Diagnosis of retinal diseases using the vision transformer model based on optical coherence tomography images

Zenan Zhou, Chen Niu, Huanhuan Yu, Jiaqing Zhao, Yuchen Wang, Cuixia Dai*
Shanghai Institute of Technology, Shanghai, China.
*Corresponding author: sdadai7412@163.com.

ABSTRACT

Deep learning is rapidly becoming the state of the art, leading to enhanced performance in various medical diagnosis. In recent years, convolutional neural network (CNN) has been used to diagnose retinal disease and has proven its superiority in detection and classification tasks. Vision transformer is a new image classification model that has been proposed in 2020. It does not rely on any CNN and completely performs based on the transformer structure which has a different feature extraction method from CNN. In this study, diagnosis of retinal disease using vision transformer was presented using optical coherence tomography (OCT) images. A multi-class classification layer in the vision transformer model was used to group the OCT images into the normal and three abnormal type, Choroidal Neovascularization (CNV), Drusen, and Diabetic Macular Edema (DME). The proposed method achieved a accuracy of 95.76%, sensitivity of 95.77% and specificity of 98.59% in detecting CNV, DME and DRUSEN. Results showed that the classification accuracy of vision transformer is higher than that of other traditional CNN models. The performance of vision transformer was evaluated with different performance metrics like accuracy, sensitivity, and specificity, which proved that vision transformer is a statistically significant method than other standard CNN architectures in classifying retinal diseases using OCT images. This technology enables early diagnosis of retinal diseases, which may be useful for optimal treatment to reduce vision loss.

Keywords: retinal image classification, deep learning, vision transformer, choroidal neovascularization, diabetic macular edema, drusen

1. INTRODUCTION

Retina diseases which are prominently seen in a huge proportion of the population may lead to loss of vision. Vision loss [1] is a huge threat which needs to be detected early and properly diagnosed and treated. From the statistics in ophthalmic study [2], it can be found that most of the eye impairments occurred in the age group over 40 years old. In the human eye, the macular is responsible for creating vivid central vision. However, in the case of macular disease, the vision is disturbed due to the injury to the blood vessels which causes accumulation of vascular fluid beneath the macula region [3]. As shown in Fig. 1, the main ocular diseases affecting macular include choroidal neovascularization (CNV), diabetic macular edema (DME), and drusen.

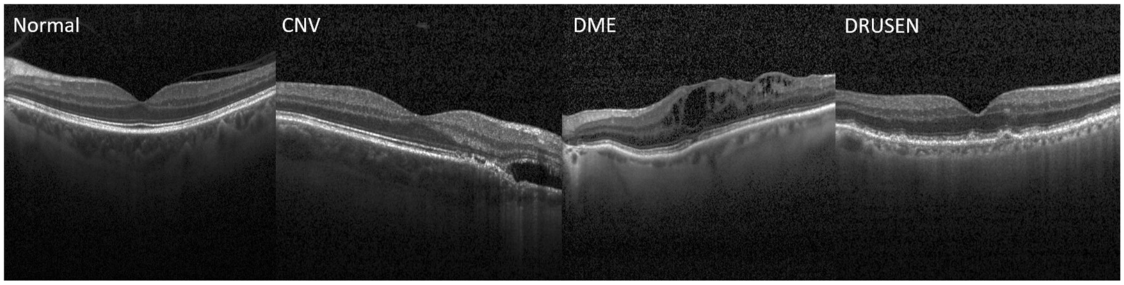


Fig. 1 Sample retina disease images.

