FEW-SHOT TRANSFER LEARNING FOR HEREDITARY RETINAL DISEASES RECOGNITION

Siwei Mai ⋆, *Qian Li* †, *Qi Zhao* †, *Mingchen Gao* ⋆

⋆ Department of Computer Science and Engineering, University at Buffalo
† Department of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital
Capital Medical University, Beijing Key Laboratory of Ophthalmology and Visual Sciences

ABSTRACT

This project aims to recognize a group of rare retinal diseases, the hereditary macular dystrophies, based on Optical Coherence Tomography (OCT) images, whose primary manifestation is the interruption, disruption, and loss of the layers of the retina. The challenge of using machine learning models to recognize those diseases arises from the limited number of collected images due to their rareness. To handle the problems caused by the lack of labeled data. We formulate this as a Student-Teacher (S-T) learning task with a discriminative feature space and knowledge distillation (KD). OCT images have large variations due to different types of macular structural changes, capturing devices, and angles. To alleviate such issues, a pipeline of preprocessing is first utilized for image alignment. Tissue images at different angles can be roughly calibrated to a horizontal state for better feature representation. Extensive experiments on our dataset demonstrate the effectiveness of the proposed approach.

Index Terms— Hereditary Retinal Diseases Recognition, Student-Teacher Learning, Knowledge Distillation, Transfer Learning

1. INTRODUCTION

Visual impairment and blindness caused by inherited retinal diseases (IRDs) are increasing due to the global prolonged life expectancy. There was no treatment for IRDs until recently, a number of therapeutic approaches such as gene replacement and induced pluripotent stem cells transplantation has been proposed, developed, and shown promising potential in some of the ongoing therapeutic clinical trials. Spectral-domain Optical coherence tomography (SD-OCT) has been playing a crucial role in the evaluation of the retina of IRDs in diagnosis, progression surveillance as well as strategy exploration and response assessment of the treatment. However, the recognition, interpretation, and comparison of the minimal changes on OCT as shown in IRDs sometimes could be difficult and time-consuming for retinal physicians. Recently automated image analysis has been successfully applied in the detection of changes on fundus and OCT images of multiple

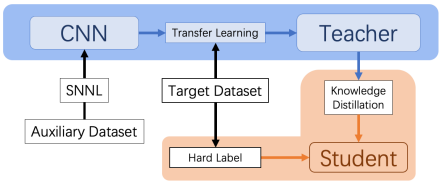


Fig. 1. The overview of the proposed method

retinal diseases such as diabetic retinopathy and age-related macular degeneration, which are of higher prevalence among the population to enable acquisition of large volumes of training data for traditional machine learning approaches including deep learning. On the contrary, for rare diseases like IRDs, acquiring a large volume of high-quality data representative of the patient cohorts is challenging. These datasets also require accompanying annotations generated by experts which are time-consuming to produce. This hinders the applying state-of-the-art image classification, which usually requires a relatively large number of images with annotations for training. The goal of this project is to design a computer-aided diagnosis algorithm when only a very limited number of rare disease samples can be collected.

The methods of diagnosing ocular diseases like age-related macular degeneration (AMD), diabetic macular edema (DME), etc., through the spectral domain OCT images can be roughly categorized into the traditional machine learning methods and deep learning-based methods. There are lots of works on OCT image analysis based on the traditional machine learning methods like Principal Components Analysis (PCA) Support Vector Machine (SVM) , or Random Forest, segmenting each layer of the OCT images [15] or learns global representation directly [16, 19].

Lots of researches also focus on the deep-learning-based methods. Some of them are dedicated to improving existing mature and pre-trained frameworks such as Inception-v3 [6, 20], VGG16 [13, 20], PCANet [2], GoogLeNet [7, 8], ResNet [7, 11], DenseNet [7] to classify OCT images. Others unify multiple networks together to make classification more robust for diagnosing, for example four-parallel-ResNet system [11] and four-network with two-stage system [12]. Com-

pared with the traditional feature engineering, deep learning has the advantage of learning the hierarchical features.

