Journal of Computers Vol. 35 No. 1, February 2024, pp. 255-266
doi: 10.53106/199115992024023501021

Retinal OCT Image Classification Based on CNN-RNN Unified Neural Networks

Xue-Feng Jiang[1], Ken-Cheng[2], Zhi-De Li[2*]

[1] Department of Artificial and Intelligence, Shenzhen Polytechnic,
Nanshan District, Shenzhen City, Guangdong Province
jiangxuefeng@szpu.edu.cn

[2] Shenzhen International Graduate School Tsinghua University,
Tsinghua Campus, Xili University Town, Nanshan District, Shenzhen
{chengken, lizhide}@sz.tsinghua.edu.cn

Received 15 December 2023; Revised 15 January 2024; Accepted 31 January 2024

Abstract. Computer-aided diagnosis of retinopathy is a hot research topic in the field of medical image classification, where optical coherence tomography (OCT) is an important basis for the diagnosis of ophthalmic diseases. Traditional approaches to multi-label image classification learn independent classifiers for each category and employ ranking or thresholding on the classification results. These techniques, although working well, fail to explicitly exploit the label dependencies in an image. In this paper, two publicly available retinal OCT image datasets are integrated and screened. Then, an end-to-end deep learning algorithmic framework based on CNN-RNN Unified Neural Networks was proposed to automatically and reliably classify six categories of retinal OCT images. Numerical results suggest that the proposed algorithm works well in terms of accuracy, precision, sensitivity and specificity, approaching or even partially surpassing the performance of clinical experts. It is valuable in promoting computer-aided diagnosis towards practical clinical applications and improving the efficiency of clinical diagnosis of retinal diseases.

Keywords: optical coherence tomography (OCT), image classification, CNN-RNN

1 Introduction

The macula is an important area of the retina that is associated with visual functions such as color vision and fine vision. Once a lesion occurs in the macula, vision will be negatively affected [1]. Morphological changes in the macula can often help in the early diagnosis and treatment of certain diseases. Retinal imaging techniques can help physicians understand the pathogenesis of diseases such as age-related macular degeneration (AMD), diabetic retinopathy (DR), and macular fissures. Early monitoring these diseases can further prevent more severe vision loss as well as play an important role in exploration of new therapies.

OCT was first used for eye imaging in 1991 [2] and is now gaining rapid and widespread adoption in ophthalmology. Its principle is based on the basis of weak coherent light interference, to detect the incoming weak coherent light or the back reflection of multiple scattered signals in different levels of biological tissue, to scan and obtain the two-dimensional or three-dimensional structure image of biological tissue. Fig. 1(c) is an example of an OCT image. OCT can be used for cross-sectional visualization of retinal structures and is critical for early diagnosis of several pathologies affecting the retina and optic nerve head. Since the eye can capture both visible and near-infrared light, OCT can combine unprecedented high resolution (1-10 μm) with appropriate tissue penetration depth (1-2 mm). Studies have shown that OCT can also be used to monitor tissue function and intraretinal blood flow as well as to assess blood oxygenation levels [3].

With the development of machine learning, more and more people begin to study the application of machine learning algorithms and various neural networks in medical image analysis. In 2017, Karri [4] used the public OCT dataset proposed by Srinivasan et al. Classify and recognize OCT images. They fine-tuned GoogLeNet, a pre-trained convolutional neural network, to train and classify OCT images so that the neural network could be trained and get good results by using limited data. In 2019, Feng [5] used transfer learning to classify OCT images to reduce reliance on data set size. They fine-tuned the pre-trained VGG16 [6]. In 2019, Juan [7] conducted

* Corresponding Author

a comprehensive comparative analysis on the recognition and diagnosis performance of VGG19, GoogLeNet, ResNet50 and DeNet neural networks for glaucoma disease through color fundus images. Cheng [8] proposed a deep hash algorithm based on ResNet 50 to perform image retrieval and classification tasks. Although better results have been obtained in diagnosing OCT images of retinopathy through traditional machine learning methods [9, 10], it is still difficult to obtain higher accuracy. The study of Rasti [11] improved the network structure and combined local and global information with multi-scale methods to extract and retain as many effective features as possible, thus achieving good results. Karri's study [4] proposed transfer learning to reduce the training parameters of neural networks, thereby reducing the dependence on the size of data sets, but this method has poor adaptability to the differences between different data sets.

