Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230823

Research on pathologic myopia recognition based on vision transformer

Chen Yang

School of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 102200, China

2020040267@mail.buct.edu.cn

Abstract: Currently, the diagnosis of pathological myopia is mostly done through manual diagnosis, which not only requires experienced ophthalmologists but is also time-consuming and labour-intensive. In order to improve the diagnostic efficiency and accuracy, and to prevent irreversible visual impairment caused by missed diagnosis, misdiagnosis, and delayed treatment, this paper presents a fine-grained image analysis task of classifying fundus images of patients with pathological myopia and non-pathological myopia. To accurately identify subtle differences in features among similar fundus images, a pathological myopia recognition model based on Vision Transformer (ViT) is proposed. The model incorporates a feature selection module using self-attention mechanism that can effectively select important features in the fundus images, thereby eliminating the influence of irrelevant regions on recognition. Experimental results demonstrate that this method outperforms traditional ViT models, achieving high accuracy in pathological myopia recognition.

Keywords: pathologic myopia, fine-grained, ViT, feature selection.

1. Introduction

Pathological myopia, also known as degenerative myopia, is a type of near-sightedness that goes beyond the normal range and can lead to vision loss. It is caused by excessive elongation of the eyeball, leading to structural changes in the eye. Symptoms of pathological myopia may include blurred vision, difficulty seeing objects at a distance, eye strain or fatigue, headaches, and an increased risk of retinal detachment or other eye complications. As the condition progresses, patients may experience a progressive loss of vision that cannot be corrected with eyeglasses or contact lenses. Early diagnosis and treatment are important to prevent complications and preserve vision.

The difficulty in classifying the fundus images of patients with pathological myopia and non-pathological myopia lies in two main aspects: similarity and variability. For similarity, in both pathological and non-pathological myopia fundus images, the macular area may show some degree of deformation or thinning. In terms of variability, the interpretation of fundus images may be influenced by different factors such as intraocular pressure, age, genetic factors, etc. This can pose a challenge in identifying and classifying pathological myopia and non-pathological myopia.

The current diagnosis of pathological myopia relies mainly on experienced ophthalmologists who manually diagnose patients based on a comprehensive eye examination. This process is not only time-consuming and labor-intensive but also difficult to achieve accurate diagnosis, especially in developing

DOI: 10.54254/2755-2721/15/20230823

countries or impoverished areas with a shortage of experienced ophthalmologists and inadequate medical facilities. This can lead to irreversible vision loss due to delayed treatment. As approximately 89% of people with visual impairments live in low- and middle-income countries [1], visual impairment and blindness remain significant challenges in underdeveloped countries and related poverty-stricken areas. Therefore, an efficient and automated machine diagnosis method is required to assist doctors in making timely diagnosis decisions without requiring massive human involvement or medical device intervention, which can lay a solid foundation for future remote medical assistance.

This paper proposes a ViT-based fundus feature selection model for pathological myopia recognition tasks. The model first divides a complete image into equally sized image patches. Then, these image patches are inputted into the encoder of the Transformer, which learns the self-attention weights between each image patches. Next, the model selects the image patches with higher contribution to the classification based on the self-attention weights. Finally, the selected image patch features are inputted into the classifier. The proposed method using self-attention to select image patches can reduce the influence of similar parts between different categories of fundus images and achieves good classification performance on the iChallenge-PM which is a dataset of pathological myopia recognition.

2. Related Work

Traditional convolutional neural networks, such as AlexNet [2], ResNet [3], and GoogleNet [4], have achieved good classification results for the coarse classification of images. However, traditional classification models extract features for the entire image without focusing on subtle differences in features, which results in these classification networks performing poorly on fine-grained image classification.

In order to extract local detail features with discrimination, many previous methods need to rely on local feature labels of images. Part-based R-CNNs extends R-CNN architecture by introducing a new branch to predict object part locations, in addition to bounding boxes and class labels [5]. This allows for more fine-grained localization and segmentation of objects in images. Besides, based on manually labeled strong part annotations, PS-CNN uses a fully convolutional network to locate parts and a dual-stream classification network to encode features of objects and parts [6]. However, the most common situation of clinical medical image data is with a small amount of labelled data and a large number of raw images and these approaches requires labeling local features of the target, which consumes a significant amount of human labor to annotate data [7].

