SEMI-SUPERVISED DOMAIN ADAPTATION VIA CONVOLUTIONAL NEURAL NETWORK

Pengcheng Liu, Cheng Cheng, Youji Feng, Xiaohu Shao, Xiangdong Zhou*

Intelligent Media Technique Research Center,
Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences

ABSTRACT

Semi-supervised visual domain adaptation is devoted to adapting a model learned in source domain to target domain where there are only a few labeled samples. In this paper, we propose a semi-supervised cross-domain image recognition method which unifies the feature learning and recognition model training into a convolutional neural network framework. Based on a few labeled samples and massive unlabeled samples in the source and target domains, we specially design three branches for class label, domain label and similarity label prediction which simultaneously optimizes the network to generate image features that are domain invariance and inter-class discriminative. Experimental results demonstrate that our method is effective for learning robust cross-domain image recognition model, and achieves the state-of-the-art performance on the widely used visual domain adaptation benchmark.

***Index Terms*—** Domain adaptation, feature learning, image recognition

1. INTRODUCTION

In practice, training (source domain) and test (target domain) samples are always drawn from different distributions, which is usually caused by the situation that samples in the source and target domains are acquired under different sets of background, lighting, view point, resolution, etc. Cross-domain image recognition [1] is devoted to minimizing the distribution difference between the source and target domains so that the pre-trained recognition model could be adapted to the target domain without much performance loss. In order to reduce the distribution difference, most of the previous work [2, 3, 4, 5, 6, 7] focus on transforming the given hand-craft feature presentation of samples in the source and target domains to a new kind of domain-invariant feature representation. However, due to the limited representation ability of the hand-craft features, it is hard to learn an optimal transformation for feature presentation of samples in both source and target domains.

In recent years, neural network based feature learning methods [8, 9, 10, 11] have shown great success in cross-domain image recognition. Based on a deep convolutional neural network (DCNN) model that is pre-trained with a large scale of data in ImageNet [12], the study of [13] shows that the extracted deep features of samples from different domains are more similar with each other while comparing with the hand-craft features. Since it is a hard task for learning an DCNN with a small-scale dataset, many Fine-tuning technologies [14, 15, 16] are proposed to adapt the parameters of a pre-trained DCNN model to a new task. The Fine-tuning methods need a large number of labeled samples in the target domain, and show good performance on cross-domain image recognition. However, in practice of cross-domain image recognition, there is no labeled (unsupervised) or only a few labeled samples (semi-supervised) in the target domain, which is impossible for Fine-tuning method. Inspired by work [17, 18], we propose to learn an DCNN model for semi-supervised cross-domain image recognition even if there are only a few labeled samples in the target domain.

In our proposed semi-supervised cross-domain DCNN method, loss functions for class label, domain label and similarity label predictors are optimized with stochastic gradient descent (SGD) simultaneously to ensure that the learned image features are domain invariance and inter-class discriminative. First, the learned features are inter-class discriminative by minimizing the loss of class label prediction based on the labeled samples. Second, the unlabeled samples are useful for learning domain invariance features by maximizing the loss of domain label prediction, i.e., make the feature distribution of different domains be similar with each other. In addition, the learned features are further enhanced to be domain invariance and inter-class discriminative by minimizing the loss of similarity label prediction for samples from different domains. Extensive experimental results demonstrate that our method is beneficial to reducing the domain difference, and shows the state-of-the-art performance on the widely used domain adaptation benchmark dataset.

2. PROPOSED METHOD

In order to learn features that are domain invariance and inter-class discriminative, we propose a semi-supervised cross-domain DCNN method which simultaneously optimizes the loss functions for class label, domain label and

* The corresponding author (E-mail: chengcheng@cigit.ac.cn).

similarity label predictors based on the labeled and unlabeled samples in the source and target domains. The framework of our method is summarized in Figure 1. The rest of this section will describe our method in detail.

2.1. Network model

In this paper we mainly focus on the semi-supervised cross-domain image recognition, where the class labels Y^s of samples X^s in source domain (S) are all known while the data in target domain (T) consist of only a few labeled samples (X^{lt}, Y^{lt}) and massive unlabeled samples X^{ut}. The label set of S (Y^s) and T (Y^t) are the same, i.e., $Y^s = Y^t$, but the data distribution of S ($P(X^s)$) and T ($P(X^t)$) are different, i.e., $P(X^s) \neq P(X^t)$. Since there are only a few labeled samples in T, it is hard to learn a robust model for cross-domain image recognition.