Lecun [12] used convolutional neural network (CNN) structures in practical tasks for the first time such as handwritten digit recognition, and etc. by applying the idea of local perceptual space downsampling to achieve feature extraction and dimensionality reduction. Yuan Liu [13] propose a deep ensemble network with attention mechanism that detects glaucoma using optic nerve head stereo images. And the approach increases recall (sensitivity) from the state-of-the-art 88.89% to 95.48%, while maintaining precision and performance stability. Szegedy [14] proposed version 3 (V3) of a CNN structure referred to as Inception to improve network classification performance by increasing the width and depth of the network in a parallel manner. Gong [15] adopted a domain adaptation CNN structure to improve the performance of the model with small data sets and prevent overfitting. They used pre-trained VGG16 for fixed feature extraction, and then fine-tuned the domain adaptation CNN with medical images to realize transfer learning. This network was applied to the graded assessment of knee osteoarthritis and achieved higher accuracy than train-from-scratch.

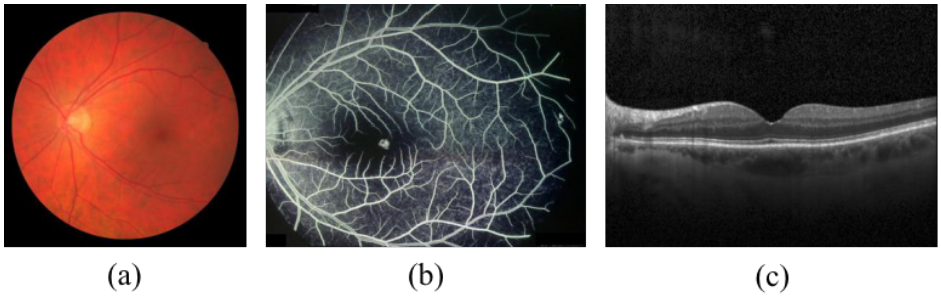
While deep convolutional neural networks (CNNs) have shown a great success in single-label image classification, it is important to note that real world images generally contain multiple labels, which could correspond to different objects, scenes, actions and attributes in an image. Traditional approaches to multi-label image classification learn independent classifiers for each category and employ ranking or thresholding on the classification results. These techniques, although working well, fail to explicitly exploit the label dependencies in an image.

Every real-world image can be annotated with multiple labels, because an image normally abounds with rich semantic information, such as objects, parts, scenes, actions, and their interactions or attributes. Modeling the rich semantic information and their dependencies is essential for image understanding. As a result, multi-label classification task is receiving increasing attention [16, 17]. Inspired by the great success from deep convolutional neural networks in single-label image classification in the past few years, which demonstrates the effectiveness of end-to-end frameworks, we explore to learn a unified framework for multi-label image classification.

A common approach that extends CNNs to multi-label classification is to transform it into multiple single-label classification problems, which can be trained with the ranking loss or the cross-entropy loss. However, when treating labels independently, these methods fail to model he dependency between multiple labels. Previous works have shown that multi-label classification problems exhibit strong label co-occurrence dependencies. For instance, sky and cloud usually appear together, while water and cars almost never co-occur.

To model label dependency, most existing works are based on graphical models, among which a common approach is to model the co-occurrence dependencies with pairwise compatibility probabilities or co-occurrence probabilities and use Markov random fields to infer the final joint label probability. However, when dealing with a large set of labels, the parameters of these pairwise probabilities can be prohibitively large while lots of the parameters are redundant if the labels have highly overlapping meanings. Moreover, most of these methods either cannot model higher-order correlations, or sacrifice computational complexity to model more complicated label relationships. In this paper, we explicitly model the label dependencies with recurrent neural networks (RNNs) to capture higher-order label relationships while keeping the computational complexity tractable. We find that RNN significantly improves classification accuracy.

For the CNN part, to avoid problems like overfitting, previous methods normally assume all classifiers share the same image features. However, when using the same image features to predict multiple labels, objects that are small in the images are easily get ignored or hard to recognize independently. In this work, we design the RNNs framework to adapt the image features based on the previous prediction results, by encoding the attention models implicitly in the CNN-RNN structure. The idea behind it is to implicitly adapt the attentional area in images so the CNNs can focus its attention on different regions of the images when predicting different labels. For example, when predicting multiple labels for images in Fig. 1, our model will shift its attention to smaller ones (i.e. Runway, Person, Hat) after recognizing the dominant object (i.e. Airplane, Great Pyrenees, Archery). These small objects are hard to recognize by itself, but can be easily inferred given enough contexts.



