Computer Methods and Programs in Biomedicine 220 (2022) 106832

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Towards more efficient ophthalmic disease classification and lesion location via convolution transformer

Huajie Wen[a,b], Jian Zhao[a], Shaohua Xiang[a], Lin Lin[a], Chengjian Liu[a], Tao Wang[a], Lin An[c], Lixin Liang[a,*], Bingding Huang[a,*]

[a] College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China
[b] College of Applied Science, Shenzhen University, Shenzhen 518060, China
[c] Guangdong Vision Medical Science & Technology Co., Ltd. Foshan 528000, China

ARTICLE INFO

Article history:
Received 13 August 2021
Revised 1 April 2022
Accepted 21 April 2022

Keywords:
Retina OCT images
Ophthalmic disease classification
Convolution neural network
Transformer
Self-attention

ABSTRACT

Objective: A retina optical coherence tomography (OCT) image differs from a traditional image due to its significant speckle noise, irregularity, and inconspicuous features. A conventional deep learning architecture cannot effectively improve the classification accuracy, sensitivity, and specificity of OCT images, and noisy images are not conducive to further diagnosis. This paper proposes a novel lesion-localization convolution transformer (LLCT) method, which combines both convolution and self-attention to classify ophthalmic diseases more accurately and localize the lesions in retina OCT images.
Methods: A novel architecture design is accomplished through applying customized feature maps generated by convolutional neutral network (CNN) as the input sequence of self-attention network. This design takes advantages of CNN's extracting image features and transformer's consideration of global context and dynamic attention. Part of the model is backward propagated to calculate the gradient as a weight parameter, which is multiplied and summed with the global features generated by the forward propagation process to locate the lesion.
Results: Extensive experiments show that our proposed design achieves improvement of about 7.6% in overall accuracy, 10.9% in overall sensitivity, and 9.2% in overall specificity compared with previous methods. And the lesions can be localized without the labeling data of lesion location in OCT images.
Conclusion: The results prove that our method significantly improves the performance and reduces the computation complexity in artificial intelligence assisted analysis of ophthalmic disease through OCT images.
Significance: Our method has a significance boost in ophthalmic disease classification and location via convolution transformer. This is applicable to assist ophthalmologists greatly.[1]

1. Introduction

Optical coherence tomography (OCT) imaging technology is now considered the standard technique used in clinical ophthalmology to examine the retina and evaluate the response to treatment [1]. Many people suffer from Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR.), which associate with increased Alzheimer's disease risk [2]. The common clinical characteristic of AMD is the existence of drusen, and advanced AMD is called Choroidal Neovascularization (CNV). CNV is usually resulted from subretinal and intraretinal fluid accumulation, Retinal Pigment Epithelium (RPE) detachment and fibrotic scars [3]. Diabetic Macular Edema (DME) is caused by abnormal leakage from damaged retinal blood vessels, whose manifestations are fluid-filled cysts and hard exudates in the retina, as well as retinal thickening [4]. The characteristic of drusen is asymptomatic deposition of extracellular substances between the RPE and the Bruch's membrane collagen layer [5]. Typical diseased OCT images of CNV, DME, and drusen in the macular area are shown alongside a normal image in Fig. 1.

Retina imaging is based on the interference of near infrared weak coherent light irradiating to the surface of retina. The surface roughness at different positions of the retinal tissue is different, and there are a large number of scattering particles inside the tissue. These particles will scatter the incident light multiple times

Abbreviations: AMD, Age-related Macular Degeneration; CNN, Convolutional Neutral Network; CNV, Choroidal Neovascularization; DME, Diabetic Macular Edema; DR, Diabetic Retinopathy; LLCT, Lesion-Localization Convolution Transformer; OCT, Optical Coherence Tomography.

* Corresponding authors.
E-mail addresses: lianglixin@sztu.edu.cn (L. Liang), huangbingding@sztu.edu.cn (B. Huang).

https://doi.org/10.1016/j.cmpb.2022.106832
0169-2607/

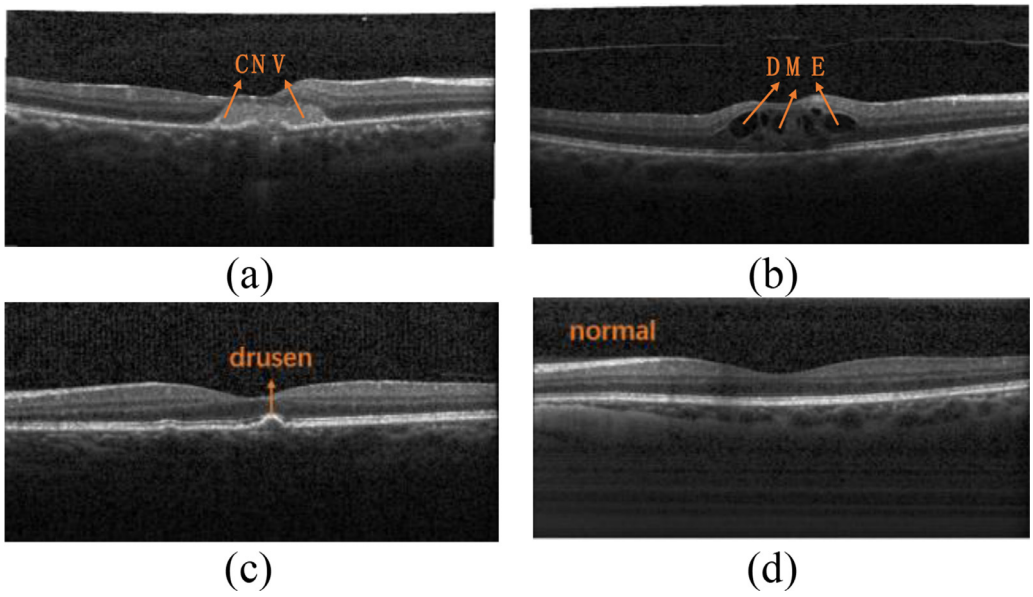


Fig. 1. Four different types of OCT images: (a) CNV, (b) DME, (c) drusen, and (d) normal.

without directionality, creating a large amount of speckle noise in the image, which affects the downstream classification performance and the lesion locations. Many efforts have been made to detect different ophthalmology diseases. However, it is not easy to achieve good detection performance as retinal OCT images have significant speckle noise, irregularity, and inconspicuous features. Schmitt et al. analyzed characteristics of speckle noise in OCT images and present a method to suppress speckle noise, including polarization diversity, spatial compounding, frequency compounding, and digital signal processing [6].