In this paper, we take full advantage of all the label and unlabeled samples in S and T. First, we construct a huge number of similar and dissimilar sample pairs based on the labeled samples in S and T. Let s_{ij} denotes the similarity label of one sample pair, if one sample (x_i^s, y_i^s) in S has same class label with one sample (x_j^{lt}, y_j^{lt}) in T, i.e, $y_i^s = y_j^{lt}$, the similarity of the sample pair (x_i^s, x_j^{lt}) is $s_{ij} = 1$, otherwise, $s_{ij} = 0$. Second, let d_i denotes the domain label of samples, the domain label of the sample x_i^s in S is $d_i = 0$. On the contrary, the domain label of the sample x_i^t in T is $d_i = 1$.

During the process of forward propagation, our network mainly predicts the class labels (y_i, y_j), domain labels (d_i, d_j) and similarity label s_{ij} for one sample pair (x_i^s, x_j^t), see Figure 1. It consists of four parts. First, for one sample pair (x_i^s, x_j^t), the feature extractor G_f outputs their features $(f_i^s, f_j^t) = G_f(x_i^s, x_j^t; \theta_f)$, Here, θ_f denotes the parameters of feature extraction layers. Then based on f_i^s and f_j^t, the class label predictor G_y outputs their labels $(y_i, y_j) = G_y(f_i^s, f_j^t; \theta_y)$, here let θ_y denotes the parameters of class label prediction layers. In addition, the similarity predictor G_s with layer parameters θ_s outputs the similarity s_{ij} of that pair, i.e., $s_{ij} = G_s(f_i^s, f_j^t; \theta_s)$. Finally, let θ_d denotes the parameters of domain label prediction layers, the predictor G_d output the domain label (d_i, d_j) for the pair (f_i^s, f_j^t), i.e., $(d_i, d_j) = G_d(f_i^s, f_j^t; \theta_d)$.

At the training stage, we first aim to minimize the loss L_y of class label predictor $G_y(\cdot; \theta_y)$ based on the labeled samples X^s and X^{lt} in both domains, so that the learned features are inter-class discriminative enough. At the same time, our network is devoted to maximizing the domain invariance ability of learned features based on all the labeled and unlabeled samples. Similar with [18], we maximize the loss L_d of domain label predictor $G_d(\cdot; \theta_d)$. This would make samples from different domains are indistinguishable, i.e., the extracted features are domain invariance. In addition, we minimize the loss L_s of similarity label predictor $G_s(\cdot; \theta_s)$ based on the similar and dissimilar sample pairs, which would further improves the domain invariance and inter-class discriminative ability of learned features. More formally, the optimization of our network is equivalent to solving the following minimization problem

$$\begin{aligned} E(\theta_f, \theta_y, \theta_d, \theta_s) = & \sum_{i,j=1}^{n_s,n_t} L_y(G_y(G_f(x_i^s, x_j^t; \theta_f); \theta_y), (y_i, y_j)) \\ & - \alpha \sum_{i,j=1}^{n_s,n_t} L_d(G_d(G_f(x_i^s, x_j^t; \theta_f); \theta_d), (d_i, d_j)) \\ & + \beta \sum_{i,j=1}^{n_s,n_t} L_s(G_s(G_f(x_i^s, x_j^t; \theta_f); \theta_s), s_{ij}) \\ = & \sum_{i=1}^{n_s+n_t} L_y^i(\theta_f, \theta_y) - \alpha \sum_{i=1}^{n_s+n_t} L_d^i(\theta_f, \theta_d) \\ & + \beta \sum_{i=1}^{n_s \times n_t} L_s^i(\theta_f, \theta_s) \end{aligned} \tag{1}$$

where α and β are tradeoff parameters to avoid overfitting. We transform the problem (1) into a new problem so that it can be solved with the standard SGD method, which will be detailed in the following.

2.2. Model optimization

The problem (1) is solved based on the following stochastic updates method

$$\theta_f \leftarrow \theta_f - \mu(\frac{\partial L_y^i}{\partial \theta_f} - \alpha \frac{\partial L_d^i}{\partial \theta_f} + \beta \frac{\partial L_s^i}{\partial \theta_f}) \tag{2}$$

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial L_y^i}{\partial \theta_y} \tag{3}$$

$$\theta_d \leftarrow \theta_d - \mu(-\alpha) \frac{\partial L_d^i}{\partial \theta_d} \tag{4}$$

$$\theta_s \leftarrow \theta_s - \mu\beta \frac{\partial L_s^i}{\partial \theta_s} \tag{5}$$

where μ represents the learning rate. The update process of equations (2)-(5) is formally like the widely used SGD method except the factor $-\alpha$ in (2) and (4), which makes the problem hard to solve with standard SGD method directly.

