Conditional Adversarial Transfer for Glaucoma Diagnosis

Jingwen Wang, Yuguang Yan, Yanwu Xu, Wei Zhao, Huaqing Min, Mingkui Tan*, Jiang Liu

***Abstract*— Deep learning has achieved great success in image classification task when given sufficient labeled training images. However, in fundus image based glaucoma diagnosis, we often have very limited training data due to expensive cost in data labeling. Moreover, when facing a new application environment, it is difficult to train a network with limited labeled training images. In this case, some images from some auxiliary domains (*i.e.*, source domain) could be exploited to improve the performance. Unfortunately, direct using the source domain data may not achieve promising performance for the domain of interest (*i.e.*, target domain) due to reasons like distribution discrepancy between two domains. In this paper, focusing on glaucoma diagnosis, we propose a deep adversarial transfer learning method conditioned on label information to match the distributions of source and target domains, so that the labeled source images can be leveraged to improve the classification performance in the target domain. Different from the most existing adversarial transfer learning methods which consider marginal distribution matching only, we seek to match the label conditional distributions by handling images with different labels separately. We conduct experiments on three glaucoma datasets and adopt multiple evaluation metrics to verify the effectiveness of our proposed method.**

I. Introduction

Glaucoma is caused by structural changes in the optic nerve head, and it is often caused by the progressive degeneration of optic nerve fibres [7]. Due to the incurableness of glaucoma, it is extremely important to diagnose glaucoma from digital fundus photographs in time. Recently, deep convolutional networks have shown promising performance in classifying fundus images into glaucoma and normal fundus ones given sufficient labeled training images [3] due to their powerful representative ability.

However, in fundus image based glaucoma diagnosis, due to the expensive cost in data collection and annotation, we usually have only limited labeled training images. As a result, it is difficult to train a network to diagnose glaucoma from fundus images. Although the domain of interest (*i.e.*, target domain) includes only limited images, some labeled images in an auxiliary domain (*i.e.*, source domain) can be used to enhance the classification performance in the target domain. For example, in order to classify the images in ORIGA dataset [16], we can leverage the images in iSee or REFUGE datasets to train a better classifier for ORIGA. Due to the domain discrepancy between different datasets, direct using source images may not achieve promising performance in the target domain. To alleviate this, transfer learning is proposed to match the data distributions of different domains [10]. As a result, the source images can be better leveraged for the classification task in the target domain.

There have been some works investigating deep learning for domain transfer [8], [9]. Among them, deep adversarial learning has attracted much attention for distribution matching [4], [15]. However, most of them only match the marginal distributions of domains. This means that the approaches for distribution matching omit the label information, which is vital for classification. Motivated by this, we propose a deep adversarial transfer learning method called **C**onditional **A**dversarial **T**ransfer (CAT) to reduce the label conditional distributions for glaucoma diagnosis.

Specifically, we introduce two domain discriminators for positive and negative samples, respectively. The domain discriminators are trained to distinguish between source and target images, and encoders are trained to make source and target images indistinguishable by domain discriminators. By this adversarial learning approach, the label conditional distributions of two domains are matched, and label information extracted from the source images can be transferred into the target domain. To demonstrate the effectiveness of our proposed method, we adopt three glaucoma datasets and multiple evaluation metrics in our experiments.

A. Related Studies

In the last decade, deep learning has been widely applied in medical image analysis especially for fundus image processing [1], [7]. M-net applies a multi-scale U-net for fundus segmentation [3]. In [13], Sun *et al.* use an object detection network for optic disc segmentation. In [14], R-CNN is used to localize optic disc and cup for glaucoma diagnosis.

Transfer learning has been widely studied in the communities of machine learning, data mining, medical imaging, *etc* [10], [12], [11]. Adversarial learning has also attracted much attention since the proposal of generative adversarial networks [5]. ADDA [15] performs adversarial learning between a domain discriminator and two encoders to match the source and target distributions. A gradient reversal layer (GRL) is proposed in [4] to reverse the gradient from a domain discriminator for distribution matching. These methods aim to match the marginal distributions without considering the label information, which is vital for classification. Compared with them, our proposed method exploit label information to perform label conditional distribution matching between domains, thus can achieve better performance in glaucoma diagnosis.