(a) Examples of fundus camera (b) Fluorescein angiography (c) OCT images

Fig. 1. Morphology of fundus/retina under **different** imaging techniques

Finally, many image labels have overlapping meanings. For example, cat and kitten have almost the same meanings and are often interchangeable. Not only does exploiting the semantic redundancies reduce the computational cost, it also improves the generalization ability because the labels with duplicate semantics can get more training data.

The label semantic redundancy can be exploited by joint image/label embedding, which can be learned via canonical correlation analysis, metric learning, or learning to rank methods. The joint image/label embedding maps each label or image to an embedding vector in a joint low-dimensional Euclidean space such that the embedding of semantically similar labels are close to each other, and the embedding of each image should be close to that of its associated labels in the same space. The joint embedding model can exploit label semantic redundancy because it essentially shares classification parameters for semantically similar labels. However, the label co-occurrence dependency is largely ignored in most of these models.

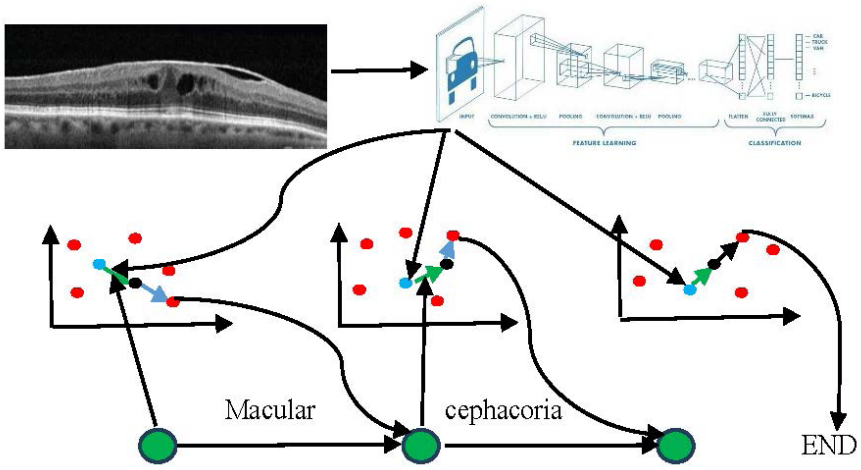


Fig. 2. An illustration of the CNN-RNN for retinal OCT image classification

In this paper, we propose a unified CNN-RNN framework for multi-label image classification, which effectively learns both the semantic redundancy and the co-occurrence dependency in an end-to-end way. The framework of the proposed model is shown in Fig. 2. The multi-label RNN model learns a joint low-dimensional image-label embedding to model the semantic relevance between images and labels. The image embedding vectors are generated by a deep CNN while each label has its own label embedding vector. The high-order label co-occurrence dependency in this low-dimensional space is modeled with the long short term memory recurrent neurons, which maintains the information of label context in their internal memory states. The RNN framework

computes the probability of a multi-label prediction sequentially as an ordered prediction path, where the a priori probability of a label at each time step can be computed based on the image embedding and the out-put of the recurrent neurons. During prediction, the multi-label prediction with the highest probability can be approximately found with beam search algorithm. The proposed CNN-RNN framework is a unified framework which combines the advantages of the joint image/label embedding and label co-occurrence models, and it can be trained in an end-to-end OCT image classification algorithm for retinopathy, which effectively improves the classification accuracy of retinal OCT images with six categories and performs well on small datasets.

2 Method

RNN with LSTM can effectively model the long-term temporal dependency in a sequence. It has been successfully applied in image captioning, machine translation, speech recognition], language modeling, and word embedding learning. Since we aim to characterize the high-order label correlation, we employ long short term memory (LSTM) neurons as our recurrent neurons, the LSTM is shown in Fig. 3.