The goal of this project is to classify a group of macular-involved IRDs. We have collected 1128 diseased OCT images from 60 patients. We propose a Student-Teacher Learning framework to overcome the limited training sample problems. In the teacher part, the teacher model is trained by the Soft Nearest Neighbor Loss (SNNL) [3] and Transfer Learning methods using the large-scale labeled auxiliary OCT dataset [9] which contains 3 common retinal degenerative diseases with 84484 images. While for the student model, a smaller size model is used to accommodate the limited sample size. The student model can learn from both the teacher and hard label information by Knowledge Distillation [5]. Demonstrated by the experiments, even under the circumstance of limited training samples, the student model can catch a better performance than the teacher model and the vanilla classification model with the same architecture.

2. METHODS

The proposed pipeline consists of three parts: data preprocessing, training teacher model, training student model. The first step, data preprocessing, mainly includes image angle adjustments and vertical pixel columns movements. The teacher model is trained to adapt to the auxiliary OCT dataset first with a 4-category classification designed to learn the nuances from each disease by SNNL in the fixed-dimensional feature space, then transfer to the target OCT dataset with 5-categories. The student model learning from the information from the soft label with the teacher and the hard label from real labeled data by Knowledge Distillation.

2.1. Image Alignment

As shown in Fig. 3, the original OCT images are usually captured from diverse angles. We adjust all images to the horizontal position without destroying the original pathological information following a adapted OCT image preprocessing strategy from [16]. The main idea of process is to generate a mask to attain retina layered structure.

The process begins with noise reduction as shown in Fig.2(b) to reduce the irregularly distributed noise. We use Block-matching and 3D filtering(BM3D)[1], the state-of-the-art noise-reducing algorithm. The noise reduction preprocessing helps subsequent steps better capture retina structure. The Otsu algorithm in Fig.2(c) allocates the location and morphology of the black background. The median filter in Fig.2(d) further reduces the noise area within the tissue. The morphological operations opening and closing in Fig.2(e) are used to clean noises inside and outside the tissue area so that the binary mask of the target tissue is complete and clear. After the contours are obtained, we use a polynomial curve fitting to represent the degree of curvature of the tissue area.

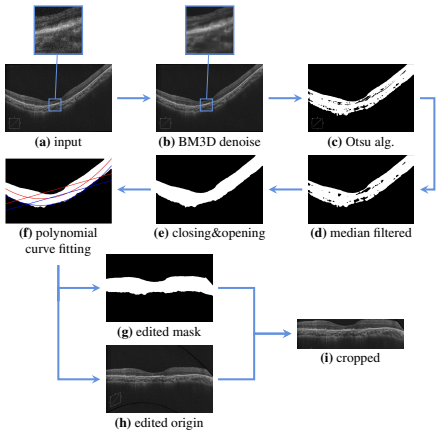


Fig. 2. Pipeline for the image alignment

At last, adjusting and cropping both the mask and the original image as shown in Fig.2(i).

2.2. Feature Space Learning

Soft Nearest Neighbor Loss (SNNL) [3] is applied for the teacher model training in the feature space before the classifier. Teacher model is the backbone structure for absorbing and learning the information from an auxiliary dataset, more specifically, the textures, patterns, and pixel distributions in the end-level convolutional layers. There are bound to be objective differences between the target and auxiliary datasets, and the great separation between the categories in the feature space will facilitate the subsequent transfer learning [3, 18] with the target dataset.

$$\mathcal{L} = \sum_{j} y_j \log f^k(x_j) + \alpha \cdot \sum_{i \in k-1} l'_{sn}(f^i(x), y) \quad (1)$$

$$l'_{sn} = \arg\min_{T \in \mathbb{R}} -\frac{1}{b} \sum_{i \in 1 \ldots b} \log \left(\frac{\sum\limits_{\substack{j \in 1 \ldots b \\ j \neq i \\ y_i = y_j}} e^{-\frac{\|x_i - x_j\|^2}{T}}}{\sum\limits_{\substack{k \in 1 \ldots b \\ k \neq i}} e^{-\frac{\|x_i - x_k\|^2}{T}}} \right) \quad (2)$$

Eqn. 1 shows the total loss function, which consists of the cross-entropy loss on logits and the soft nearest neighbor loss for the representation learning controlled by the hyper-parameter α. In Eqn. 2, b is the batch size, T is the temperature. When the temperature is large, the distances between widely separated points can influence the soft nearest neighbor loss more. Moreover, the numerator of the $\log$ function implies the distance between the target, and similar samples in each category, while the denominator is the distances between the target and other samples in the batch. Usually, the use of cosine distances in training results in a smoother training process.