In order to update the parameters in equations (2)-(5) with the standard SGD method, we introduce a gradient reversal layer (GRL)[18] between the feature extractor $G_f(\cdot; \theta_f)$ and the domain label predictor $G_d(\cdot; \theta_d)$, as shown in Figure 1. During the forward propagation, the GRL layer is an identity transform. During the back propagation, the GRL layer takes the gradient from the subsequent layer, multiplies it by $-\alpha$ and passes it to the preceding layer. To summarize in mathematical, we formulate the GRL as a "pseudo-function" $R_\alpha(x)$

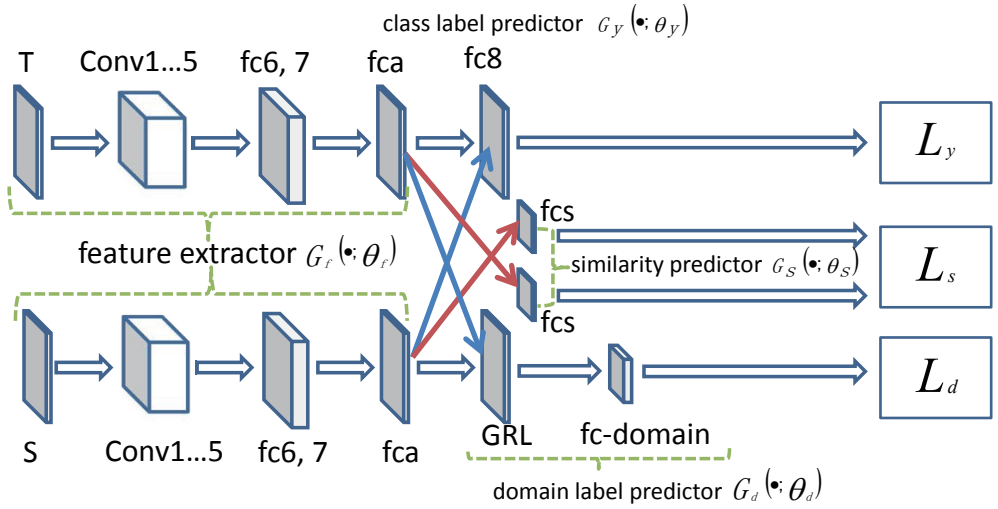

Fig. 1. Framework of proposed method.

defined by the following two equations describing its forward and back propagation.

$$R_\alpha(x) = x \tag{6}$$

$$\frac{dR_\alpha}{dx} = -\alpha \mathbf{I} \tag{7}$$

where $\mathbf{I}$ represents an identity matrix. Based on the pseudo-function $R_\alpha(x)$, the equation (1) can be transformed into the following form.

$$\begin{aligned} E(\theta_f,\theta_y,\theta_d,\theta_s) = & \sum_{i,j=1}^{n_s,n_t} L_y(G_y(G_f(x_i^s,x_j^t;\theta_f);\theta_y),(y_i,y_j)) \\ & + \sum_{i,j=1}^{n_s,n_t} L_d(G_d(R_\alpha(G_f(x_i^s,x_j^t;\theta_f));\theta_d),(d_i,d_j)) \\ & + \beta \sum_{i,j=1}^{n_s,n_t} L_s(G_s(G_f(x_i^s,x_j^t;\theta_f);\theta_s),s_{ij}) \end{aligned} \tag{8}$$

The update process of (2)-(5) can then be implemented as doing standard SGD for (8), and leads to the emergence of features that are domain invariance and inter-class discriminative. After the convergence of model optimization, the class label predictor $y(x) = G_y(G_f(x;\theta_f);\theta_y)$ would be a robust model for cross-domain image recognition.

3. EXPERIMENTS

In this section, we first introduce the exact architecture of our network model, and then evaluate our method on the widely-used Office-Caltech benchmark [5, 6, 7, 10] for cross-domain image recognition. We compare the proposed method with several competitive ones. Experimental results show that our model is effective for cross-domain image recognition, and achieves the state-of-the-art performance.

