where source and target can be well-coupled to the similar distribution with the same networks weights. This is very similar to transfer subspace learning methods [13, 14], which assumes the target and source have the same or similar distribution in one common latent subspace.

Furthermore, to achieve an discriminative classifier with respect to the original features, we employ a task-driven approach in order to minimize the classification error, similar to [9]. We incorporate the deep feature $f^{(L)}(x_{\mathrm{S},i})$ as the input to train the classifier parameter $\mathcal{A}$. The function to be minimized is as follows:

$$\min_{\mathcal{A}} \mathcal{L}(Y_{\mathrm{S}}, \mathcal{A}, f^{(L)}(X_{\mathrm{S}})) = \min_{\mathcal{A}} \sum_{i=1}^{n_{\mathrm{S}}} \mathcal{L}(y_{\mathrm{S},i}, \mathcal{A}, f^{(L)}(x_{\mathrm{S},i})),$$

where $\mathcal{L}(y_{\mathrm{S},i}, \mathcal{A}, f^{(L)}(x_{\mathrm{S},i}))$ is the classification error for a training pair $(x_{\mathrm{S},i}, y_{\mathrm{S},i})$. For simplicity, we deploy a linear regression to measure the classification error as:

$$\mathcal{L}(y_{\mathrm{S},i}, \mathcal{A}, f^{(L)}(x_{\mathrm{S},i})) = \frac{1}{2}\|y_{\mathrm{S},i} - \mathcal{A}f^{(L)}(x_{\mathrm{S},i})\|_2^2. \quad (3)$$

To sum up, we develop a framework to simultaneously learn a classifier and a deep structure by combining (2) and (3) into a unified objective function J as follows:

$$J = \frac{1}{2}\|Y_{\mathrm{S}} - \mathcal{A}f^{(L)}(X_{\mathrm{S}})\|_{\mathrm{F}}^2 + \lambda\|f^{(L)}(\widehat{X}_{\mathrm{S}}) - f^{(L)}(X_{\mathrm{T}})\|_{\mathrm{F}}^2 \quad (4)$$

where λ is the balanced parameter. We adopt *sigmoid* functions for the non-linear function in this paper.

3.3. Training the Proposed TDTL

To address the nonlinear problem in (4), we propose to adopt the stochastic sub-gradient descent algorithm to optimize the weight, bias and regression variables $W^{(l)}$, $b^{(l)}$ and $\mathcal{A}$. Specifically, we could calculate the gradients of J in (2) with respect to $W^{(l)}$, $b^{(l)}$ and $\mathcal{A}$ in the following way:

Updating $W^{(l)}$:

$$\frac{\partial J}{\partial W^{(l)}} = \mathcal{S}^{(l)}\mathbf{h}_{\mathrm{S}}^{(l-1)^{\mathrm{T}}} + \lambda\widehat{\mathcal{S}}^{(l)}\widehat{\mathbf{h}}_{\mathrm{S}}^{(l-1)^{\mathrm{T}}} + \lambda\mathcal{T}^{(l)}\mathbf{h}_{\mathrm{T}}^{(l-1)^{\mathrm{T}}}, \quad (5)$$

where $\widehat{\mathbf{h}}_{\mathrm{S}}^{(l-1)} = f^{(l-1)}(\widehat{X}_{\mathrm{S}})$, $\mathbf{h}_{\mathrm{S}}^{(l-1)} = f^{(l-1)}(X_{\mathrm{S}})$ and $\mathbf{h}_{\mathrm{T}}^{(l-1)} = f^{(l-1)}(X_{\mathrm{T}})$.