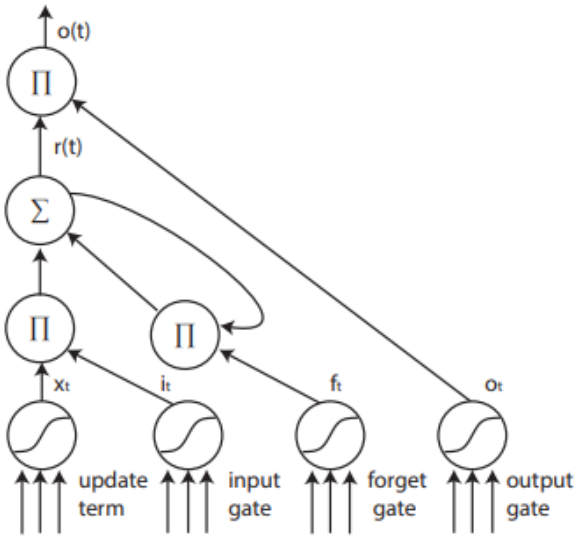


Fig. 3. A illustration of a LSTM neuron

2.1 Long Short Term Memory Networks (LSTM)

RNN is a class of neural network that maintains internal hidden states to model the dynamic temporal behaviour of sequences with arbitrary lengths through directed cyclic connections between its units. It can be considered as a hidden Markov model extension that employs nonlinear transition function and is capable of modeling long term temporal dependencies. LSTM extends RNN by adding three gates to an RNN neuron: a forget gate f to control whether to forget the current state; an input gate i to indicate if it should read the input; an output gate o to control whether to output the state. These gates enable LSTM to learn long-term dependency in a sequence, and make it is easier to optimize, because these gates help the input signal to effectively propagate through the recurrent hidden states r(t) without affecting the output. LSTM also effectively deals with the gradient vanishing/exploding issues that commonly appear during RNN training [18].

$$x_t = \delta(U_r \cdot r(t-1) + U_w w_k(t)). \quad (1)$$

$$i_t = \delta(U_{i_r} r(t-1) + U_{i_w} w_k(t)). \tag{2}$$

$$f_t = \delta(U_{f_r} r(t-1) + U_{f_w} w_k(t)). \tag{3}$$

$$o_t = \delta(U_{o_r} r(t-1) + U_{o_w} w_k(t)). \tag{4}$$

$$r(t) = f_t \odot r(t-1) + i_t \odot x_t. \tag{5}$$

$$o(t) = r(t) \odot o(t). \tag{6}$$

where δ(.) is an activation function, $\odot$ is the product with gate value, and various W matrices are learned parameters. In our implementation, we employ rectified linear units (ReLU) as the activation function.

2.2 Model

We propose a novel CNN-RNN framework for multi-label classification problem. The illustration of the CNN-RNN framework is shown in Fig. 4. It contains two parts: The CNN part extracts semantic representations from images; the RNN part models image/label relationship and label dependency [19].

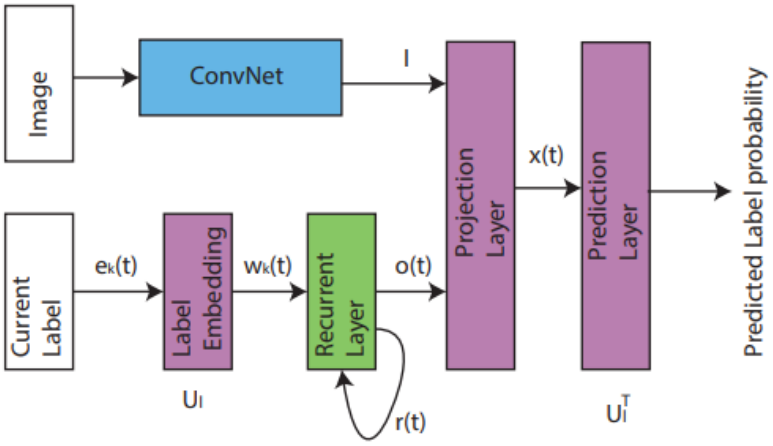


Fig. 4. The architecture of the proposed RNN model for multi-label classification

We decompose a multi-label prediction as an ordered prediction path. For example, labels “zebra” and “elephant” can be decomposed as either (“zebra”, “elephant”) or (“ele- phant”, “zebra”). The probability of a prediction path can be computed by the RNN network. The image, label, and recurrent representations are projected to the same low- dimensional space to model the image-text relationship as well as the label redundancy. The RNN model is employed as a compact yet powerful representation of the label co-occurrence dependency in this space. It takes the embedding of the predicted label at each time step and maintains a hidden state to model the label co-occurrence information. The a priori probability of a label given the previously predicted labels can be computed according to their dot products with the sum of the image and recurrent embeddings. The probability of a prediction path can be obtained as the product of the a-prior probability of each label given the previous labels in the prediction path.