Optical coherence tomography (OCT) is a high-resolution imaging technique that uses the coherent light to capture the micrometer-resolution, three-dimensional tomography images of biological tissues. For the human eye, the OCT have the

SPIE-CLP Conference on Advanced Photonics 2022, edited by Xu Liu,
Anatoly Zayats, Xiaocong Yuan, Proc. of SPIE Vol. 12601, 1260102
© 2023 SPIE · 0277-786X · doi: 10.1117/12.2665918

capability of imaging cross-sections of tissue layers, which provides a direct method for evaluating cellular tissue [4] and axonal thickness of macular degeneration [5] and other eye diseases or systemic diseases. Therefore, the OCT imaging technique has been widely used in the identification and treatment of retinal diseases in recently years. In recent years, various deep learning algorithms have been proposed for multi-class retinal diseases classification using OCT images [6]. Rasti et al. trained a CNN ensemble model to predict three retinal states including DME, AMD and healthy [7]. Similarly, Lee et al. presented a convolutional neural network (CNN) based on VGG-16 backbone for the classification of AMD and normal OCT images [8]. Furthermore, Fang et al. proposed a CNN model based on lesion perception for OCT classification [9]. In addition, several proposed schemes used transfer learning to predict retinal diseases [10,11]. In this paper, a method based on vision transformer [12] is proposed to diagnose retina diseases. The vision transformer model aims to learn a mapping for a sequence of image patches to the corresponding semantic label, relying on the attention mechanism.

2. METHOD

2.1 Dataset

The dataset is retrieved from the Kaggle repository named kermany2018. It has three folders (train, test, validation), each with subfolders for four categories of images (NORMAL, CNV, DME, DRUSEN). The dataset consists of 83,484 training images, 41,741 validation images, and 968 test images.

2.2 Vision Transformer

The vision transformer model was proposed by Dosovitskiy in 2020 [12]. The model contains three part: token generation, transformer encoder and classification layer. Fig.2 shows the structure of vision transformer. The input image is separated into several patches called tokens, firstly. Then the linear embedding is performed on each token to consider the position information. A new token called the class token is added to the sequence of the tokens which plays the role of feature expression for each scene. Next the token sequence is sent into the encoder to consider the interaction of the tokens using multi-head self-attention mechanism. Finally, the output class token is classified into different scene categories by a MLP layer. The vision transformer model directly considers the contextual information and the spatial distribution of the objects contained in the image.

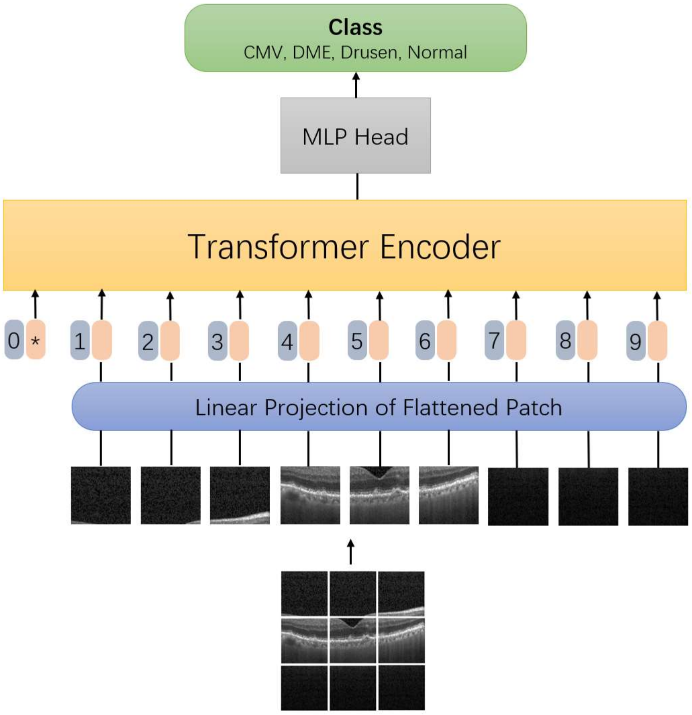


Fig. 2 The structure of vision transformer.