Updating $b^{(l)}$:

$$\frac{\partial J}{\partial b^{(l)}} = \mathcal{S}^{(l)} + \lambda\widehat{\mathcal{S}}^{(l)} + \lambda\mathcal{T}^{(l)}, \quad (6)$$

where we define the variables in Eqs. (5),(6) in the following:
$\mathcal{S}^{(L)} = \mathcal{A}^{\mathrm{T}}(Y_{\mathrm{S}} - \mathcal{A}\mathbf{h}_{\mathrm{S}}^{(L)}) \odot \varphi'(Z_{\mathrm{S}}^{(L)})$,
$\widehat{\mathcal{S}}^{(L)} = (\widehat{\mathbf{h}}_{\mathrm{S}}^{(L)} - \mathbf{h}_{\mathrm{T}}^{(L)}) \odot \varphi'(\widehat{Z}_{\mathrm{S}}^{(L)})$,
$\mathcal{T}^{(L)} = (\mathbf{h}_{\mathrm{T}}^{(L)} - \widehat{\mathbf{h}}_{\mathrm{S}}^{(L)}) \odot \varphi'(Z_{\mathrm{T}}^{(L)})$,
$\mathcal{S}^{(l)} = (W^{(l+1)^{\mathrm{T}}}\mathcal{S}^{(l+1)}) \odot \varphi'(Z_{\mathrm{S}}^{(l)})$,
$\widehat{\mathcal{S}}^{(l)} = (W^{(l+1)^{\mathrm{T}}}\widehat{\mathcal{S}}^{(l+1)}) \odot \varphi'(\widehat{Z}_{\mathrm{S}}^{(l)})$,
$\mathcal{T}^{(l)} = (W^{(l+1)^{\mathrm{T}}}\mathcal{T}^{(l+1)}) \odot \varphi'(Z_{\mathrm{T}}^{(l)})$,
where the operation $\odot$ denotes the element-wise multiplication, and $Z_{\mathrm{S}}^{(l)}$, $\widehat{Z}_{\mathrm{S}}^{(l)}$ and $Z_{\mathrm{T}}^{(l)}$ are given as follows: $Z_{\mathrm{S}}^{(l)} = W^{(l)}\mathbf{h}_{\mathrm{S}}^{(l-1)} + b^{(l)}$, $\widehat{Z}_{\mathrm{S}}^{(l)} = W^{(l)}\widehat{\mathbf{h}}_{\mathrm{S}}^{(l-1)} + b^{(l)}$, $Z_{\mathrm{T}}^{(l)} = W^{(l)}\mathbf{h}_{\mathrm{T}}^{(l-1)} + b^{(l)}$.

Updating $\mathcal{A}$:

$$\frac{\partial J}{\partial \mathcal{A}} = (Y_{\mathrm{S}} - \mathcal{A}\mathbf{h}_{\mathrm{S}}^{(L)})\mathbf{h}_{\mathrm{S}}^{(L)^{\mathrm{T}}}. \quad (7)$$

We list the gradient steps in the above, then we adopt L-BFGS optimizer [15] to optimize this unconstrained problem (4), since L-BFGS can be stabilized real-world large-scale dataset.

To classify a new testing data, we could apply the learned deep networks to extract the features, then input it to the regression classifier $y_{test} = \mathcal{A}x_{test}$ to achieve the label; also we can apply the deep features to train a classifier model, then predict the labels of real testing data.

4. EXPERIMENTAL RESULTS

In the experiments, we evaluate on two datasets. We first provide the data description and experimental settings. Then we present the comparison experiments with several competing algorithms. Finally, we evaluate our proposed deep NN.

4.1. Datasets & Experimental Settings

FLIR dataset contains two subsets: Sig and Roi of military vehicles. The images of Sig dataset were captured under very favorable environments. Specifically, the Sig dataset is consisted of around 874 to 1468 image samples for each target. The Roi dataset contains the images which were collected in less favorable environments compared to Sig. For example, Roi images show various weather environments and changing backgrounds; hence, these two datasets are very challenging. There are five common classes between Sig and Roi (Fig. 2). The images of both datasets are cropped to the size 40×75 and the pixel value is adopted as the input. In the training stage, we adopt Sig as the source, while half of Roi as the target and the rest half of Roi as the testing data.