When it comes to specific diagnostic methods for pathologic myopia, Liu et al. proposed the PAMELA (Pathological Myopia Detection Through Peri-papillary Atrophy) system, which automatically receives retinal fundus images, performs region of interest (ROI) extraction and optic disc segmentation, and uses support vector machine (SVM) to automatically diagnose pathological myopia based on the feature of peripapillary atrophy (PPA) in a dataset containing 80 fundus images [8]. Zhang et al. utilized the Minimum Redundancy-Maximum Relevancy (mRMR) feature selection technique to select and rank candidate features, and then used SVM classifier to diagnose pathological myopia [9-10]. However, these approaches belong to machine learning, which requires manual feature extraction and selection, resulting in relatively high workload.

3. Method

3.1. Vision Transformer (ViT)

Transformer [10] was originally proposed for the field of natural language processing (NLP) and achieved great success in this area. Dosovitskiy et al. were inspired by this and introduced the Vision Transformer (ViT) model [11]. Without modifying the original Transformer structure, this model applies Transformer to the field of vision by dividing images into patches, and has achieved very good results.

An overview of ViT is depicted in Figure 1. The model first crops the original image into fixed-sized patches, these patches are flattened and converted into sequences of embeddings which are fed into a

DOI: 10.54254/2755-2721/15/20230823

transformer encoder. The transformer encoder processes the embeddings through multiple layers of self-attention and feed-forward networks, allowing the model to learn the relationships between the patches and capture spatial information across the image. The output embeddings from the final transformer layer are fed into a feedforward neural network (classifier head) that predicts the class label of the input image.

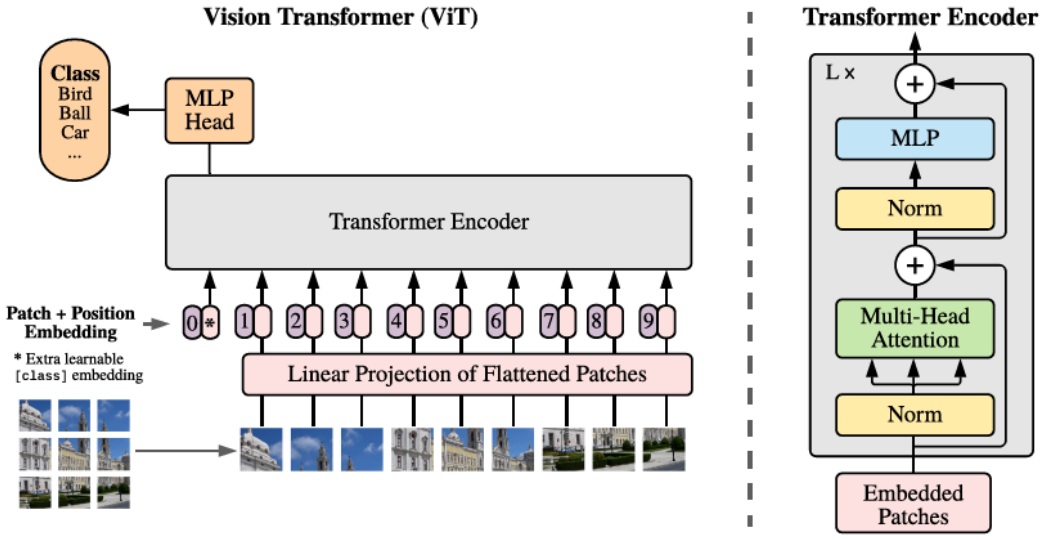

Figure 1. ViT overview.

3.2. Feature Selection Module

Attention mechanism is a computational technique used in machine learning and artificial intelligence to enable models to focus on specific parts of input data while processing information. The attention mechanism assigns weights to different parts of the input data based on their relevance to the task at hand.