J. Wang, Y. Yan, H. Min, M. Tan are with South China University of Technology, China. Y. Xu is with Baidu, inc., China. J. Liu is with Southern University of Science and Technology, Chinese Academy of Sciences, China. W. Zhao is with CVTE Research, China. This work was done when J. Wang and Y. Yan were interns at Medical Image and Signal Processing Group, CVTE Research. M. Tan is the corresponding author (mingkuitan@scut.edu.cn).

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

II. Methodology

A. Problem Statement

We now present the notations used in this paper. Let $\mathbf{S} = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ be the set of source images and their corresponding labels, where $y_i^s \in \{0, 1\}$, and n_s is the number of source images. The target images and the corresponding labels are denoted as $\mathbf{T} = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t}$, where $y_i^t \in \{0, 1\}$, and n_t is the number of target images. $\mathbf{S}$ and $\mathbf{T}$ are drawn from two different distributions, *e.g.*, two different glaucoma datasets collected from two practical applications. Among the target images, let $\mathbf{L} = \{\mathbf{x}_i^l, y_i^l\}_{i=1}^{n_l}$ be the set of labeled images and their labels, and $\mathbf{U} = \{\mathbf{x}_i^u\}_{i=1}^{n_u}$ be the set of unlabeled images for testing, where n_l and n_u are the numbers of images in $\mathbf{L}$ and $\mathbf{U}$. In this paper, we leverage both source and target labeled images (*i.e.*, $\mathbf{S}$, $\mathbf{L}$) to learn a classifier for predicting target test images (*i.e.*, $\mathbf{U}$).

B. Overview

Fig. 1 presents an overview of the proposed method CAT. We use a source encoder E_s and a target encoder E_t to extract high-level representations for source and target images, respectively. A classifier C is trained on images $\mathbf{S}$ and $\mathbf{L}$ to generate the probability $\Pr(y = 1|\mathbf{x})$ given an image $\mathbf{x}$ from the source or target domain. Domain discriminators D_+ and D_- are trained to generate the probability that the input comes from the source domain, where D_+ is trained for positive images, and D_- is trained on negative images. By performing adversarial learning between the encoders and domain discriminators, the distributions of source and target representations can be matched. As a result, the classifier can be used to classify target images in the testing set. In practice, we share the parameters of two encoders E_s and E_t, thus a better encoder is expected because of more training images.

C. Conditional Adversarial Transfer

To obtain an effective classifier, we leverage labeled images from both source and target domains to train the encoders and classifier by minimizing the following loss function:

$$\begin{aligned}\mathcal{L}_{bce}(E_s, E_t, C) = \\ -\sum_{i=1}^{n_s} y_i^s \log C(E_s(\mathbf{x}_i^s)) + (1 - y_i^s) \log\big(1 - C(E_s(\mathbf{x}_i^s))\big) \\ -\sum_{i=1}^{n_l} y_i^l \log C(E_t(\mathbf{x}_i^l)) + (1 - y_i^l) \log\big(1 - C(E_t(\mathbf{x}_i^l))\big), \quad (1)\end{aligned}$$

where the classifier C consists of a fully-connected layer for classification, and encoders E_s and E_t are convolutional networks.

Due to the domain discrepancy, the direct use of source and target images for training the model cannot achieve promising performance on target test images. To alleviate this issue, one of the common methods is to introduce a domain discriminator, and then perform adversarial learning between the discriminator and encoders. Specifically, let D be a domain discriminator network trained by the binary cross-entropy loss to calculate the probability that the input comes from the source domain, the goal of E_s and E_t is to mislead D to maximize its loss function. In this way, the domain discriminator finally cannot distinguish the source and target embeddings, so that the marginal distribution of source and target domain can be aligned, and the classifier C is trained on a unified data distribution. Adversarial learning between (E_s, E_t) and D can be formulated as the following adversarial loss function:

$$\begin{aligned}\mathcal{L}_{adv}(E_s, E_t, D) = \\ -\sum_{i=1}^{n_s} \log D(E_s(\mathbf{x}_i^s)) - \sum_{i=1}^{n_l} \log\big(1 - D(E_t(\mathbf{x}_i^l))\big). \quad (2)\end{aligned}$$