3.1. Network architecture

The baseline network architecture is the 8-layers AlexNet proposed in [19], as shown in Figure 1. Since there are only a few samples in the widely-used benchmark for cross-domain image recognition, we insert a bottleneck layer fca (128-dimensional, fully connection) between the $fc7$ and $fc8$ layers to avoid overfitting. We adopt $softmaxloss$ and $contrastiveloss$ for the loss function of L_y and L_s, respectively. For the domain label prediction branch, we attach 3 fully connection layers ($1024 \rightarrow 1024 \rightarrow 1$) to the GRL layer, and adopt the $crossentropyloss$ for the loss function of L_d.

Our network is initialized with a model pre-trained on the ImageNet dataset [12]. The learning rate of layers subsequent to fca is 10 times that of the preceding layers. The factor α and β in equation (1) is set to 1 and 10 by cross validation, respectively. During training, images are preprocessed by the mean subtraction. A half of samples in each batch are from source domain, and the rest consists of samples from the target domain.

3.2. Experimental results

The widely-used Office-Caltech dataset for cross-domain image recognition is composed of four domains: Amazon (denoted by **A**), DSLR (denoted by **D**), Webcam (denoted by **W**) and Caltech (denoted by **C**). The first three domains are from the Office dataset [1], and share 31 common object categories, while the Caltech domain is introduced in [20] and

Method	A→C	A→D	A→W	W→A	W→C	W→D
Baseline	80.8	78.4	77.3	70.1	64.8	95.1
SA [21]	79.9	78.5	77.7	71.1	64.1	95.5
GFK [22]	79.1	80.4	78.5	70.0	67.3	96.1
Our method	**85.4**	**91.7**	**92.2**	**89.7**	**81.4**	**97.5**

Table 1. Recognition accuracies on 6 pairs of unsupervised cross-domain image recognition.

there are 10 common classes among the four domains. The number of images per class ranges from 8 to 151.

In order to validate the generalization performance of our method, we first evaluate our method based on the standard unsupervised domain adaptation experimental protocol from [21, 22], i.e., no labeled sample in target domain is available while learning the recognition model. For the proposed method, it degrades to an unsupervised method after removing the similarity prediction branch during training. In this paper we focus on 6 pairs of source (S) and target (T) domains among the Office-Caltech dataset. We denote a cross-domain image recognition problem by the notation S $\rightarrow$ T. The recognition accuracies on the 10 common classes are shown in Table 1.

The results in Table 1 are given based on DCNN features extracted from the **fc7** layer of the AlexNet model [19]. For the baseline method, we use the source data only to train classifiers (linear SVM).The two most popular domain adaptation methods (SA [21] and GFK [22]) show a slight performance increment over the baseline. The proposed method takes full advantage of all the labeled and unlabeled samples in both domains and learns an end-to-end cross-domain image recognition model, which significantly improves the recognition accuracies and performs the best on all pairs.

We further evaluate our method in the semi-supervised way (a few labeled source and target samples are available while learning the classification model). We follow the standard semi-supervised domain adaptation experimental protocol from [1, 8, 13], and use the Office dataset consisting of three domains: **A**, **D** and **W**. There are 31 common object categories among the three domains. We randomly select 20 samples per category from the source domain **A** and 8 samples per category from **D** or **W** as the source domain while 3 labeled target samples per category are randomly selected. We evaluate our method based on 10 random splits across all domain adaptation pairs. The average cross-domain image classification accuracies and the standard deviation are reported in Table 2.

In Table 2, based on the output feature presentation of **fc7** layer in AlexNet, the classification model (linear-SVM) of the baseline method is directly trained with the labeled source and target samples, i.e., without domain adaptation method. The recognition results of the compared deep learning methods are directly quoted from their papers. In contrast to the DLID [8], DeCAF6 [13] and DaNN [9] methods that finish the feature learning and recognition model training separately, the DDC [17] and DAB [18] methods unify the feature learning and recognition model training into a CNN framework, and they perform a much better cross-domain image recognition performance. Compared with the DAB [18] method, our method proposes to exploit the similarity of sample pairs among different domains, and presents a different network framework. Experimental results demonstrate that our method takes full advantage of all samples in the source and target domains to learn domain invariance and inter-class discriminative features, and achieves the state-of-the-art performance on all the three domain pairs.