With the advancement of artificial intelligence technology, it has become possible to use machine learning or deep learning for OCT image analysis. Because the sizes and shapes of lesions in the macular area are different, and the structure of a lesion is complex, visual or statistical methods for analysis are not conducive to improving recognition accuracy, sensitivity, specificity, and robustness. Our work designs a novel learning architecture combining both convolution neural network (CNN) and transformer network to form a lesion-localization convolution transformer network named LLCT. This LLCT network is used to classify OCT images and output the corresponding probability of CNV, DME, drusen, and normal cases, respectively. Experimental results show that our proposed method can effectively improve classification accuracy, precision, sensitivity, and specificity. The contributions of our work are summarized as follows:

First, we creatively apply the pixels in the feature map generated by CNN as the input sequence of the transformer network, which can recognize the underlying features of the image, reduce length of the input sequence of the transformer, and take spatial relations and long-term dependency into considerations during learning. This will enhance computing efficiency and analysis performance. As we know, CNN cannot consider spatial relations and long-term dependence.

Second, extensive experiments demonstrate that our proposed method achieves improvement of about 7.6% in overall accuracy, 10.9% in overall sensitivity, and 9.2% in overall specificity compared to other architectures such as HOG-SVM, transfer learning, VGG16, AlexNet, and LACNN [7–11], which has a significance boost in performance.

Third, our LLCT network can localize lesions according to the gradient of self-attention, without the labeling of lesion locations (depending only on the disease class label), enabling ophthalmologists to focus attention on the target region of OCT images for further diagnosis.

The remainder of this paper is organized as follows. We introduce related work in Section 2. Our LLCT method for ophthalmic disease classification and location is presented in Section 3. Experimental results and analysis are performed in Section 4. We conclude our work in Section 5.

2. Related work

Automatic diagnosis of retinal OCT images usually takes machine learning or deep learning approaches. With machine learning approaches, Thomas et al. proposed a method to detect the retinal pigment epithelium (RPE) layer and use statistical methods and randomization to analyze AMD diseases. The method has relatively high accuracy in detecting drusen and requires no training, but some thresholds must be set manually [12]. Srinivasan et al. used the multi-scale histogram of the oriented gradient descriptor (HOG) to extract features from DME, dry AMD, and normal images. This method uses a support vector machine (SVM) and can detect DME images with 81.6% accuracy [13]. Naz et al. proposed a method to detect DME diseases using different feature sets and used an SVM classifier to achieve an accuracy of 79.25% [7]. These are all machine learning approaches. However, machine learning methods usually require us to set parameters manually and it is difficult to improve the accuracy and performance.

With deep learning approaches, researchers usually apply CNN with different architectures. For example, Lee et al. proposed to use VGG16 for OCT image classification of AMD disease, achieving an accuracy of 86.73% [14]. Karri et al. proposed to use GoogleNet for the recognition of OCT images and used the AdaGrad optimizer to achieve 99%, 89%, and 86% accuracy for normal, AMD, and DME classifications, respectively on their dataset [8]. Kermany et al. used transfer learning techniques and achieved 96.6% accuracy on the classification of CNV, DME, drusen, and normal [15]. Rasti et al. introduced and analyzed the multi-scale convolutional mixture of expert (MCME) model in the retinal OCT classification problem, achieving average precision of 98.86% on a dataset of 148 and 45 retinal OCT scans, including dry AMD, DME, and normal [16]. Seebock et al. proposed an unsupervised identification of retinal OCT imaging, with accuracy of 81.4% on normal, early AMD, and late AMD [17]. For the possibility of gradient loss and explosion caused

by too many network layers, He et al. proposed residual connection [18]. These deep learning methods represented by CNN cannot take the long-term dependence and spatial relationship into considerations, and their computation complexity is usually high.

Transformer network proposed by Vaswani et al. uses an attention mechanism. It is mainly applied in the natural language processing (NLP) field, and later developed into visual field [19, 20]. Recently some researchers apply it in medical field. Chen et al. proposed TransUnet, which combines a transformer and U-net, as a strong alternative for medical image segmentation [21], solving long-term dependency and achieving performance superior to various methods on medical applications including multi-organ and cardiac segmentation. Wu et al. proposed a feature map as the transformer input, extracting visual tokens from all feature maps, merging them with one transformer, and projecting them back to the original feature maps, with significant accuracy improvements across tasks and datasets [22]. However, Wu et al. introduced convolutions to vision transformers without position embedding. In each stage, they progressively decreased the token sequence length while increased the token feature dimension, achieving superior performance with good computation efficiency [23]. Zhang et al. combined transformers and CNNs with simple late fusion for medical image segmentation, consisting of parallel CNN and transformer branches, achieving state-of-the-art performance on polyp segmentation while reaching a compromise between the size of model parameters and the speed of inference [24]. Since CNN performs image classification or segmentation based on local features, Transformer-based semantic modeling is based on image global features for classification or segmentation. The current Transformer-based method applied to the field of medical image segmentation considers global features and long-distance semantic relationships, so the semantic segmentation performance of lesions is improved compared to the CNN case. However, B-Scans images contain a lot of speckle noise, experiments show that the classification results are close to CNN when only Transformer is used for OCT image classification. Therefore, we use CNN for low-level feature extraction, and perform global semantic modeling of OCT images in Transformer, which is beneficial to the classification and lesion localization performance of the model. Our method converts pixels of a feature map into tokens as the input sequence of the transformer. This can significantly reduce the length of the input sequence and the computation complexity. The network of LLCT includes not only forward propagation for image analysis, but also partial backward propagation for calculating the gradient of self-attention to locate the lesions. Meanwhile, our method can improve the semantic relationship among tokens because of the overlap between convolutions.

3. Methods

3.1. System overview

Our system is composed of four modules:

- The **Input Module** performs operations such as size cropping and data format conversion on OCT images.
- The **Image Preprocessing Module** selects and subdivides the region of interest (ROI), extracts image features with crucial information, and enhances the features to improve the signal-to-noise ratio (SNR).
- The **Pathological Analysis Module** is used for the pathological analysis of preprocessed images to distinguish the CNV, DME, drusen, and normal cases and output the corresponding probabilities. In addition, the self-attention gradient is calculated based on the analysis results to locate the lesions.
- The **Result Output Module** displays images and analysis results to facilitate further manual judgment.

The output is the lesion location, the classification of CNV, DME, drusen, and normal cases and the corresponding probability.

The process of ophthalmology disease inference on OCT images through our system includes three steps. First, it extracts the ROI of OCT images, i.e., the area of retinal information. This reduces the influence of non-ROI areas on model training and disease inference. Second, it applies a CNN to generate feature maps as the input sequence to the transformer network, which improves performance by reducing the size of inputs and enhancing the semantic relationship between tokens. Finally, it applies the self-attention method to infer ophthalmology diseases through OCT images.

3.2. Image preprocessing

We first preprocess the OCT images for training. This includes extraction of the ROI and data augmentation through ROI enhancement.

We use computer vision techniques to generate an image mask. The ROI is easily extracted by an AND operation between the original image and the mask image. Such ROI extraction is able to remove the impact of the noise in the non-ROI area on model training and inference. Meanwhile, we perform feature enhancement on ROI of OCT images to augment training dataset.