A label k is represented as a one-hot vector e_k = [0, … , 0, 1, 0, …, 0], which is 1 at the k-th location, and 0 elsewhere. The label embedding can be obtained by multiplying the one-hot vector with a label embedding matrix U_I. The k-th row of U_I is the label embedding of the label k.

$$w_k = U_I \cdot e_k. \tag{7}$$

The dimension of w_k is usually much smaller than the number of labels.

The recurrent layer takes the label embedding of the previously predicted label, and models the co-occurrence dependencies in its hidden recurrent states by learning non-linear functions:

$$o(t) = h_o(r(t-1), w_k(t), r(t) = h_r(r(t-1), w_k(t)), \tag{8}$$

where $r(t)$ and $o(t)$ are the hidden states and outputs of the recurrent layer at the time step t respectively, $w_k(t)$ is the label embedding of the t-th label in the prediction path, and $h_o(.)$, $h_r(.)$ are the non-linear RNN functions, which will be described in details in Sec. 3.1.

The output of the recurrent layer and the image representation are projected into the same low-dimensional space as the label embedding.

$$x_t = h(U_o^x o(t) + U_I^x I. \tag{9}$$

Where U_o^x and U_I^x are the projection matrices for recurrent layer output and image representation, respectively.

The convolutional neural network is employed as the image representation, and the recurrent layer captures the information of the previously predicted labels. The output label probability is computed according to the image representation and the output of the recurrent layer.

Number of columns of U_o^x and U_I^x are the same as the label embedding matrix U_I , I is the convolutional neural network image representation.

The label scores can be computed by multiplying the transpose of U_I and x_t to compute the distances between x_t and each label embedding.

$$s(t) = U_I^T x_t. \tag{10}$$

The predicted label probability can be computed using softmax normalization on the scores.

2.3 Inference

A prediction path is a sequence of labels (l_1, l_2, l_3, ..., l_N), where the probability of each label can be computed with the information of the image I and the previously predicted labels (l_1, ..., l_{t-1}). The RNN model predicts multiple labels by finding the prediction path that maximizes the a priori probability.

$$\begin{aligned} l_1,...,l_k &= \arg\max_{l_1,...,l_k} P(l_1,...,l_k \,|\, I) \\ &= \arg\max_{l_1,...,l_k} P(l_1 \,|\, I) \times P(l_2 \,|\, I, l_1) \\ &...P(l_k \,|\, I, l_1, ..., l_{k-1}). \end{aligned} \tag{11}$$

Since the probability $P(l_k|I, l_1, ..., l_{k-1})$ does not have Markov property, there is no optimal polynomial algorithm to find the optimal prediction path. We can employ the greedy approximation, which predicts label $l_t = \mathrm{argmax}_{l_t} P(l_t|I, l_1, ..., l_{t-1})$ at time step t and fix the label prediction l_t at later predictions. However, the greedy algorithm is problematic because if the first predicted label is wrong, it is very likely that the whole sequence cannot be correctly predicted. Thus, we employ the beam search algorithm to find the top-ranked prediction path.

An example of the beam search algorithm can be found in Fig. 5. Instead of greedily predicting the most probable label, the beam search algorithm finds the top-N most probable prediction paths as intermediate paths $S(t)$ at each time step t.

$$S(t) = \{P_1(t), P_2(t), ..., P_N(t)\}. \quad (12)$$

At time step t + 1, we add N most probable labels to each intermediate path Pi(t) to get a total of N × N paths. The N prediction paths with highest probability among these paths constitute the intermediate paths for time step t + 1. The prediction paths ending with the END sign are added to the candidate path set C. The termination condition of the beam search is that the probability of the current intermediate paths is smaller than that of all the candidate paths. It indicates that we cannot find any more candidate paths with greater probability.

3 Implementation Program

3.1 Datasets

Optical coherence tomography image data for this study was selected from the University of California, San Diego. San Diego) Shiley Eye Institute, California Retinal Research Foundation, Shanghai First People's Hospital Eye Medicine Center and Beijing Tongren Eye Center Adult patient cohort between July 1, 2013 and March 1, 2017. There were a total of 83484 OCT images of choroidal neovascularization, diabetic macular edema, choroidal hyaline wart and healthy fundus in the dataset, 37205, 8616, 11348 and 26315 images for each type, respectively. Test sets of 250 sheets for each category.