The attention mechanism is commonly used in natural language processing (NLP) tasks such as machine translation, sentiment analysis, and text classification. In these tasks, the attention mechanism can help the model to identify the most important words or phrases in a sentence or document. For instance, the self-attention mechanism proposed by Vaswani et al. has improved the performance of Transformers in many natural language processing tasks compared to previous RNN-based approaches [10]. The original ViT model uses self-attention mechanism in the transformer layers. Every transformer layer takes all image patches as input and does not eliminate the influence of irrelevant image regions on the classification results. Pathological myopia recognition belongs to fine-grained image classification, which requires focusing on representative features of the image rather than all regions of the image.

Using the self-attention maps outputted by the Encoder, this paper proposes a self-attention-based feature selection module that selects feature vectors that contribute more to pathological myopia recognition, reducing the influence of useless image patches. An overview of the proposed model is shown in Figure 2.

DOI: 10.54254/2755-2721/15/20230823

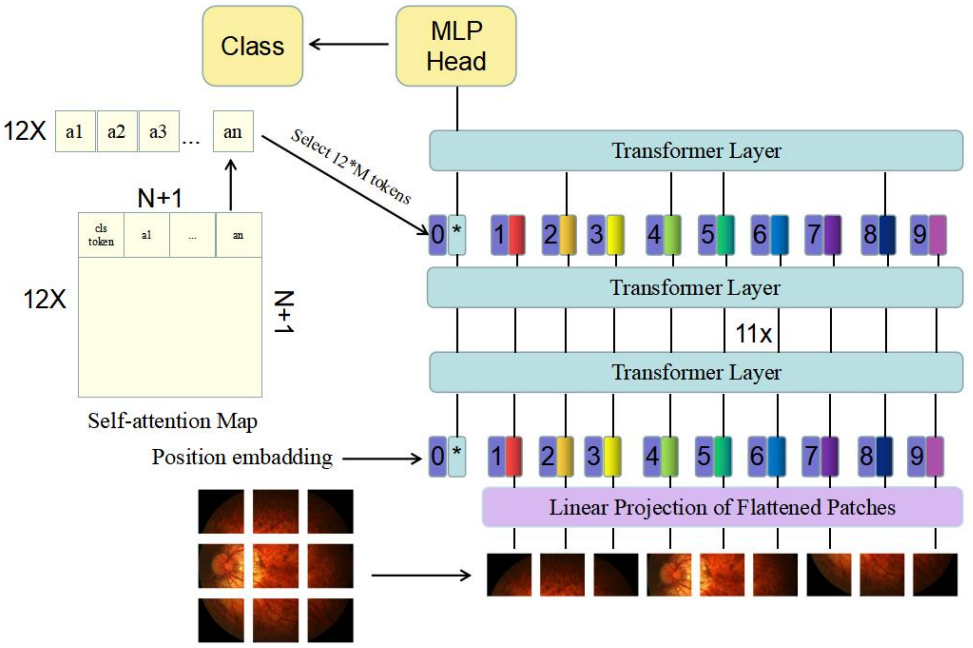

Figure 2. Proposed model overview.

The input image is transformed into (N+1) high-dimensional feature vectors (including a vector that represents the class of the input image) through linear projection. Next, the N feature vectors are inputted into the Encoder of the Transformer. The Encoder uses its Multi-Head self-attention module to calculate the weight size between (N+1) feature vectors and obtains K (K is the number of heads) (N+1) x (N+1) self-attention maps. These self-attention maps represent the correlation between each feature vector and other feature vectors, with a larger attention value between two feature vectors indicating a stronger relationship.