3.2.1. ROI extraction

Extracting the ROI includes the following steps: median filtering, binarization, contours and filling, opening and closing, AND operation.

Step 1 Median Filter: The median filter is applied to the original images to reduce noise.

Step 2 Binarization: The maximum-variance algorithm in all classes is applied to calculate the threshold for image binarization [25]. By setting the threshold k, the image pixel is considered as the foreground when its gray level is greater than k, otherwise it is considered as the background. The variance of the foreground and background pixels is denoted as σ^2, we have

$$\sigma^2 = p(f_a - g)^2 + (1-p)(b_a - g)^2 \tag{1}$$

where g is the average gray level of the image, which is $\sum_{i=0}^{L-1} i\mu_i$, L is the gray level of OCT images, and μ_i is the probability of a pixel in gray level i; f_a and b_a are the average gray level of pixels in the foreground and background images respectively, and p is the probability that a pixel is in the foreground.

Variance σ^2 is a measure of the uniformity of the grayscale distribution. The larger the inter-class variance between the background and foreground, the more difference between the two parts of the image. When part of the foreground is wrongly considered in the background, or vice versa, the difference between the two parts becomes smaller. We use Algorithm 1 to determine the optimal threshold k which achieves the largest variance.

Algorithm 1: Image Foreground and Background Threshold Determination.

Initialization: $\sigma^2 \leftarrow 0$, $k \leftarrow 0$
for all $k^* \in [0, 255]$ **do**
  Calculate the variance $\sigma^2(k^*)$ according to formula (1).
  if $\sigma^2(k^*) > \sigma^2$ **then**
    Update $\sigma^2 \leftarrow \sigma^2(k^*)$, $k \leftarrow k^*$
  end if
end for
Binarize the image according to the following formula:

$$f(x) = \begin{cases} 255, & x > k \\ 0, & x \le k \end{cases}$$

Step 3 Contours and Filling: By searching and filling the contour of the binarized image, a closed continuous area is formed inside the mask image, reducing the impact of feature loss.

Step 4 Opening and Closing: The noise of the mask image is reduced by setting the structure element with a size of 5 × 5 to open and close the mask image.

Step 5 AND Operation: AND operation is performed on the original and mask images to extract the ROI.

3.2.2. Data augmentation through ROI enhancement

To increase the diversity of training dataset, we use both feature extraction and enhancement to increase the number of training images. We take the method of enhancing the image contrast to obtain new images. Both original and new images are applied together to train the model. The preprocessed OCT images can also be applied in the inference phase.

The contrast of OCT images is enhanced as follows:

Step 1: Count the number of pixels in different gray level of the original image and sort the gray values from small to large.

Step 2: Calculate the probability of the gray level for pixels.

Step 3: Calculate the cumulative probability distribution and convert the pixel gray level,

$$g(t) = L\sum_{i=0}^{t} p_i \tag{2}$$

where L is the maximum gray level of OCT images, t means gray level t, and p_i is the probability of the gray level i in the original image, and $g(t)$ is the converted pixel gray level.

OCT image feature extraction and enhancement are described in Algorithm 2.

Algorithm 2: OCT image feature extraction and enhancement.

1 **Initialization:** image mask ∈ ←0

2 **Input:** the image for processing $\phi$with the size [*W, H, C*].
3 **Output:** the image after processing δ with the size [*W, H, C*].
4 $t \leftarrow$ *m*median_filter (ϕ)
5 ((t) according to Algorithm 1.
6 Obtain all the contour set *con_set* ← contours (t)
7 **for** *c* in *con_set* **do**:
8 Update ∈← *draw_contours*(c)
9 **end for**
10 Update ∈← *contours_filling*
11 Update ∈← *open_operation*(∈)
12 Update ∈← *close_operation*(∈)
13 Update $\delta \leftarrow$ *and_operation*(φ, ∈)
14 Update $\delta \leftarrow$ *Convert*the gray level δ according to formula (2)

3.3. Lesion-localization convolution transformer (LLCT) for OCT image analysis

We design a novel self-attention deep learning network named LLCT to effectively detect various macular diseases and precisely locate the lesions. The LLCT architecture is illustrated in Fig. 2. The network consists of the two parts of convolution neural network and transformer. The former part is used to recognize the underlying features of the image, such as contours, edges, and colors. The convolution operation outputs feature maps. By treating each pixel of a feature map as one token in the input sequence for the transformer, the length of the input sequence will be reduced and the semantic relationship between tokens will be enhanced. The latter part is a transformer, which encodes pixels in the feature map to generate tokens and embeds learnable position information to tokens, and then uses a multi-head attention mechanism to achieve self-supervised learning. It solves the long-term dependence problem of CNN, and improves the ability to detect macular lesions. The locations of CNV, DME and drusen lesions are computed based on the classification results and gradient of self-attention.

3.3.1. Feature map generation through CNN

In our designed LLCT network, we first apply a customized CNN to obtain feature maps from OCT images. The CNN first performs convolution with kernel size of 7 × 7, batch normalization, and ReLU activation on the input OCT image with the width of 512, the height of 512 and the channel of 3 (512 × 512 × 3, W × H × C), which outputs feature maps downsampled by a factor of 2, i.e., outputs feature maps with size of 256 × 256 × C, and the output channel C is 64. The max pooling layer performs a similar downsampling operation. The output of max pooling is then connected to multiple residual blocks. Each residual block is composed of two groups of convolutions with kernel size of 3 × 3, batch normalization and ReLU activation. To recognize as many low-level features of OCT images as possible, the number of convolutions in the network is increased. In order to accelerate convergence and reduce the possibility of gradient loss and explosion, the skip connection technique was employed to connect the feature maps in every residual block. In the fourth, ninth, and fifteenth residual blocks, a convolution kernel with strides of 2 is used to achieve downsampling, and the feature maps' channel dimensions are increased by a factor of 2, i.e., the feature map's size changes from 128 × 128 × 64 to 16 × 16 × 512 at last. The remaining residual blocks will unchange the feature map in all dimensions. Finally, the output of the convolutional layer is a feature map with the size of 16 × 16 × 512 (W × H × C).

3.3.2. Ophthalmology disease classification through transformer

The feature map output by CNN is flattened into a sequence of pixels, i.e., the feature map's size changes from 16 × 16 × 512 to 256 × 1 × 512 (W × H × C). In the linear projection with a flattened pixel E, we adjust the dimension of the feature map to 256 × 1 × 768 (W × H × C), and regard each pixel in the feature map as a token in the input sequence for the transformer, so the length of the sequence is 256, and each token, x_p^i ($1 \le i \le 256$), in the sequence is represented by a 768-dimensional vector. We add position embedding ϕ to the input sequence and realize the learnable position coding of the token information. The model can take position dependence into considerations:

$$\mu = [x_1E + \varphi_1;\ x_2E + \varphi_2; \ldots;\ x_{256}E + \varphi_{256}] \tag{3}$$

The transformer encoder consists of 12 layers of blocks. Each block consists of multi-head self-attention and multilayer perceptron (MLP), which are connected with layer normalization and skip connections. The mechanism of multi-head attention is used to improve the performance of the self-attention layer. It is implemented on different heads through different queries, keys, and value matrices. The MLP contains two hidden layers with GELU activation. In the transformer encoder, we use 12 heads in multi-head attention, and the dimension of each layer of the MLP is 3072. The attention mechanism of multi-head attention is

$$Attention(Q_i, K_i, V_i) = Softmax\left(\frac{\left[\mu w_q^i\right]\left[\mu w_k^i\right]^T}{\sqrt{d_k}}\right)\left[\mu w_v^i\right] \tag{4}$$

where w_q^i, w_k^i, w_v^i are the weight matrices of head i; d_k is the dimension of single head; and μ is the feature map.