2.3. Knowledge Distillation & Student-Teacher Learning

In order to overcome the obstacles caused by the lack of training data, we use the combination of Knowledge Distillation and Student-Teacher Learning for knowledge transfer [5, 21].

$$\begin{aligned}\mathcal{L}(x;W) =& \alpha \cdot H(y, \sigma(z_s; T=1)) \\ &+ \beta \cdot H(\sigma(z_t; T=\tau), \sigma(z_s; T=\tau))\end{aligned} \tag{3}$$

By the S-T architecture, the smaller size student model with the blank background is able to accept the knowledge from the fine-tuned teacher as well as information from labels. In Eqn. 3, α and β control the balance of information coming from the two sources, which generally add up to 1. τ denotes the temperature of adapted softmax function, each probability p_i of class i in the batch is calculated from the logits z as:

$$p_i = \frac{\exp(\frac{z_i}{T})}{\sum_j \exp(\frac{z_j}{T})} \tag{4}$$

As the T in the Eqn. 4 increases, the probability distribution of the output becomes "softer", which means the differences among the probability of each class decreased and more information will provide.

3. EXPERIMENTS

We conduct experiments to classify five categories of IRDs with the help of a much larger auxiliary dataset.

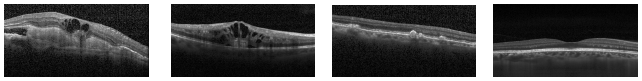
3.1. Datasets

3.1.1. Auxiliary Dataset

We use a publicly available dataset of OCT images as shown in Fig. 3 [9] for the teacher model training. The data set contains four categories including NORMAL, CNV, DME and DRUSEN. They have a total of 109,309 samples, of which 1,000 are used for testing and the rest are used for training. There are two kinds of sizes in those images, $1536 \times 496 \times 1$ and $1024 \times 496 \times 1$. For the experiments, they are all preprocessed and resized to $224 \times 224 \times 1$.

3.1.2. Target Dataset and Data Acquisition

In the target dataset as shown in Fig. 4, we have 1128 samples from 60 patients's 94 eyes, of which 236 are central IRDs, 204 are excluded central IRDs, 209 are extensive IRDs, 185 are normal and 294 are control samples (CNV, DME, MH, ERM...). The size of the images is $1180 \times 786 \times 1$. The data were collected from Beijing Tongren Eye Center with a clinical diagnosis of IRDs involving the macular area were included in the current study. SD-OCT data were acquired using a Cirrus HD-OCT 5000 system (Carl Zeiss Meditec Inc., Dublin, CA, USA). Extracted macular OCTs containing at least one OCT scan providing a cross section of the fovea were included in this study. The B scan OCT images with evidence of retinal disease as determined by two retinal specialists were defined as controls. For the experiments, they are all preprocessed and resized to $224 \times 224 \times 1$.



(a) CNV (b) DME (c) DRUSEN (d) Normal

Fig. 3. Four types of samples in the auxiliary dataset. (a) choroidal neovascularization (CNV) with neovascular membrane (unusually cavernous part) and associated subretinal fluid. (b) Diabetic macular edema (DME) with retinal-thickening-associated intraretinal fluid (abnormal separation of tissues and fluid content in the layers). (c) Multiple drusen (white irregular bulge) present in early AMD. (d) Normal retina with preserved foveal contour and absence of any retinal fluid/edema.

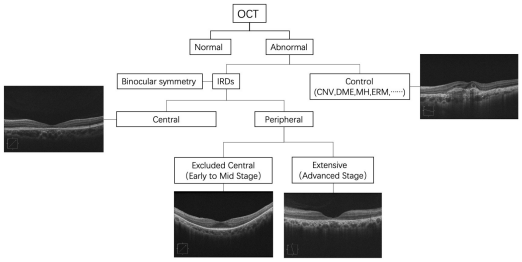


Fig. 4. Relationship of categories in the target dataset

3.2. Baseline

We conducted baseline experiments on the model's adaptability to the data and vanilla classification capabilities.