Although N image patches contain different parts of the original image, the contribution of different image patches to pathological myopia recognition is not the same. For example, image patches containing important parts such as the macula, retinal vessels, and optic disc are more important than other image patches. To select important image patches, a feature selection module was added to the penultimate transformer layer, which can pick out image patches that are strongly associated with the class vector. According to the method in TransFG [12]. The original output of layer L-1 is

$$Z_{L-1} = [Z_{L-1}^0; Z_{L-1}^1, Z_{L-1}^2, \dots, Z_{L-1}^N] \tag{1}$$

The attention weights of the previous layers can be written as follows:

$$a_l = \left[a_l^0, a_l^1, a_l^2, \dots, a_l^K\right] \; l \in 1,2, \dots, L-1 \tag{2}$$

The attention weight in each head is

$$a_l^i = \left[a_l^{i_0}; a_l^{i_1}, a_l^{i_2}, \dots, a_l^{i_N}\right] \; i \in 0,1, \dots, K \tag{3}$$

DOI: 10.54254/2755-2721/15/20230823

A matrix multiplication to the raw attention weights in all the layers as

$$a_{\text{final}} = \prod_{l=0}^{L-1} a_l \tag{4}$$

The a_{final} contains K (N+1) x (N+1) self-attention maps. The feature selection module selects the first row of attention values except the first value from each self-attention map. These N attention values represent the correlation between the N feature vectors outputted by the Encoder and class vector. The tokens corresponding to A1, A2,..., AK*M, which are the index of top M maximum values in each row, are selected and concatenated with the classification token as input sequence which is denoted as

$$\boldsymbol{z}_{\text{local}} = \left[z_{L-1}^{0}; z_{L-1}^{A_1}, z_{L-1}^{A_2}, \cdots, z_{L-1}^{A_{K*M}}\right] \tag{5}$$

The feature selection module not only preserves global information, but also allows the model to pay more attention to subtle differences between different categories.

4. Experiments

4.1. Datasets

The dataset used in this experiment is the iChallenge-PM dataset, which is a medical dataset on Pathologic Myopia (PM) provided during the iChallenge competition jointly organized by Baidu Brain and Zhongshan Ophthalmic Center of Sun Yat-sen University. The dataset includes 800 fundus retina images from subjects categorized into two classes: pathologic myopia and non-pathologic myopia. The non-pathologic myopia class includes two sub-classes: highly myopic and normal vision. In this experiment, the dataset was split into a training set of 480 images a validation set of 140 images and a test set of 140 images, with a ratio of 6:2:2.

4.2. System environment and experimental setup

The system is based on a 64-bit Windows operating system, and equipped with an AMD Ryzen 7 4800H CPU and a NVIDIA GeForce RTX 2060 GPU. The image in this experiment is resized to a size of 224*224 and divided into 196 image patches, with each block consisting of 16×16 pixels. The number of heads in the multi-head attention mechanism is set to 12, which means that in the feature selection module, 12 most important feature vectors will be selected from the 196 feature vectors.

The SGD optimizer provided by PyTorch is used in the model training. The hyperparameters of the learning rate are set to 0.001, momentum factor to 0.9, and weight decay to 5E-5. In addition, a learning rate scheduler based on the cosine annealing strategy is also utilized. The period of the cosine function is set to 10, and the scheduler includes a lower bound value, 0.01, which represents the minimum value of the learning rate. This optimizer and scheduler are chosen because they have performed well on similar tasks and datasets, and can effectively control the learning rate and convergence speed of the model. The model is trained for 10 epochs and a pre-trained ViT model—ViT-B_16 provided by Google is used in this experiment.

4.3. Result

In order to compare with the method proposed in this paper, ViT is tested for classification accuracy on the iChallenge-PM dataset. The experimental results are shown in Table 1. The classification accuracy of the pathological myopia recognition model based on ViT proposed in this paper is 94.9%, which is 2.5% higher than using the ViT model alone. This demonstrates that the feature selection module has made good selections of important local features in the images and eliminated the influence of irrelevant features on classification.

DOI: 10.54254/2755-2721/15/20230823

Table 1. Comparison of different methods on iChallenge-PM dataset.

Method	Backbone	ACC (%)
ViT	ViT-B_16	93.1%
Ours	ViT-B_16	97.5%

Figure 3 shows the accuracy and loss values of two models on the validation set.