2.3 Symmetric cross-entropy loss function

The symmetrical cross-entropy loss function can minimize the effect of noise on the training set and prevent overfitting [13]. Therefore, we used the symmetric cross-entropy loss function as the loss function of the model to reduce the impact of noise during model training. The definition of the symmetric cross-entropy loss function is:

$$l_{sce} = l_{ce} + l_{rce} \tag{1}$$

l_{ce} here is the cross-entropy loss function and l_{rce} is the reverse cross-entropy function. Their definitions are shown as follows:

$$l_{ce} = -\sum_{k=1}^{K} q(k|x) \log p(k|x) \tag{2}$$

$$l_{rce} = -\sum_{k=1}^{K} p(k|x) \log q(k|x) \tag{3}$$

Where, q(k|x) is the ground truth class distribution conditioned on sample x, whilst p(k|x) is the predicted distribution over labels by the classifier.

3. RESULTS AND DISCUSSION

3.1 Experimental environment

The vision transformer model is simulated using PyCharm software installed on 64-bit windows operating system, with Intel (R) Xeon (R) W-2245 CPU, NVIDIA GeForce RTX 3090 and 32GB memory. Further, we conducted various experiments to evaluate its performance in predicting the candidate retinal conditions (healthy, CNV, DME, and Drusen) using OCT scans.

3.2 Evaluation standard

We adopt the evaluation metrics of classification accuracy, sensitivity and specificity to evaluate the performance of the vision transformer model. Accuracy is defined as the ratio of correctly recognized positive instances and negative instances to total instances. Sensitivity is the detection rate of positive instances and specificity is the detection rate of negative instances. Their definitions are as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{4}$$

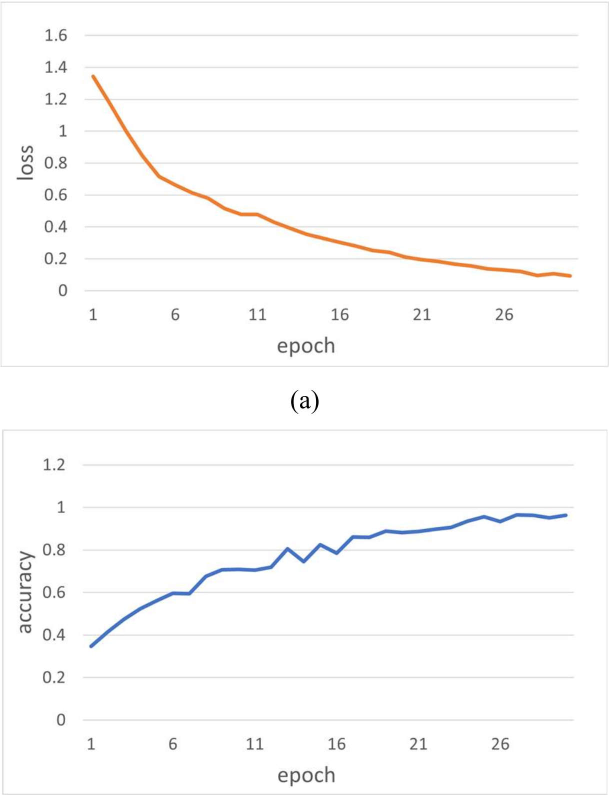
$$Sensitivity = \frac{TP}{TP+FN} \tag{5}$$

$$Specificity = \frac{TN}{FP+TN} \tag{6}$$

where, TP is the true positive instance, TN is the true negative instance, FP is the false positive instance, and FN is the false negative instance. In addition, we also used the confusion matrix, the comparison matrix between the predicted labels and the real labels, to analyze the performance of each category of OCT image.

3.3 Results

We trained the model 30 times to obtain the best results and saved the parameters of the model with the highest accuracy of the model on the validation set. To verify the reliability of the model, accuracy and loss of iterations are shown in Fig. 3. The best classification accuracy rate is 96.57% and the loss is 0.0929.



(a)

(b)

Fig. 3 The loss changes (a) and the accuracy changes (b) of the validation set.