However, this approach has a severe drawback that it omits label conditional information involved in images, which is important to train a classifier for glaucoma diagnosis. Motivated by this, we propose to match the label conditional distribution of the source and target domains by handling positive and negative data separately. To this end, we employ two domain discriminators, D_+ and D_-, to respectively calculate the probability that the input comes from the source domain, and perform adversarial learning between the encoders and the discriminators to match the distributions of positive and negative images. Let $\mathbf{x}_i^{s+}$ and $\mathbf{x}_i^{l+}$ be positive images, n_{s+} and n_{l+} be the numbers of positive source images and positive target images, respectively. The adversarial objective related to D_+ is given by

$$\begin{aligned}\mathcal{L}_{adv}^{+}(E_s, E_t, D_+) = \\ -\sum_{i=1}^{n_{s+}} \log D_+(E_s(\mathbf{x}_i^{s+})) - \sum_{i=1}^{n_{l+}} \log\big(1 - D_+(E_t(\mathbf{x}_i^{l+}))\big). \quad (3)\end{aligned}$$

Similarly, define $\mathbf{x}_i^{s-}$ and $\mathbf{x}_i^{l-}$ as negative images, n_{s-} and n_{l-} as the numbers of negative source images and negative target images, respectively. The adversarial objective related to D_- is given by

$$\begin{aligned}\mathcal{L}_{adv}^{-}(E_s, E_t, D_-) = \\ -\sum_{i=1}^{n_{s-}} \log D_-(E_s(\mathbf{x}_i^{s-})) - \sum_{i=1}^{n_{l-}} \log\big(1 - D_-(E_t(\mathbf{x}_i^{l-}))\big). \quad (4)\end{aligned}$$

Overall, the optimization problems *w.r.t.* C, D_+ and D_- are given by

$$\min_C \ \mathcal{L}_{bce}(E_s, E_t, C), \quad (5)$$

$$\min_{D_+} \ \mathcal{L}_{adv}^{+}(E_s, E_t, D_+), \quad (6)$$

$$\min_{D_-} \ \mathcal{L}_{adv}^{-}(E_s, E_t, D_-), \quad (7)$$

respectively. The encoders E_s and E_t receive losses from the classifier C and the domain discriminators (D_+, D_-), and are optimized by minimizing optimization problem:

$$\begin{aligned}\min_{E_s, E_t} \mathcal{L}_{bce}(E_s, E_t, C) \\ - \mathcal{L}_{adv}^{+}(E_s, E_t, D_+) - \mathcal{L}_{adv}^{-}(E_s, E_t, D_-). \quad (8)\end{aligned}$$

III. Experiments

In order to evaluate the performance of the proposed method, we conduct experiments on different fundus image datasets for glaucoma diagnosis, and adopt multiple metrics for evaluation and several state-of-the-art transfer learning methods for comparison.

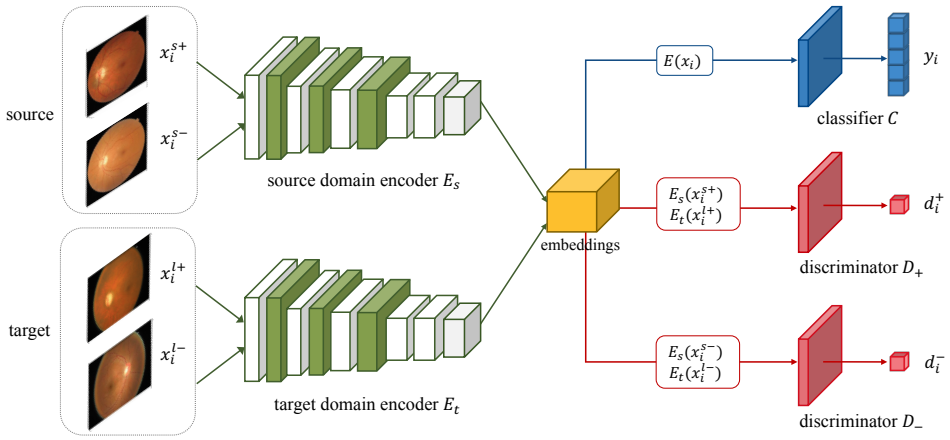

Fig. 1. An illustration of our proposed method. The source (*resp.* target) domain encoder E_s (*resp.* E_t) extracts high-level embeddings from source (*resp.* target) images, and the classifier C aims to classify source and target images. The domain discriminator D_+ (*resp.* D_-) is trained to distinguish between positive (*resp.* negative) source and positive (*resp.* negative) target images, and two encoders E_s and E_t are trained to hamper the classification performance of the domain discriminators D_+ and D_-. As a result, the embeddings of source and target images are indistinguishable by the domain discriminators, so that the label conditional distributions of source and target domains are matched.