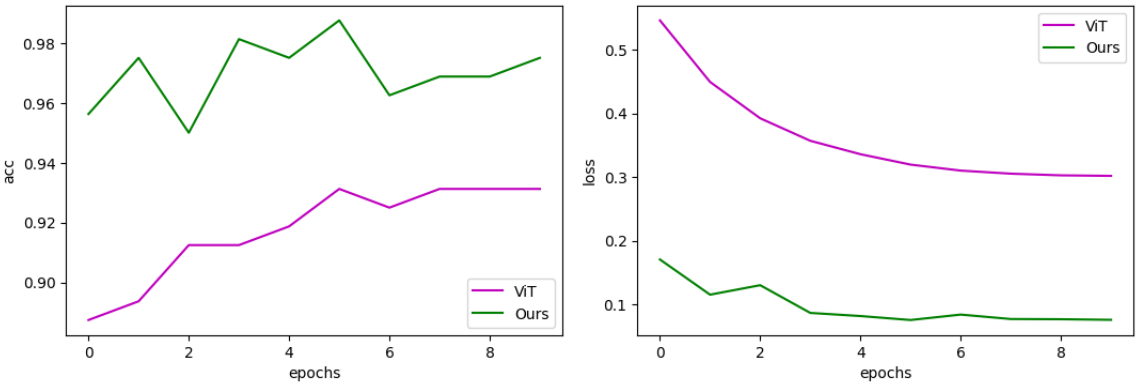

Figure 3. Performance of two models on the validation set.

This experiment also tested the impact of different M values on accuracy in the feature selection module, as shown in Table 2. When M is set to 6 or 8, the model achieves the highest accuracy on the iChallenge-PM dataset, which is 97.5%.

Table 2. Ablation study on value of M on iChallenge-PM dataset.

Value of M	ACC(%)
1	96.9
2	96.9
6	97.5
8	97.5
10	97.3

5. Conclusion

This paper proposes a ViT-based model which contains the feature selection module and study the impact of the number of selected features on the classification accuracy to address the issue of pathological myopia recognition. The feature selection module uses self-attention mechanisms to select important parts of the image, reducing the impact of irrelevant areas on classification. This approach demonstrates good recognition of similar but different classes of fundus images. On the iChallenge-PM dataset, the proposed model has higher classification accuracy compared to the ViT model.

References

[1] Zheng L, Yang Y, Tian Q. SIFT meets CNN: A decade survey of instance retrieval. 2017, IEEE transactions on pattern analysis and machine intelligence, 40(5): 1224-1244.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural

DOI: 10.54254/2755-2721/15/20230823

networks. 2017, Communications of the ACM, 60(6): 84-90.
[3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition, 2016, IEEE conference on computer vision and pattern recognition. 770-778.
[4] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. 2015, IEEE conference on computer vision and pattern recognition. 1-9.
[5] Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection,2014, Computer Vision–ECCV 2014: 13th European Conference: 834-849.
[6] Huang S, Xu Z, Tao D, et al. Part-stacked CNN for fine-grained visual categorization, 2016, Proceedings of the IEEE conference on computer vision and pattern recognition. 1173-1182.
[7] Wang Z, Li T, Zheng J Q, et al. When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation, Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, 424-441.
[8] Liu J, Wong D W K, Lim J H, et al. Detection of pathological myopia by PAMELA with texture-based features through an SVM approach. 2010, Journal of Healthcare Engineering, 1(1): 1-11.
[9] Zhang Z, Cheng J, Liu J, et al. Pathological myopia detection from selective fundus image features. 2012 7th IEEE Conference on Industrial Electronics and Applications: 1742-1745.
[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 2017, Advances in neural information processing systems, 30.
[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2020 arXiv preprint arXiv:2010.11929.
[12] He J, Chen J N, Liu S, et al. Transfg: A transformer architecture for fine-grained recognition. 2022, Proceedings of the AAAI Conference on Artificial Intelligence. 36(1): 852-860.