Data set II is from the Theoretical and Experimental Cognitive Theory Laboratory at the University of Waterloo. The images for the database were obtained from a raster scanning protocol with a 2 mm scan length and a resolution of 512 × 1024 pixels, taken using a Cirrus HD-OCT machine (supplied by Carl Zeiss) at Sankara Nethralaya Eye Hospital, Chennai, India. During each volume scan, an image centered on the foveal area is selected by an experienced clinical optometrist. The axial and transverse resolution of the tissue were 5 μm and 15 μm, respectively. The light source is a superlight-emitting diode with a wavelength of 840 nm. Then resize the image to 500 x 750 pixels. The cause of the disease is diagnosed by the clinician, and the image classification label is determined according to the diagnosis of the retinal clinical experts of SN Hospital. The dataset contains images of different stages of different diseases, including mild, moderate and severe stages. The severity of the disease is random in the data set, and the images for each disease are distributed in no order.

In terms of dataset screening, I excluded images with poor quality, and in order to ensure the balance of the dataset and reduce the network's bias against certain anomalies, I only selected 150 OCT images for each type of fundus morphology in dataset 1, and all 100 images of MH and SMD in dataset 2 were included in my own database. In the division of data set, 15% images were randomly selected from the whole data set as the verification set, which did not participate in the training process, to ensure the reliability of the experimental results. Each image was labeled according to the location of the images. The labels were in the form of unique thermal coding, and five abnormal and healthy fundus OCT images were labeled in lexicographical order.

3.2 Data Preprocessing and Data Augmentation

OCT images are subject to random noise due to differences in the physical properties of the device they are taking, which limits the accuracy of any quantitative measurements in the data. The purpose of non-local mean filtering is to reduce speckle noise in OCT amplitude images, improve image clarity, and lay a foundation for later feature extraction and image recognition tasks. Non-local mean filtering algorithms use redundant information that is common in natural images to remove noise. Different from the commonly used bilinear filter and median filter, which use the local information of the image to filter, it uses the whole image to denoise, finds the similar regions in the image block as the unit, and then averages these regions, which can better remove the Gaussian noise in the image. In addition, for the problem that the blurred edges of OCT images after denoising will affect the feature extraction, Laplace filtering is used for image enhancement to ensure the integrity of relevant information. Due to the large amount of raw fundus OCT image data and the difficulty of manual annotation, the existing publicly available single dataset is often small in terms of disease types and the number of images. In this paper, we first screen, integrate and pre-process the two datasets, and then augment the datasets to make the model better fit the real, complex clinical environment and reduce the overfitting phenomenon of the network. Taking the data set used in this work as an example, the specific processing steps are as follows.

(a) For Each image v = {v(i) | i ∈ I} containing discrete noise, the estimate NL[v](i) after non-local mean filtering for each pixel i is calculated from the weighted average of all pixels in the image, i.e.,

$$NL[v](i) = \sum_{j \in I} w(i, j)v(j), \tag{13}$$

Where weights $\{w(i, j)\}_j$ depend on the similarity between pixel i and pixel j, measured by the weighted Euclidean distance $\|v(N_i) - v(N_j)\|_{2,\alpha}^2$. The specific calculation formula is $w(i, j)\frac{1}{Z(i)}\exp = \left(\frac{\|v(N_i) - v(N_j)\|_{2,\alpha}^2}{h^2}\right)$, where $Z(i) = \sum_j \exp\left(-\frac{\|v(N_i) - v(N_j)\|_{2,\alpha}^2}{h^2}\right)$ is the normalization The non-local mean filtering is highly dependent on the parameter settings. We balance the computation time and processing effect, and finally choose to find a similar patch of 7×7 size in a search window of 21×21 pixels.

(b) The denoised image is Laplacian sharpened by the filtering template shown in Fig. 5. The results of pre-processing are shown in Fig. 6.