From the above loss changes and accuracy changes, it can be determined that the transformer model is normally fitted without overfitting. The results were summarized in the form of a confusion matrix according to real labels. The columns of the matrix represent true labels, and the rows represent predicted labels. Confusion matrix for diagnosing retinal diseases is shown in Fig. 4.

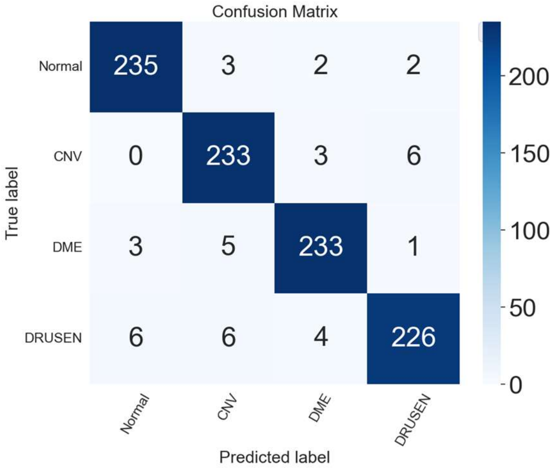
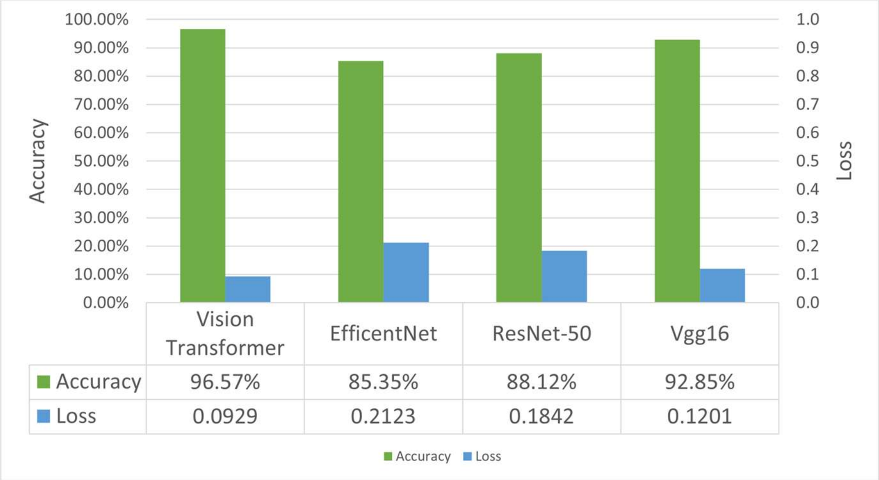


Fig. 4 The confusion matrix of the test set.

As shown in Fig. 5, the proposed method can achieve classification accuracy of 95.76%, sensitivity of 95.77% and specificity of 98.59% on the test set. In this study, the model presented in this paper was compared with the traditional CNN to verify the effectiveness of the method proposed. We choose EfficentNet, Resnet50, and VGG16 as the comparison models. The experimental steps and hyperparameters of these models are consistent with the vision transformer. From the graphical comparison of the validation set presented in Fig. 5, it is shown that the validation loss of vision transformer for classifying retinal diseases is 0.0929, which is less than that of Vgg16 by 0.0272 and ResNet50 by 0.0913. Compared with Vgg16 and ResNet50, the accuracy is increased by 3.72% and 8.45%, respectively.



	Vision Transformer	EfficentNet	ResNet-50	Vgg16
🟩 Accuracy	96.57%	85.35%	88.12%	92.85%
🟦 Loss	0.0929	0.2123	0.1842	0.1201

Fig. 5 The accuracy and loss comparison of vision transformer with the other CNN models.

The accuracy comparison of validation set and test set between vision transformer and five CNN models are shown in Table 1.