The attention mechanism in single head enables us to look at the lesion from the aspect of the shape and thickness of the retina with considerations of long-term dependence. This can classify the OCT images according to the global features instead of the local features, and improve the ability to detect the lesion. Further by using the multi-head attention, which consists of multiple independent single head self-attention, different position and different angles of OCT images can be focused on, the ability of spatial interpretation is improved.

The output sequence of the last block in the transformer encoder is used as the input for the average pooling to reduce the

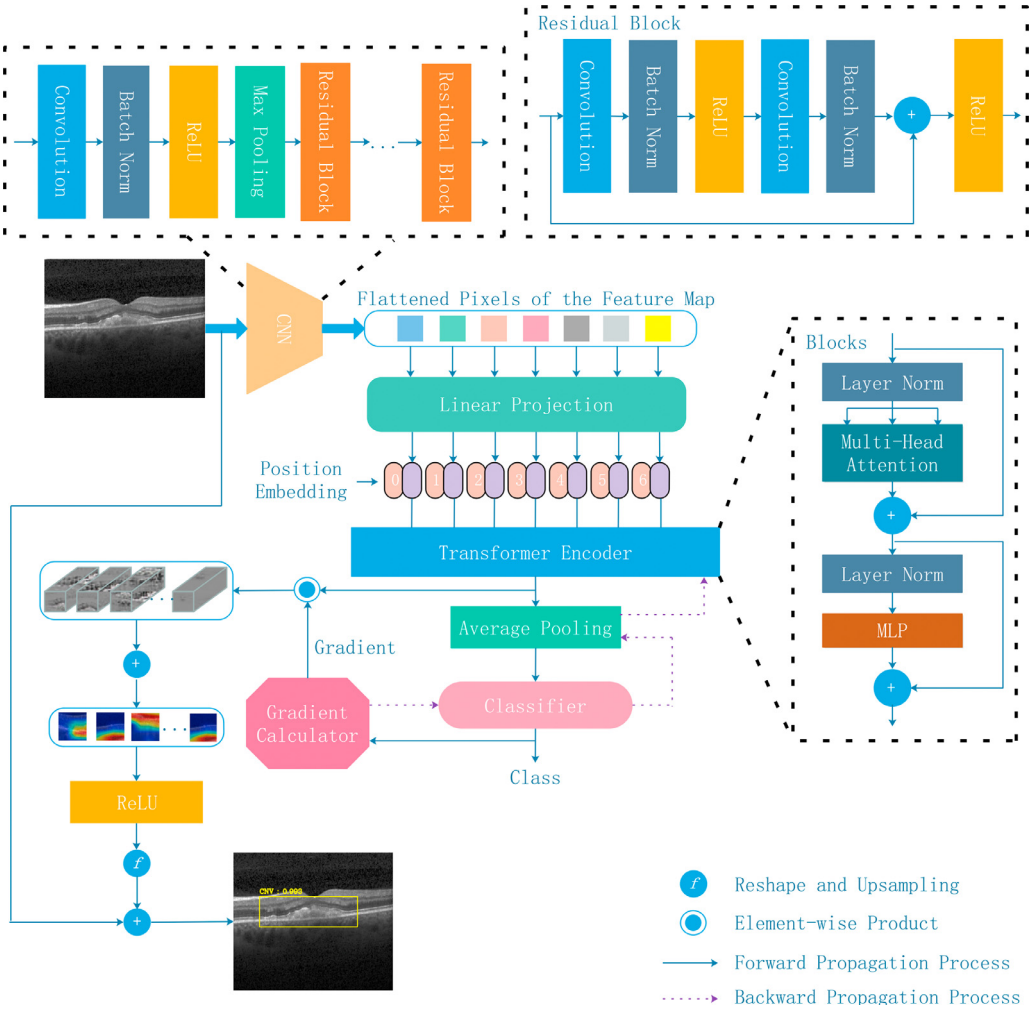


Fig. 2. The architecture of LLCT. The number of residual blocks is 17 and transformer encoder blocks is 12. The input resolution of the image is 512 × 512 with RGB. The skip connection is used on every residual block. The gradient calculator calculates the self-attention gradient of the last Transformer encoder block according to the backward propagation of the classification result, and multiplies it with the forward propagation to locate the lesion.

length of sequence, and then connect to the classifier to classify the four cases of CNV, DME, drusen and normal. The classifier is composed of a single hidden layer. The output of the classifier is used as the input for the Softmax to compute the corresponding probability and output the classification with the highest probability.

The loss function is used to evaluate the degree of difference between the predicted and true values of the model. Currently LLCT uses cross entropy as the loss function.

3.3.3. Lesion localization

Besides ophthalmology disease classification, our designed LLCT network is able to locate the lesion of CNV, DME, and drusen.

The detail steps for lesion localization are as follows:

Step 1: According to formula (5), calculate the average gradient α_k^c of dimension k in sequence A of the last transformer encoder block. Here y is the output score, c is the disease type, and A_i^k is the token i on the k-th dimension of sequence A.

$$\alpha_k^c = \frac{1}{i}\sum_i \frac{\partial y^c}{\partial A_i^k} \tag{5}$$

The computation process of $\partial y^c / \partial A_i^k$ is as follows: according to the classification results, the computer graph of partial derivative is calculated from the classifier to the last block of transformer encoder.

Step 2: Multiply α_k^c by A^k.

Step 3: Sum up $\alpha_k^c A^k$ of each dimension k,

$$S = \sum_k \alpha_k^c A^k \tag{6}$$

Step 4: The ReLU activation function is used on S getting from formula (6). This will get a map of the lesion location. By reshaping and upsampling the lesion location map to the input size of the original image, the lesion location of the original image is obtained.

Step 5: After processing the output result of step 4 with algorithm 1, we will get a rectangular convex hull, which is used to locate the lesion.