0	-1	0
-1	5	-1
0	-1	0

Fig. 5. Laplacian filter template

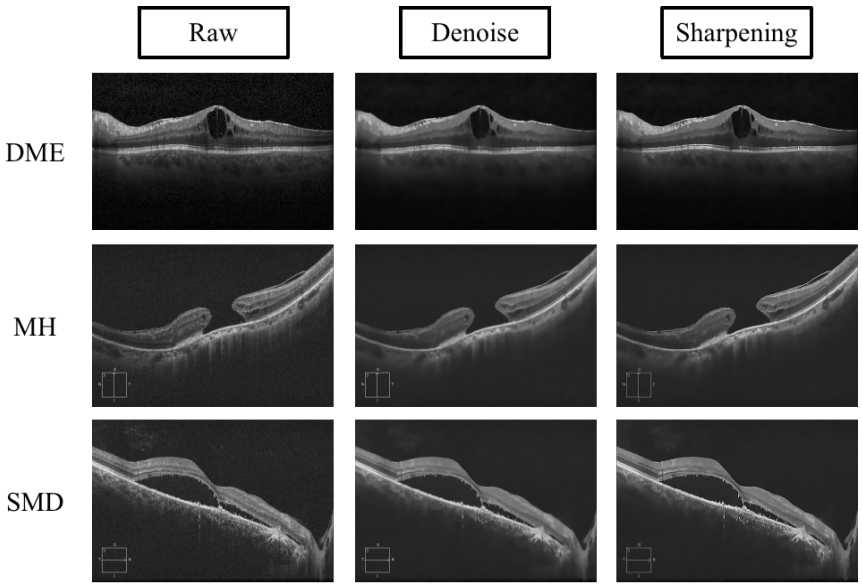


Fig. 6. Example of pre-processing results

(c) In order to deal with intensity variations in OCT images from different patients, the images are normalized so that they have equal variance. For image data, we divide the pixel value by 255, which is a fast way to approximate normalized data.

(d) For data augmentation, Keras uses the function ImageDataGenrator(), whose parameters can be configured to enable a series of random transformations of the OCT image, including rotation, cropping, scaling, and horizontal flipping. Since the RGB images of the dataset have different resolutions, we resize them all to 299×299×3. Finally, 8100, 1200 and 946 pieces of data are obtained for the training set, validation set and test set, respectively. For ease of explanation, For the sake of illustration, the input data after processing in the way above is uniformly denoted as A_in hereafter.

3.3 Training and Prediction

We aim to classify retinal OCT images with CNN-RNN. The classification network uses Inception V3, which has the following advantages: (a) it parallelly increases the width and depth of convolution module to allow it to extract more representational features and obtain higher accuracy; (b) it supports a small amount of data to train the model with pre-trained weights on ImageNet provided, which helps to speed up convergence and shorten training time during the fine-tuning process; (c) it is fast to classify images with the trained model.

The dimensions of the label embedding and of LSTM RNN layer are 64 and 512, respectively. We employ weight decay rate 0.0001, momentum rate 0.9, and dropout [4] rate 0.5 for all the projection layers.

We evaluate the proposed method on datasets and the evaluation demonstrates that the proposed method achieves superior performance to state-of-the-art methods. We also qualitatively show that the proposed method learns a joint la- bel/image embedding and it focuses its attention in different image regions during the sequential prediction.

4 Performance Analysis

4.1 Evaluation Indicators

The classification model is evaluated using the classification accuracy (Accuracy, Acc), sensitivity (Sensitivity, SE), and specificity (Specificity, SP), and each metric was calculated as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{14}$$

$$SE = \frac{TP}{TP + FN}, \tag{15}$$

$$SP = \frac{TN}{FP + TN}. \tag{16}$$

Where TP denotes true positive, TN denotes true negative, FP denotes false positive, and FN denotes false negative.

The F1-score is chosen as the model metric, which is a weighted reconciling average of precision and recall taking into account both precision and recall of the classification model.