Fig. 2. Five common Target classes within Sig and Roi in FLIR datasets: BMP, HMMWV, M35, M60, T72.

ATR dataset contains two modalities: long wave (LW) and middle wave (MW) IR imagery, each has 6 classes, 614 samples. The images are also cropped and resized to 40×75. We still use the pixel value as the input of the coupled deep NN. In the training stage, we use LW/MW modality as the source domain, and half of MW/LW modality as the target domain. Then the remaining half of MW/LW is for testing. Since the dataset involves two modalities, this is heterogenous transfer learning.

For both datasets, the source domain is well-labeled while the target is unlabeled. For our method, we can use two classification strategies for testing. One strategy is to use the deep structure to extract the features, then NNC to do classification, which is not task-driven and named as NDTL. The other strategy is to use a linear regression model, that is, we can apply the regression matrix $\mathcal{A}$ (Eq. (3)) to the extracted features to achieve the final classification results, named as TDTL.

4.2. Comparison Results

We compare our method with some competing transfer learning methods and subspace methods, LDA [16], JDA [17], GFK [18], DASA [19], mSDA [2] and LTSL [14]. Among them, only LDA is a traditional supervised subspace learning algorithm, so that we train a subspace on labeled source data then predict the labels of the real testing data. For other transfer learning methods, we train the model on source and target, then apply the model to label the real testing data. NNC is adopted for comparison algorithms in the testing stage. For our method, we adopt three-layer NN with layer-size as 2000 for the second layer and 200 for the third layer.

From our results (Table 1), we observe that the two proposed models outperform other comparisons. Transfer learning methods can improve the classification performance more than the traditional methods, e.g., LDA. Most of comparison transfer learning methods are shallow-layer ones thus they cannot uncover the rich information across the two domains. mSDA is also a deep learning method, which outperforms other methods, hence deep structure is definitely essential in learning better feature representation for source and target. However, mSDA only focus on learning a deep structure, therefore, it cannot make full use of the label information. Our method provides a joint classifier learning and deep feature learning task in a unified fashion.

4.3. Deep NN Architecture Evaluation

In this part, we verify one property of our coupled NN, i.e., the influence of dimension in each layer. From our experiments, we observe that three layers are usually enough to obtain the best results. So we adopt three-layer scheme and set the size of the second layer as 2000, then evaluate different sizes of the third layer from 50 to 400. The dataset we use is the FLIR datasets. The results are presented in Fig. 3.

Table 1. Recognition rate (%) of seven different algorithms on two datasets, where Case 1 is for FLIR, Case 2 means that LW of ATR is the source, while Case 3 denotes that MW of ATR is adopted as the source.

	Case 1	Case 2	Case 3
LDA [16]	37.23±2.05	48.86±2.03	51.47±1.02
GFK [18]	49.66±2.15	83.77±1.31	77.60±2.12
DASA [19]	49.77±2.04	82.14±1.04	76.95±1.25
LTSL [14]	49.43±2.02	82.67±2.02	77.34±1.14
JDA [17]	50.32±1.18	72.15±2.15	77.50±3.02
mSDA [2]	53.42±1.28	85.32±1.25	84.87±2.13
NDTL	56.99±2.12	88.96±2.18	89.24±3.05
TDTL	**57.44±1.07**	**89.26±2.09**	**90.15±2.05**

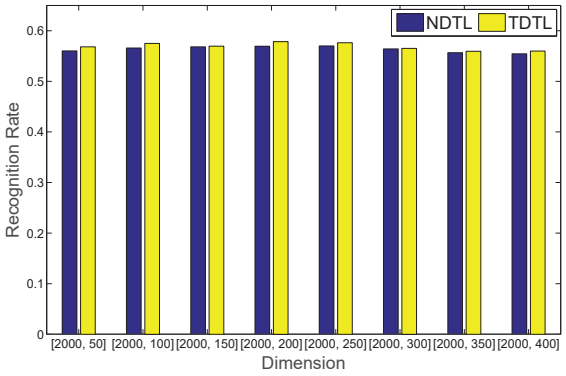

Fig. 3. Recognition results of NDTL and TDTL on FLIR dataset for different dimension cases.