4. Experimental results

4.1. Validation dataset

We validate our proposed model on the publicly available OCT dataset [15], which consists of 84,484 OCT B-scans (37,205 CNV, 11,348 DME, 8616 drusen, and 51,140 normal) acquired from 4686 adult patients without exclusion criteria based on age, gender, or race at the Shiley Eye Institute of the University of California San Diego, California Retinal Research Foundation, Medical Center Ophthalmology Associates, Shanghai First People's Hospital, and Beijing Tongren Eye Center between July 1, 2013, and March 1, 2017 [9]. To

evaluate the accuracy, precision, sensitivity, and specificity of classification, we randomly selected 1000 B-scans (250 each of CNV, DME, drusen, and normal) as the testing set, and ensured that the scans from the same patient appear either in the training dataset or the testing dataset. The rest constituted the training set; hence, there is an imbalanced number of training datasets among different classes.

4.2. Experimental settings

The model was optimized using the Adam optimizer [26], training on randomly selected samples from the training set. The initial learning rate was set to 0.002, the batch size to 256, the weight decay factor to 0.0001, the attention dropout rate to 0, the MLP dropout rate to 0.1, and the iterative training time to 80. The Xavier algorithm [27] was used for weight initialization of the transformer encoder, and the Kaiming normalization algorithm was used for convolution initialization. To enhance the diversity of the training dataset with limited photos, we used data enhancement methods to process the images, including random horizontal flips and random resized crops. Source images and preprocessed images were combined and transmitted to the model,

$$T_{training\_set} = p + D(p) \tag{7}$$

where $T_{training\_set}$ is the training set, p is the original dataset and $D(p)$ is the dataset after preprocessing. The original OCT images were resized to 512 × 512 as input. The hardware of the training machine consisted of an Intel Xeon E5–2698 v4 with 256 GB RAM and four Nvidia Tesla V100 GPUs. The operating system was Ubuntu 16.04 LTS, and the deep learning framework was PyTorch 1.7.0.

The performance of the model in classification was evaluated using accuracy, precision, sensitivity, and specificity,

$$Accuracy = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \tag{8}$$

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{9}$$

$$Sensitivity = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{10}$$

$$Specificity = \frac{N_{TN}}{N_{TN} + N_{FP}} \tag{11}$$

where N_{TP}, N_{TN}, N_{FP}, N_{FN} are the numbers of true positive, true negative, false positive, and false negative, respectively. At the same time, the time required for each iteration of training and inference for a typical model and the proposed model was recorded.

4.3. Compared methods

We compared the performance of LLCT with that of other networks in OCT image classification, such as histogram of oriented gradients—support vector machine (HOG-SVM) [7], Inception V3 [8], visual geometry group 16 (VGG16) [10], lesion-aware convolutional neural network (LACNN) [9], vision Transformer (ViT) [19], and AlexNet [11]. HOG-SVM is based on machine learning, and the others are based on deep learning. In the HOG-SVM method, the feature vector of the OCT image was extracted with a multiscale HOG descriptor, and transmitted to the SVM for classification. The Inception V3 model was pre-trained on the ImageNet dataset, utilizing transfer learning to freeze all convolutional layers and update the last full connection layer for the classification of the OCT image. It can reduce computation and accelerate convergence. LACNN applies the method of multi-task learning. Firstly, it uses the lesion detection network constructed by CNN to detect the lesion, extract the feature map of B-scans image, and then use the lesion-attention model composed of CNN for classification. ViT uses image patch for embedding and adds positional encoding, and uses self-attention mechanism for image global feature modeling. It adds a class token before the Token of image patches, and classifies according to the class token in the classifier. Vit and LLCT have the same hyperparameter settings. The VGG16 and AlexNet models consist of multiple convolutional, pooling, and fully connected layers with different architectures. The hyperparameters of VGG16 and the Inception V3 are set as the same in LACNN [9]. The average training time per iteration and inference time in classification of the network based on the classic CNN structure (Inception V3, VGG16 and AlexNet), the ViT network based on the self-attention mechanism, and the proposed model are shown in Table 2. The reason why ViT model requires highest computational cost during training is that it uses image patches as embeddings and the semantic relationship between tokens is weak. The training cost of our proposed model LLCT is between typical CNNs and ViT. It shows that the inference performance of our proposed model LLCT is closed to that of three typical CNN models, and self-attention-based ViT model.

4.4. Learning efficiency during training

Our proposed method LLCT and the classical methods (including VGG16, Xception [28], ViT, and AlexNet) were validated on the same training and testing datasets. The number of iterations was 80. The loss and accuracy were recorded every four iterations. The average training loss and testing accuracy of each batch are shown in Fig. 4.

We can see that our proposed model needs more iterations than the other models to converge in training. The reason may be that self-attention needs more computation to converge than CNN. At the same time, the proposed model achieves convergence faster than ViT. The possible reason is that convolution is used to extract the underlying features of B-Scan images, overlapping convolutions can enhance the semantic relationship between tokens, and treating a single pixel in the feature map as a single token are beneficial for global feature modeling. The proposed model has higher test accuracy than the ViT model. The possible reason is that in the Vit model, the original image is directly used for patch embedding without filtering, and then the Transformer is used for global feature modeling, and the classifier directly uses class token for classification, and its performance will be affected by the speckle noise of the image. As iterations increase, our proposed model has better testing accuracy, and the result tends to be more stable.

4.5. Performance results

The lesion localization in OCT images is shown in Fig. 3. The results demonstrate that the LLCT can provide localization of lesions in classification without lesion labeling in the training dataset. For our proposed method, the dataset is randomly divided into a training set and testing set with no intersection. This process is repeated six times for training and testing. Table 1 shows the classification of four cases on OCT images with different methods. The overall accuracy, overall sensitivity, and overall specificity with different models are shown in Table 3. The confusion table of the test dataset is shown in Fig. 5.

As can be observed in Table I, the accuracy, specificity and sensitivity of LLCT in the classification of CNV, DME, drusen and normal are better than other methods in general. The average accuracy, precision, sensitivity, and specificity of LLCT are improved by 3.8%, 11.7%, 10.9%, and 2.6% in four types (CNV, DME, drusen and normal), respectively. The LLCT based on self-attention has the most noticeable improvement in precision and sensitivity. One of