A. Datasets

- **REFUGE**[1] contains 800 fundus images in total, comprising of 80 with glaucoma and 720 normal ones.
- **iSee** contains fundus images with four diseases, including AMD, DR, glaucoma and myopia. We randomly pick up 800 images with glaucoma and 800 normal ones from the dataset as a domain.
- **ORIGA** [16] contains 650 fundus images with 168 glaucomatous eyes and 482 normal eyes. Following the setting in [2], the dataset is divided into 325 labeled images for training and 325 unlabeled images for testing.

These three datasets are collected by different devices from different application environments. Therefore, they are taken as three different domains. In our experiments, we take ORIGA as the target domain, and the other two datasets as the source domain, respectively. As a result, we perform two transfer learning tasks, *i.e.*, REFUGE $\rightarrow$ ORIGA and iSee $\rightarrow$ ORIGA.

B. Compared Methods

To demonstrate the effectiveness of our proposed method, we compare against several state-of-the-art transfer learning methods: ADDA [15], GRL [4], DAN [8], and JAN [9].

- ADDA is a generalized framework for domain adaptation. It combines discriminative modeling and GAN loss for asymmetric adaptation.
- GRL addresses domain adaptation by making the source and target samples indistinguishable for a domain discriminator by adversarial training.
- DAN learns transferable features by embedding deep features of multiple task-specific layers to reproducing kernel Hilbert spaces(RKHSs) and matching different distributions optimally using multi-kernel MMD.
- JAN extends DAN by matching the joint distributions of the deep activations in multiple domain-specific layers using Joint MMD.

For the above unsupervised domain adaptation methods, we firstly training the networks in an unsupervised domain adaptation paradigm, and then fine-tune the classifier to take better advantage of labeled target images for training. The fine-tuned versions of the above methods are denoted as ADDA-FT, GRL-FT, DAN-FT and JAN-FT.

In addition, we further conduct the following baseline methods for completed comparison:

- The source-only model (SO) only uses the source images for training without considering the domain discrepancy and knowledge transfer.
- The target-only model (TO) only uses labeled target data for training. This method is a classical supervised learning method without leveraging source data for learning.
- The source+target model (S+T) directly trains the network on all the source and target training data without considering the domain discrepancy.

C. Implementation details

In our experiments, all the methods are implemented on PyTorch platform[2]. We use ResNet-50 [6] pretrained on ImageNet dataset to initialize the encoders, and use a fully-connected layer as the classifier and the domain discriminators. The network is trained on source and target training images for 200 epochs using a batch size of 16. We apply SGD to train the network with the learning rate being 0.0001.

D. Evaluation Metrics

In glaucoma diagnosis, datasets are usually highly imbalanced, which means that the number of normal images

[1]https://refuge.grand-challenge.org

[2]https://pytorch.org

TABLE I

RESULTS ON REFUGE → ORIGA TASK.

Method	recall	F1	G-mean
SO	0.5895	0.5308	0.6601
TO	0.5604	0.5730	0.6873
S+T	0.7419	0.6699	0.7734
ADDA [15]	0.6842	0.5909	0.7106
GRL [4]	0.6421	0.5701	0.6930
DAN [8]	0.7684	0.6728	0.7776
JAN [9]	0.7474	0.6762	0.7774
ADDA+FT	0.7340	0.6188	0.7366
GRL+FT	0.7263	0.6301	0.7434
DAN+FT	0.7789	0.6916	0.7915
JAN+FT	0.7895	0.6757	0.7817
CAT (ours)	**0.8191**	**0.6968**	**0.8007**

TABLE II

RESULTS ON ISEE → ORIGA TASK.