Table 1. Model comparison of accuracy

Model	Accuracy	
	Validation set	Test set
EfficientNet	85.35%	83.25%
ResNet50	88.12%	85.39%
Vgg16	92.85%	91.42%
Vision transformer	96.57%	95.76%

As we can see from Table 1, in the CNN image classification models, VGG16 has the highest classification accuracy of 91.42% in test set which is still less than the classification accuracy of vision transformer. Through the attention mechanism, vision transformer can concentrate on the image region related to the semantics of the classification target, thereby obtaining higher accuracy. Considering the recognition accuracy, vision transformer is superior to CNN models in diagnosis of retinal disease.

4. CONCLUSIONS

As shown above, the vision transformer model outperformed the EfficentNet, RESNET-50, and Vgg16 models in classification of NORMAL, CNV, DME, and DRUSEN OCT images. The vision transformer model was properly trained and high accuracy in multiclass classification up to 95.76% was obtained. The performance of vision transformer was evaluated with different performance metrics like accuracy, sensitivity, and specificity, which proved that vision transformer is a statistically significant method than other standard CNN architectures in classifying retinal diseases using OCT images. This technology enables early diagnosis of retinal diseases, which may be useful for optimal treatment to reduce vision loss.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (62175156, 81827807), Science and Technology Commission of Shanghai Municipality (22S31903000) and Collaborative Innovation Project of Shanghai Institute of Technology (XTCX2022-27).

REFERENCES

[1] D. G. Miller, and L. J. Singerman, “Vision loss in younger patients: a review of choroidal neovascularization,” Optometry and Vision Science, 83(5), 316-25 (2006).

[2] Y. Ikuno, Y. Jo, T. Hamasaki et al., “Ocular Risk Factors for Choroidal Neovascularization in Pathologic Myopia,” Investigative Ophthalmology & Visual Science, 51(7), 3721-3725 (2010).

[3] X. He, L. Fang, H. Rabbani et al., “Retinal optical coherence tomography image classification with label smoothing generative adversarial network,” Neurocomputing, 405, 37-47 (2020).

[4] N. Cuenca, I. Ortuño-Lizarán, and I. Pinilla, “Cellular Characterization of OCT and Outer Retinal Bands Using Specific Immunohistochemistry Markers and Clinical Implications,” Ophthalmology, 125(3), 407-422 (2018).

[5] P. A. Keane, P. J. Patel, S. Liakopoulos et al., “Evaluation of age-related macular degeneration with optical coherence tomography,” Survey of Ophthalmology, 57(5), 389-414 (2012).

[6] B. Hassan, T. Hassan, B. Li et al., “Deep Ensemble Learning Based Objective Grading of Macular Edema by Extracting Clinically Significant Findings from Fused Retinal Imaging Modalities,” Sensors (Basel), 19(13), (2019).

[7] R. Rasti, H. Rabbani, A. Mehridehnavi et al., “Macular OCT Classification Using a Multi-Scale Convolutional Neural Network Ensemble,” IEEE Transactions on Medical Imaging, 37(4), 1024-1034 (2018).

[8] C. S. Lee, D. M. Baughman, and A. Y. Lee, “Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images,” Ophthalmology Retina, 1(4), 322-327 (2017).

[9] L. Fang, C. Wang, S. Li et al., “Attention to Lesion: Lesion-Aware Convolutional Neural Network for Retinal Optical Coherence Tomography Image Classification,” IEEE Transactions on Medical Imaging, 38(8), 1959-1970 (2019).

[10] D. S. Kermany, M. Goldbaum, W. Cai et al., “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” Cell, 172(5), 1122-1131.e9 (2018).

[11] S. P. Karri, D. Chakraborty, and J. Chatterjee, “Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration,” Biomedical Optics Express, 8(2), 579-592 (2017).

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., [An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale], (2020).

[13] Y. S. Wang, X. J. Ma, Z. Y. Chen et al., "Symmetric Cross Entropy for Robust Learning with Noisy Labels," IEEE International Conference on Computer Vision. 322-330 (2019).