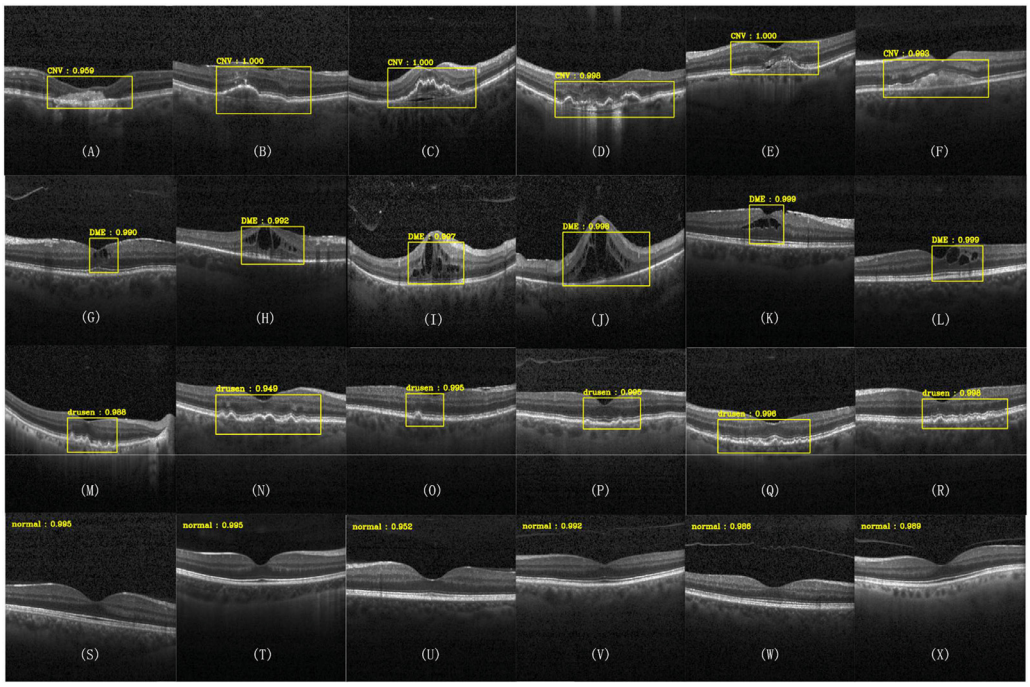


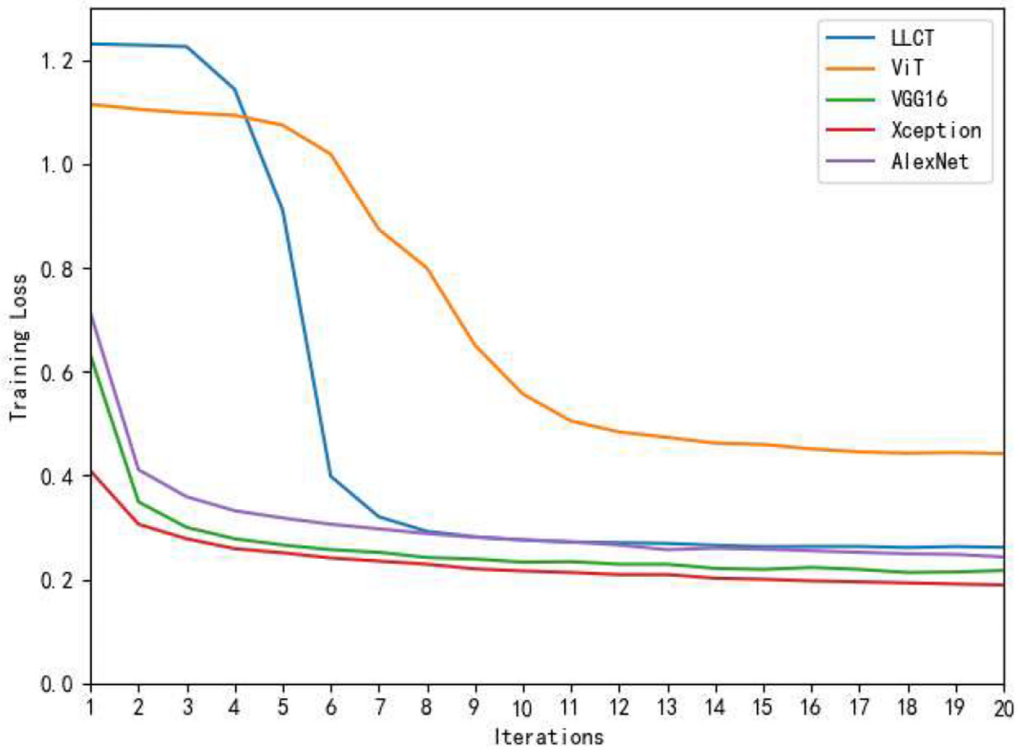
Fig. 3. Lesion localization of OCT image diagnosis. CNV is shown in pictures (A) to (F), DME in pictures (G) to (L), drusen in pictures (M) to (R), and normal in pictures (S) to (X).

Table 1
Classifier of Retina OCT image in four categories with different models. The best values in the table are bold.