From Fig. 3, we notice that different dimensions would produce different results and generally the case when the third layer is 200, it would generate the best results. By cross-validation, we have evaluated several different architectures, that is different dimensions for the second layer and more hidden layers, and we found that three layers with the second layer set as 2000 would produce the best results.

5. CONCLUSION

In this paper, we designed task-driven deep structures for better knowledge transfer. Specifically, a classifier and deep NN structures were jointly learned such that the classifier guided the deep feature learning in order to generate a more discriminative non-linear features optimized for the classifier. Through providing pseudo labels for target domain, our deep structures bridged the gap across two domains, and therefore it could transfer more discriminative information to the target domain. Experiments on two datasets showed its effectiveness by comparing with other algorithms.

6. REFERENCES

[1] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE TKDE*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha, "Marginalized denoising autoencoders for domain adaptation," *ICML*, pp. 767–774, 2012.

[3] Zhengming Ding, Ming Shao, and Yun Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *AAAI*, 2014, pp. 1192–1198.

[4] Sumit Shekhar, Vishal M Patel, Hien V Nguyen, and Rama Chellappa, "Generalized domain-adaptive dictionaries," in *CVPR*, 2013, pp. 361–368.

[5] Junlin Hu, Jiwen Lu, and Yap-Peng Tan, "Deep transfer metric learning," in *CVPR*, 2015, pp. 325–333.

[6] Zhengming Ding, Ming Shao, and Yun Fu, "Deep low-rank coding for transfer learning," in *IJCAI*, 2015, pp. 3453–3459.

[7] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He, "Supervised representation learning: Transfer learning with deep autoencoders," in *IJCAI*, 2015, pp. 4119–4125.

[8] Ming Shao, Zhengming Ding, and Yun Fu, "Sparse low-rank fusion based deep features for missing modality face recognition," in *FG*. IEEE, 2015, vol. 1, pp. 1–6.

[9] Julien Mairal, Francis Bach, and Jean Ponce, "Task-driven dictionary learning," *IEEE TPAMI*, vol. 34, no. 4, pp. 791–804, 2012.

[10] Soheil Bahrampour, Nasser M Nasrabadi, Asok Ray, and Kenneth W Jenkins, "Kernel task-driven dictionary learning for hyperspectral image classification," *ICASSP*, 2015.

[11] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu, "Boosting for transfer learning," in *ICML*. ACM, 2007, pp. 193–200.

[12] Mingsheng Long, Jianmin Wang, Guiguang Ding, S Pan, and P Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE TKDE*, vol. 26, no. 5, pp. 1076–1089, 2014.

[13] Si Si, Dacheng Tao, and Bo Geng, "Bregman divergence -based regularization for transfer subspace learning," *IEEE TKDE*, vol. 22, no. 7, pp. 929–942, 2010.

[14] Ming Shao, Dmitry Kit, and Yun Fu, "Generalized transfer subspace learning through low-rank constraint," *IJCV*, pp. 1–20, 2014.

[15] Jorge Nocedal, "Updating quasi-newton matrices with limited storage," *Mathematics of computation*, vol. 35, no. 151, pp. 773–782, 1980.

[16] Peter N. Belhumeur, João P Hespanha, and David J Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE TPAMI*, vol. 19, no. 7, pp. 711–720, 1997.

[17] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu, "Transfer feature learning with joint distribution adaptation," in *ICCV*, 2013, pp. 2200–2207.

[18] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *CVPR*, 2012, pp. 2066–2073.

[19] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *ICCV*, 2013, pp. 2960–2967.