3.2.1. Data Applicability

We conducted training without pre-training on five existing popular frames with preprocessed images. The results are shown in Table 1. ResNet-18 shows the best accuracy from for the baseline traing, and is selected as the backbone in the following experiments.

Table 1. Comparison of vanilla classification models

Methods	Top 1 Accuracy (%)
AlexNet [10]	53.29
VGG11 [14]	58.08
Inception [17]	67.66
ResNet-18 [4]	**70.06**

Table 2. Comparison of Fine-Tuning strategies

Methods	Top 1 Accuracy (%)
Features Extraction	55.69
All Parameters	59.28
High-level	**62.28**

Table 3. Test Accuracy of Different Methods

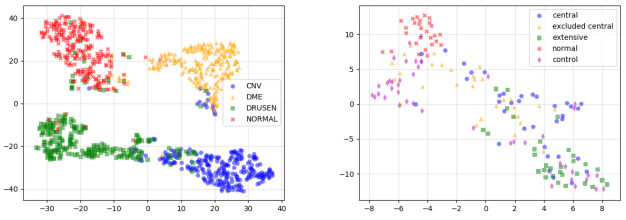
Methods	Accuracy (%)
Teacher's Fine-Tuning	59.68±2.59
Vanilla Classification	69.65±0.40
S-T Learning	**74.45±1.59**

3.2.2. Fine-Tuning

After training the Teacher model with the auxiliary dataset by SNNL loss, the ResNet-50 model get a **90.10%** test accuracy. A relative larger network architecture is selected here to handle the large auxiliary dataset as the teacher model. Based on pre-trained model parameters, we performed three different forms of fine-tuning to the target dataset. In the Features Extraction way, we freeze the parameters before the last fully-connected layer (classifier). In the High-level way, we freeze the parameters before the 5th group of convolution layers (the 129_{th} layer in the ResNet-50), left the last group of convolution to learn the high-level features from the new target dataset. In the way of All Parameters, the model can adjust all the parameters included in ResNet-50. From the data in Table 2, we pick the parameters from High-level way to play the teacher role in the S-T architecture.

3.3. Feature Space Representation

The SNNL loss function enables the model to get a better projection of the input image during training in the designed feature space, which means that inter-class samples can be clustered while intra-class samples can be separated by the distance function. From Fig. 5, we can see that when the Teacher model (ResNet-50) is trained by the auxiliary dataset.



(a) Auxiliary Dataset (b) Target Dataset

Fig. 5. 128-Dimensional Feature Space Representation (processed by T-SNE)

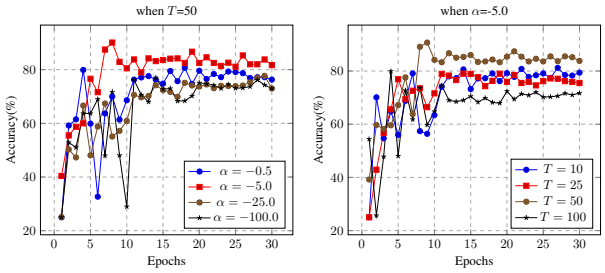


Fig. 6. Teacher model (ResNet-50) Test Experiments

It has the ability to projection the test samples from the auxiliary dataset. Meanwhile, to the target dataset, the Teacher also can cluster the normal class and control class which becomes the control class in the target dataset before fine-tuning and transfer learning.

3.4. Student-Teacher Learning

3.4.1. Teacher Model

We choose the ResNet-50 as Teacher Model to handle the auxiliary dataset. The performance is mainly controlled by the hyper-parameter α and Temperature T in Eqn. 1 and Eqn. 2. In our experiment, we fixed the dimension of Feature Space as 128. We trained them for 30 epochs with an auxiliary dataset before transfer learning and fine-tuning. We decreased the learning rate at epoch 10 and 25 with factor 0.1. There are two sets of control trials in Fig. 6. We respectively fixed the hyper-parameters α and T. Firstly, We got the best performance at α=-5.0 when T=50. Then, we fixed the α=-5.0 and got the best performance at T=50. So, the optimal hyper-parameters are α=-5.0 and T=50.