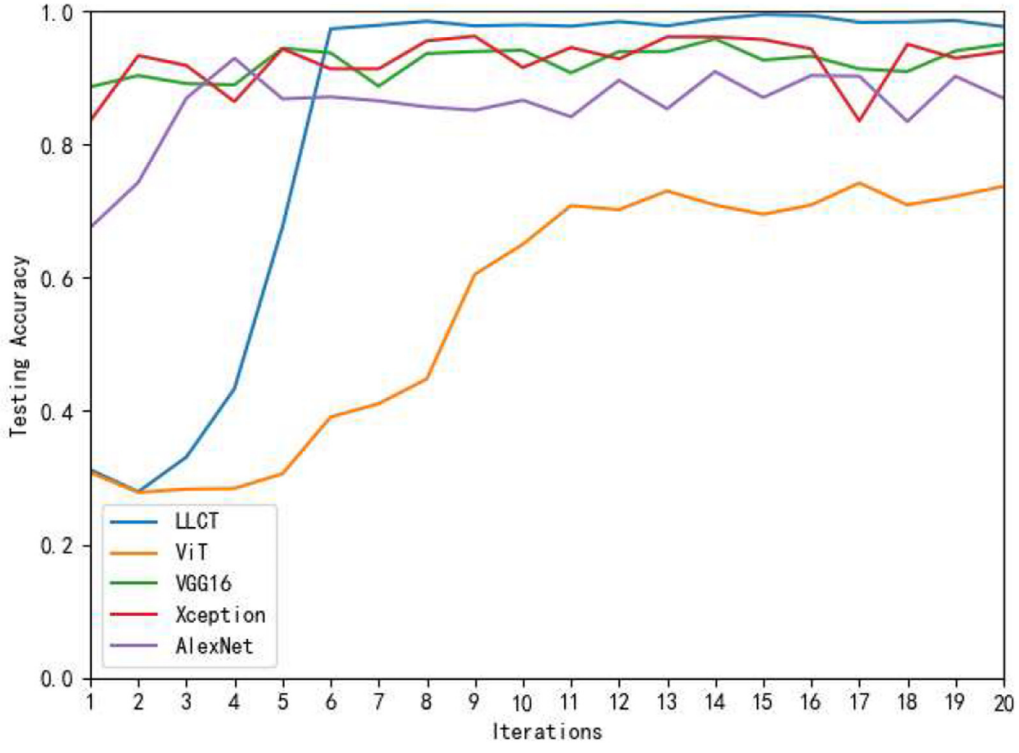
Method	Class	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	AUC (%)
HOG-SVM [7]	CNV	85.7 ± 0.2	82.0 ± 1.1	87.6 ± 2.1	84.0 ± 1.5	92.2 ± 0.2
	DME	91.4 ± 0.2	74.6 ± 0.8	53.8 ± 2.1	97.2 ± 0.2	87.3 ± 0.5
	Drusen	90.2 ± 0.4	52.6 ± 4.0	29.5 ± 3.5	97.0 ± 0.8	81.3 ± 0.9
	Normal	89.1 ± 0.8	78.1 ± 2.0	90.4 ± 1.2	88.4 ± 1.4	94.6 ± 0.4
Inception v3 [8]	CNV	86.9 ± 1.1	93.9 ± 0.7	76.2 ± 3.3	95.9 ± 0.7	96.1 ± 0.3
	DME	91.6 ± 0.7	66.9 ± 4.4	75.5 ± 3.2	94.1 ± 1.3	93.8 ± 0.5
	Drusen	87.2 ± 0.7	42.0 ± 1.7	70.7 ± 1.6	89.1 ± 0.9	89.2 ± 0.7
	Normal	93.3 ± 0.2	89.5 ± 1.0	88.9 ± 1.3	95.2 ± 0.6	98.1 ± 0.1
VGG16 [10]	CNV	91.0 ± 1.2	93.2 ± 1.5	86.6 ± 3.9	94.7 ± 1.5	97.2 ± 0.3
	DME	92.8 ± 0.5	74.6 ± 4.7	70.9 ± 4.7	96.2 ± 1.1	93.6 ± 0.6
	Drusen	90.7 ± 1.0	54.5 ± 5.3	54.7 ± 6.5	94.7 ± 1.7	88.7 ± 0.7
	Normal	91.8 ± 0.3	83.3 ± 2.6	92.6 ± 3.4	91.5 ± 2.0	97.2 ± 0.2
AlexNet [11]	CNV	90.4 ± 1.3	**94.0** ± 2.2	84.5 ± 4.3	95.3 ± 1.4	96.9 ± 0.6
	DME	92.2 ± 1.1	75.9 ± 4.7	69.8 ± 4.8	95.7 ± 1.7	93.0 ± 1.3
	Drusen	90.0 ± 1.9	51.5 ± 5.2	58.5 ± 7.3	93.9 ± 2.9	87.8 ± 2.7
	Normal	90.9 ± 1.3	86.2 ± 3.3	91.5 ± 7.8	90.7 ± 3.4	97.0 ± 0.4
LACNN [9]	CNV	92.7 ± 1.5	93.5 ± 1.3	89.8 ± 4.5	95.1 ± 1.6	97.7 ± 0.5
	DME	96.6 ± 0.2	86.4 ± 1.6	87.5 ± 1.5	98.0 ± 0.3	97.4 ± 0.4
	Drusen	93.6 ± 1.4	70.0 ± 5.7	72.5 ± 7.9	95.9 ± 2.1	**93.4** ± 1.5
	Normal	97.4 ± 0.2	94.8 ± 1.1	97.3 ± 1.0	97.4 ± 0.5	**99.2** ± 0.2
ViT [19]	CNV	87.8 ± 4.5	69.2 ± 9.3	94.6 ± 0.8	85.6 ± 6.3	95.7 ± 3.5
	DME	93.4 ± 1.1	88.5 ± 1.1	84.4 ± 6.2	96.4 ± 0.6	75.3 ± 4.1
	Drusen	88.7 ± 4.7	96.9 ± 1.7	56.8 ± 20.4	99.3 ± 0.6	58.2 ± 10.3
	Normal	94.8 ± 0.9	86.7 ± 0.4	93.4 ± 4.8	95.2 ± 0.4	94.2 ± 1.6
LLCT	CNV	**98.1** ± 1.9	93.5 ± 6.9	**99.4** ± 0.3	**97.6** ± 2.7	**99.6** ± 0.6
	DME	**99.6** ± 0.2	**98.6** ± 0.8	**99.6** ± 0.0	**99.5** ± 0.3	**99.7** ± 0.5
	Drusen	**98.1** ± 2.3	**99.6** ± 0.6	**92.8** ± 8.5	**99.9** ± 0.2	91.9 ± 9.7
	Normal	**99.6** ± 0.6	**99.6** ± 0.6	**98.8** ± 1.7	**99.9** ± 0.2	95.9 ± 0.3

Table 2
Average training time per iteration and inference time of the classical model and the proposed model. The best values in the table are bold.

Method	Average Training Time (second)	Average Inference Time (second)
Xception [28]	**971**	5.86
VGG16 [10]	1776	5.88
AlexNet [11]	1080	**5.72**
ViT [19]	3201	5.75
LLCT	1612	5.83



(a)



(b)

Fig. 4. Comparison of learning efficiency with other three models: (a) average training loss of each batch; (b) average testing accuracy of each batch.

Table 3
Classifier of Retina OCT image in overall accuracy, overall sensitivity, and overall specificity with different models. The best values in the table are bold.

Method	Overall Accuracy (%)	Overall Sensitivity (%)	Overall Specificity (%)
HOG-SVM [7]	78.1 ± 0.7	65.3 ± 0.9	71.8 ± 1.3
Inception v3 [8]	79.5 ± 1.0	77.9 ± 0.4	73.1 ± 1.1
VGG16 [10]	83.2 ± 1.2	76.2 ± 0.7	76.4 ± 1.5
AlexNet [11]	81.5 ± 2.8	75.2 ± 2.7	76.9 ± 1.4
LACNN [9]	90.1 ± 1.4	86.8 ± 1.3	86.2 ± 2.3
ViT [19]	74.1 ± 3.4	82.3 ± 0.1	90.0 ± 3.4
LLCT	**97.7** ± 2.5	**97.7** ± 1.1	**99.2** ± 1.1