Method	A→W	D→W	W→D	Average
Baseline	84.2±1.6	95.5±0.8	95.3±1.2	91.7
DLID [8]	51.9	78.2	89.9	73.3
DeCAF6 S+T [13]	80.7±2.3	94.8±1.2	-	-
DaNN [9]	53.6±0.2	71.2±0.0	83.5±0.0	69.4
DDC [17]	84.1±0.6	95.4±0.4	96.3±0.3	91.9
DAB [18]	86.0±0.8	95.3±0.8	96.0±1.0	92.4
Our method	**87.2±1.2**	**97.0±0.4**	**97.9±0.5**	**94.0**

Table 2. Recognition accuracies on 3 pairs of semi-supervised cross-domain object recognition.

4. CONCLUSION

In this paper, we have presented a novel semi-supervised cross-domain image recognition method based on the convolution neural network. The proposed method utilizes the labeled samples to learn features that are inter-class discriminative by minimizing the loss of class label prediction. Meanwhile, the unlabeled samples are exploited to learn features that are domain invariance by maximizing the loss of domain label prediction. In addition, the proposed method constructs a huge number of similar and dissimilar sample pairs to further enhanced the learned features to be domain invariance and inter-class discriminative by minimizing the loss of similarity prediction. Extensive experiments have proved the effectiveness of our method on the cross-domain image recognition. Experimental results show that our method achieves the state-of-the-art performance on the widely used visual domain adaptation benchmark.

Acknowledgment

This work is funded by the CAS "Light of West China" Program, National Natural Science Foundation of China (Grant No. 61602433, Grant No. 61502444 and Grant No. 61472386), and the Chongqing Research Program of Basic Research and Frontier Technology (No. cstc2016jcyjA0011).

5. REFERENCES

[1] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *European Conference on Artificial Intelligence*, 2010, pp. 213–226.

[2] Mahsa Baktashmotlagh, Mehrtash T. Harandi, Brain C. Lovell, and Mathieu Salzmann, "Unsupervised domain adaptation by domain invariant projection," in *International Conference on Computer Vision*, 2013, pp. 769–776.

[3] Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko, "Efficient learning of domain-invariant image representations," in *ICLR*, 2013.

[4] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," in *International Joint Conference on Artificial Intelligence*, 2009, pp. 1187–1192.

[5] Q. Qiu, V. M. Patel, P. Turaga, and R. Chellappa, "Domain adaptive dictionary learning," in *European Conference on Artificial Intelligence*, 2012, pp. 631–645.

[6] Y. Shi and F. Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," in *International Conference on Machine Learning*, 2012, pp. 1079–1086.

[7] Pengcheng Liu, Peipei Yang, Kaiqi Huang, and Tieniu Tan, "Uniform low-rank representation for unsupervised visual domain adaptation," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 216–220.

[8] Sumit Chopra, Suhrid Balakrishnan, and Raghuraman Gopalan, "Dlid: Deep learning for domain adaptation by interpolating between domains," in *ICML Workshop on Representation Learning*, 2013.

[9] Muhammad Ghifary, W. Bastiaan Kleijn, and Mengjie Zhang, "Domain adaptive neural networks for object recognition," in *arXiv:1409.6041v1*, 2014.

[10] Pengcheng Liu, Chong Wang, Peipei Yang, Kaiqi Huang, and Tieniu Tan, "Cross-domain object recognition using object alignment," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015, pp. 66.1–66.12.

[11] Hyeonseob Nam and Bohyung Han, "Learning multi-domain convolutional neural networks for visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4293–4302.

[12] Deng J., Dong W., Socher R., Li L.-J., Li K., and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[13] J Donahue, Y Jia, O Vinyals, J Hoffman, N Zhang, E Tzeng, and T Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International Conference on Machine Learning*, 2014, pp. 647–655.

[14] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1717–1724.

[15] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *arXiv preprint arXiv:1403.6382*, 2014.

[16] Judy Hoffman, Eric Tzeng, Jeff Donahue, Yangqing Jia, Kate Saenko, and TTrevor Darrell, "One-shot adaptation of supervised deep convolutional models," *CoRR*, vol. abs/1312.6204, 2013.

[17] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell, "Deep domain confusion: Maximizing for domain invariance," in *arXiv:1412.3474v1*, 2014.

[18] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015, pp. 1180–1189.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Annual Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[20] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," *Technical Report*, 2007.

[21] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *International Conference on Computer Vision*, 2013, pp. 2960–2967.

[22] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2066–2073.