Method	recall	F1	G-mean
SO	0.7579	0.6154	0.7329
TO	0.5604	0.5730	0.6873
S+T	0.8085	0.6941	0.7967
ADDA [15]	0.8000	0.6360	0.7500
GRL [4]	0.8421	0.6426	0.7558
DAN [8]	0.7684	0.6854	0.7862
JAN [9]	0.7684	0.6887	0.7883
ADDA+FT	0.7979	0.6466	0.7611
GRL+FT	0.8000	0.6255	0.7413
DAN+FT	0.8000	0.7037	0.8022
JAN+FT	0.7789	0.7048	0.8001
CAT (ours)	**0.8737**	**0.7124**	**0.8153**

is much larger than that of the images with glaucoma. Therefore, accuracy is insufficient to reflect the performance of the methods. Here we adopt multiple metrics in our experiments for evaluation. Let TP, TN, FP and FN denote the number of true positive, true negative, false positive and false negative, respectively. The metrics we will use are defined as: recall $= \frac{TP}{TP+FN}$, F1 $= \frac{2 \cdot TP}{2 \cdot TP+FN+FP}$, G-mean $= \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{TN+FP}}$.

E. Results and Discussions

TABLES I and II report the results on REFUGE → ORIGA and iSee → ORIGA, respectively. The best results are highlighted with boldface case, and the second best results are underlined. We draw several interesting observations as follows:

- Our proposed method CAT achieves the best performance in terms of the three evaluation metrics, which demonstrates the effectiveness of label conditional distribution matching in deep transfer learning.
- S+T outperforms both SO and TO, which demonstrates that although the domain discrepancy affects the learning performance, more training images are still beneficial for training an effective classifier.
- For ADDA, DAN and JAN, the fine-tuned versions usually outperform the ones without fine-tuning, which verifies the effects of labeled target images.

IV. CONCLUSION

In this paper, we study transfer learning for glaucoma diagnosis based on deep adversarial learning. We propose to match the label conditional distributions of the source and target domains. To this end, we employ two domain discriminators to handle images with positive and negative labels, respectively. The experimental results on three glaucoma datasets demonstrate the effectiveness of our proposed method in terms of multiple metrics.

V. ACKONWLEDGMENT

This work was partially supported by National Natural Science Foundation of China (NSFC) 61602185, Program for Guangdong Introducing Innovative and Enterpreneurial Teams 2017ZT07X183, Guangdong Provincial Scientific and Technological Funds under Grants (2018B010107001, 2017B090910005), and Guangzhou Shiyuan Electronics Co., Ltd. We also thank EyeSee Medical Science & Technology Chengdu Co., Ltd., iMed Team, and Singapore Eye Research Institute for providing research data.

REFERENCES

[1] Abràmoff, M.D., Garvin, M.K., Sonka, M.: Retinal imaging and image analysis. IEEE Reviews in Biomedical Engineering **3**, 169–208 (2010)
[2] Cheng, J., Zhang, Z., Tao, D., Wong, D.W.K., Liu, J., Baskaran, M., Aung, T., Wong, T.Y.: Similarity regularized sparse group lasso for cup to disc ratio computation. Biomedical Optics Express **8**(8), 3763–3777 (2017)
[3] Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. TMI (2018)
[4] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. pp. 1180–1189 (2015)
[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
[7] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
[8] Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: ICML. pp. 97–105 (2015)
[9] Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: ICML. pp. 2208–2217 (2017)
[10] Pan, S.J., Yang, Q.: A survey on transfer learning. TKDE **22**(10), 1345–1359 (2010)
[11] Patel, V.M., Gopalan, R., Li, R., Chellappa, R.: Visual domain adaptation: A survey of recent advances. IEEE Signal Processing Magazine **32**(3), 53–69 (2015)
[12] Shao, L., Zhu, F., Li, X.: Transfer learning for visual categorization: A survey. TNNLS **26**(5), 1019–1034 (2015)
[13] Sun, X., Xu, Y., Tan, M., Fu, H., Zhao, W., You, T., Liu, J.: Localizing optic disc and cup for glaucoma screening via deep object detection networks. In: OMIA. pp. 236–244 (2018)
[14] Sun, X., Xu, Y., Zhao, W., You, T., Liu, J.: Optic disc segmentation from retinal fundus images via deep object detection networks. In: EMBC. pp. 5954–5957 (2018)
[15] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR. p. 4 (2017)
[16] Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y.: Origa-light: An online retinal fundus image database for glaucoma analysis and research. In: EMBC. pp. 3065–3068 (2010)