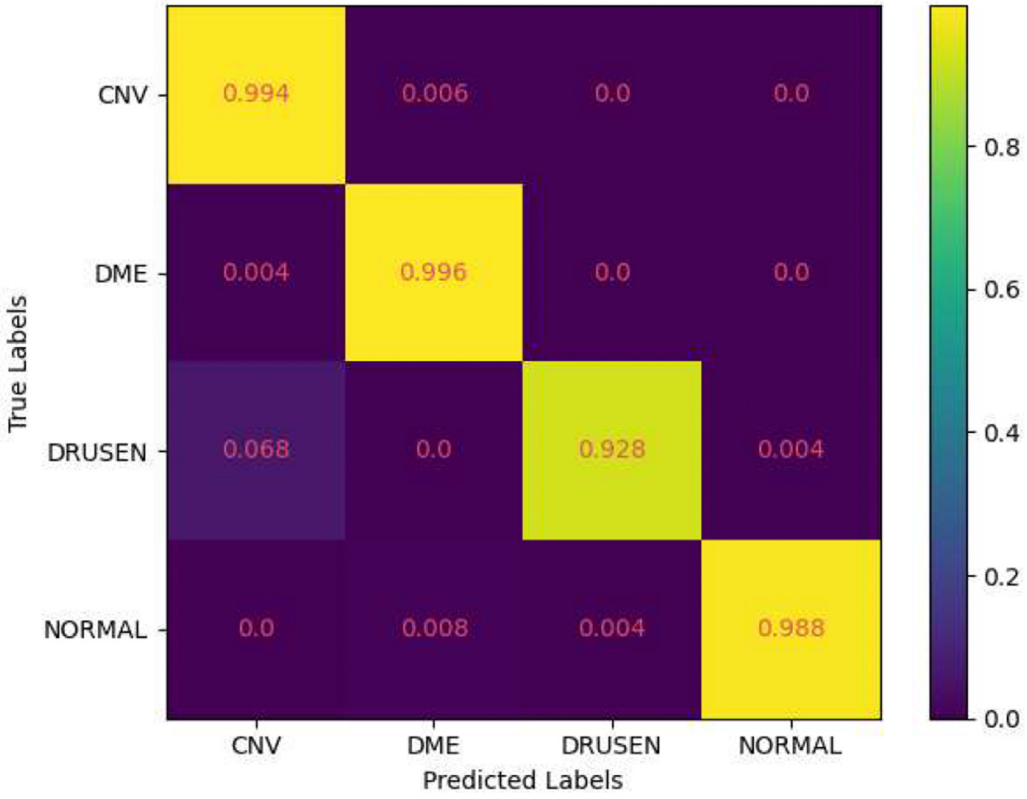


Fig. 5. Confusion table of model regarding classification result.

the reasons might be that LLCT based on self-attention can classify the image by considering the global features, while CNN only considers local features.

The results in Table III show that LLCT achieves overall accuracy of 97.7%, overall sensitivity of 97.7%, and overall specificity of 99.2%, which are better than other methods. Compared to other methods, LLCT achieves improvement of about 7.6% in overall accuracy, 10.9% in overall sensitivity, and 9.2% in overall specificity.

From Fig 8, we can find that LLCT has the highest accuracy in classifying CNV, DME and normal, followed by drusen. The reason may be that the characteristics of drusen are less obvious than CNV, DME, and normal. LLCT does not wrongly classify the CNV image as drusen, but a few wrong classifications of the drusen image as CNV. The reason may be that the feature of drusen is similar to CNV.

The limitation of the proposed model is that although the LLCT is less expensive in the training process than Transformer-based, it still requires more iterations to converge compared to the typical CNN approaches. Therefore, the computational cost in the training process is generally higher than that of typical CNNs.

5. Conclusions

In this paper, we proposed a novel lesion-localization convolution transformer for OCT image analysis, named LLCT, to improve accuracy, precision, sensitivity and specificity of ophthalmology disease classification. In LLCT, CNN is applied to generate feature maps and convert them to tokens as the input sequence for transformer, which reduces the computation complexity and enhances the semantic relationship among tokens. Transformer is applied to take global features of retinal OCT images into considerations for classification. Moreover, we have also provided a lesion location solution based on classification results and gradients without the labeled lesion location data. Extensive experiments have demonstrated that our proposed LLCT achieves improvement of about 7.6% in overall accuracy, 10.9% in overall sensitivity, and 9.2% in overall specificity compared with other methods. Our system has great potential to be applied in assist ophthalmologists for ophthalmology disease diagnosis and pay attention to lesion location. In the future, we will continue to explore the model for retinal layer segmentation on B-scan images, and optimize the model to reduce inference time.

Declaration of Competing Interest

The authors declared no conflicts of interest.

References

[1] T. Hassan, M.U. Akram, B. Hassan, A. Nasim, S.A. Bazaz, Review of OCT and fundus images for detection of Macular Edema, 2015 IEEE International Conference on Imaging Systems and Techniques (IST), 2015.

[2] C.S. Lee, et al., Associations between recent and established ophthalmic conditions and risk of Alzheimer's disease, Alzheimers Dement. 15 (1) (Jan 2019) 34–41.

[3] K.B. Freund, L.A. Yannuzzi, J.A. Sorenson, Age-related macular degeneration and choroidal neovascularization, Am. J. Ophthalmol. 115 (6) (1993) 786–791.

[4] F.E. Hirai, M. Knudtson, B. Klein, R. Klein, Clinically Significant macular edema and survival in type 1 and type 2 diabetes - ScienceDirect, Am. J. Ophthalmol. 145 (4) (2008) 700–706.

[5] P. Zarbin, Drusen in age-related macular degeneration: pathogenesis, natural course, and laser photocoagulation–induced regression, Surv. Ophthalmol. (1999).

[6] J.M. Schmitt, S.H. Xiang, K.M. Yung, Speckle in optical coherence tomography, J. Biomed. Opt. 4 (1) (1999) 95–105.

[7] S. Naz, T. Hassan, M.U. Akram, S.A. Khan, A practical approach to OCT based classification of Diabetic Macular Edema, International Conference on Signals & Systems, 2017.

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

[9] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, Z. Liu, Attention to lesion: lesion-aware convolutional neural network for retinal optical coherence tomography image classification, IEEE Trans. Med. Imaging (2019) 1-1.

[10] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv (2014).

[11] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.

[12] A. Thomas, A.P. Sunija, R. Manoj, R. Ramachandran, P. Palanisamy, RPE layer detection and baseline estimation using statistical methods and randomization for classification of AMD from retinal OCT, Comput. Methods Programs Biomed. (6) (2020) 105822.

[13] P.P. Srinivasan, et al., Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images, Biomed Opt Express 5 (10) (2014) 3568–3577.

[14] C.S. Lee, D.M. Baughman, A.Y. Lee, Deep learning is effective for classifying normal versus age-related macular degeneration OCT images, Ophthalmol. Retina vol. 1 (4) (2017) 322–327.

[15] D.S. Kermany, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.e9.

[16] R. Rasti, H. Rabbani, A. Mehridehnavi, F. Hajizadeh, Macular OCT classification using a multi-scale convolutional neural network ensemble, IEEE Trans. Med. Imaging 37 (4) (2018) 1024–1034.

[17] P. Seebock, et al., Unsupervised identification of disease marker candidates in retinal OCT imaging data, IEEE Trans. Med. Imaging (2018) 1-1.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, N. Houlsby, An image is worth 16 $\times$ 16 words: transformers for image recognition at scale, arXiv (2020).

[20] A. Vaswani, et al., Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[21] J. Chen, Y. Lu, Q. Yu, X. Luo, Y. Zhou, TransUNet: transformers make strong encoders for medical image segmentation, arXiv (2021).

[22] B. Wu, et al., Visual transformers: token-based image representation and processing for computer vision, arXiv (2020).

[23] H. Wu, et al., CvT: introducing convolutions to vision transformers, arXiv (2021).

[24] Y. Zhang, H. Liu, Q. Hu, TransFuse: fusing transformers and CNNs for medical image segmentation, arXiv (2021).

[25] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (2007) 62–66.

[26] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv (2014).

[27] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256. JMLR Workshop and Conference Proceedings.

[28] F.J.I. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, 2017.