3.4.2. Student Model

After accomplished the transfer learning and fine-tuning of Teacher model, We use the ResNet-18 as the Student Model to adapt the smaller size of target data. The ResNet-18 is totally untrained by any data before the S-T learning.

From Table 3, we can see that the student model in our trained S-T architecture gets better results than the teacher and its original classification model. This is because it incorporates knowledge from the teacher's pre-training and information from the hard-label classification.

4. CONCLUSION

In this study, we demonstrate a Student-Teacher Learning based classification model on a small dataset to distinguish several retinal diseases. This framework learns the knowledge from both ground truth labels and pretrained Teacher model to make it possible to handle limited data. Data preprocessing also plays a critical role that cannot be ignored before training.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Beijing Tongren Eye Center. (No.TRECKY2017-10, Mar.3,2017)

6. ACKNOWLEDGMENTS

Author MG is partially funded by NSF-IIS-1910492.

7. REFERENCES

[1] K. Dabov, Foi, et al. Image denoising by sparse 3D transform-domain collaborative ltering. 16(8):16.

[2] L. Fang and C. Wang. Automatic classification of retinal three-dimensional optical coherence tomography images using principal component analysis network with composite kernels. *Journal of Biomedical Optics*, 22(11):1, Nov. 2017.

[3] N. Frosst, N. Papernot, et al. Analyzing and Improving Representations with the Soft Nearest Neighbor Loss. *arXiv:1902.01889 [cs, stat]*, Feb. 2019.

[4] K. He, Zhang, et al. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, Dec. 2015.

[5] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531 [cs, stat]*, Mar. 2015.

[6] Q. Ji, W. He, et al. Efficient Deep Learning-Based Automated Pathology Identification in Retinal Optical Coherence Tomography Images. *Algorithms*, 11(6):88, June 2018.

[7] Q. Ji, J. Huang, et al. Optimized Deep Convolutional Neural Networks for Identification of Macular Diseases from Optical Coherence Tomography Images. *Algorithms*, 12(3):51, Feb. 2019.

[8] S. P. K. Karri, D. Chakraborty, et al. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. *Biomedical Optics Express*, 8(2):579, Feb. 2017.

[9] D. Kermany. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification, Jan. 2018.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.

[11] W. Lu, Y. Tong, et al. Deep Learning-Based Automated Classification of Multi-Categorical Abnormalities From Optical Coherence Tomography Images. *Translational Vision Science & Technology*, 7(6):41, Dec. 2018.

[12] N. Motozawa, G. An, et al. Optical Coherence Tomography-Based Deep-Learning Models for Classifying Normal and Age-Related Macular Degeneration and Exudative and Non-Exudative Age-Related Macular Degeneration Changes. *Ophthalmology and Therapy*, 8(4):527–539, Dec. 2019.

[13] F. Y. Shih and H. Patel. Deep Learning Classification on Optical Coherence Tomography Retina Images. *International Journal of Pattern Recognition and Artificial Intelligence*, page 2052002, Sept. 2019.

[14] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, Apr. 2015.

[15] J. Sugmk, S. Kiattisin, et al. Automated classification between age-related macular degeneration and Diabetic macular edema in OCT image using image segmentation. In *The 7th 2014 Biomedical Engineering International Conference*, pages 1–4, Fukuoka, Japan, Nov. 2014. IEEE.

[16] Y. Sun, S. Li, and Z. Sun. Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning. *Journal of Biomedical Optics*, 22(1):016012, Jan. 2017.

[17] C. Szegedy, V. Vanhoucke, et al. Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567 [cs]*, Dec. 2015.

[18] Y. Tian, D. Krishnan, et al. Contrastive Representation Distillation. *arXiv:1910.10699 [cs, stat]*, Jan. 2020.

[19] F. G. Venhuizen, B. van Ginneken, et al. Automated age-related macular degeneration classification in OCT using unsupervised feature learning. In L. M. Hadjiiski and G. D. Tourassi, editors, *SPIE Medical Imaging*, page 94141I, Orlando, Florida, United States, Mar. 2015.

[20] J. Wang, G. Deng, et al. Deep learning for quality assessment of retinal OCT images. *Biomedical Optics Express*, 10(12):6057, Dec. 2019.

[21] L. Wang and K.-J. Yoon. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *arXiv:2004.05937 [cs]*, May 2020.