$$F_1 - score = \frac{2PR}{P + R}. \tag{17}$$

Where

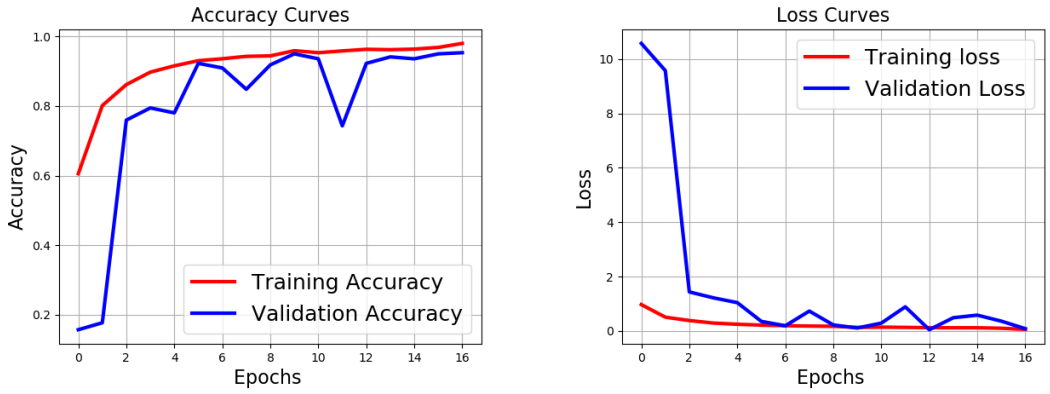
$$P = \frac{TP}{TP + FP}, \tag{18}$$

$$R = \frac{TP}{TP + FN}. \tag{19}$$

4.2 Performance Analysis

In the training process, the accuracy and loss of the network on the training set and verification set change with the increase of the number of iterations, as shown in Fig. 7. Where, (a) describes the curve of the accuracy rate changing with the increase of the number of iterations, and (b) shows the curve of the change of data set loss with the increase of the number of iterations. It can be seen that in the initial stage of training, with the increase of the number of iterations, the accuracy rate is in the rising stage and gradually approaches 1, and the loss gradually decreases and approaches 0. The weight of the network will be effectively adjusted, thus making the final classification result more accurate. After 12 iterations, the verification set loss basically no longer decreases, indicating that the network weight begins to overfit. After 16 iterations, the verification set loss does not decrease for four consecutive rounds compared to the 12th iteration, the network terminates the training process, and the weights are no longer updated.

It can be seen from Table 1 that the data are basically concentrated in the diagonal direction, indicating that our convolutional neural network model has a good classification performance.



(a) Accuracy curve (b) Loss curve

Fig. 7. Index changes and final classification results in the training process

Table 1. Confusion matrix of six types of data test results

Category	CNV	SMD	DME	Drusen	MH	Normal	Total
CNV	129	0	3	20	0	0	152
SMD	1	146	0	0	1	1	149
DME	3	0	159	3	0	3	168
Drusen	2	0	0	163	0	0	165
MH	0	1	2	0	150	0	153
Normal	0	0	0	7	0	152	159
Total	135	147	164	193	151	156	946

In order to highlight the superiority of this CNN-RNN deep learning classification method, our other research groups have compared the performance of traditional classification methods and other CNN structures (VGG16 and Inception V3) on the dataset in this paper. The results are shown in Table 2. As can be seen from the table, our deep learning method is superior to traditional classification methods in all evaluation indexes, and the CNN-RNN structure proposed by us is also superior to other CNNS. In addition, there are many kinds of data sets in this paper, and different anomalies interfere with each other, making classification difficult, which further indicates the advantages of this CNN-RNN structure algorithm.

Table 2. Comparison of classification results with traditional methods and other CNN models

Metric	Venhuizen	Lemaitre	VGG16	InceptionV3	Proposed
Acc	70.0	81.2	88.6	81.9	95.6
SE	71.4	81.2	89.5	82.1	95.0
SP	68.8	93.7	97.9	96.5	99.0
F1	70.5	78.5	89.6	81.4	95.1

Accuracy is the most intuitive evaluation index of computer-aided diagnosis in the field of medical image classification. Table 3 lists the comparison between the diagnostic accuracy rate of the model in this chapter and that of human doctors. The data of human doctors I and II came from UCSD, and the data of human doctors III and IV came from the Eye Center of Wuhan University People's Hospital. It can be seen that our automatic classification algorithm can approach the performance of human doctors in accuracy, and even exceed the performance of human experts in Drusen and MH. Data for human physicians I and II also came from UCSD. As can be seen from the table, the overall accuracy rate and sensitivity of our method are between the performance of two human doctors in two evaluation indicators, while the specificity exceeds the two doctors, indicating that the algorithm has stability and reliability comparable to that of human experts.

Table 3. Comparison of Acc with human physicians

Category	Proposed	Expert I	Expert II	Expert III	Expert IV
CNV	96.9	97.2	96.5	-	-
DME	96.6	97.3	97.6	-	-
Drusen	99.6	95.4	94.3	-	-
SMD	98.5	-	-	98.8	91.5
MH	99.6	-	-	86.7	88.9
Normal	98.8	98.9	98.7	94.0	95.0

5 Conclusion

We propose a unified CNN-RNN framework for multi-label retinal OCT image classification. The framework combines the advantages of joint image/label embedding and label co-occurrence models, and uses CNN and RNN to classify label co-occurrence dependencies in joint image/label embedding space. The method achieved accuracy and reliability comparable to or better than clinical experts in identifying the five abnormal and healthy OCT images. The experimental results show that the method can effectively automatically classify the differences between the five kinds of abnormal choroidal neovascularization (CNV), diabetic macular edema (DME), edema, macular hiatus (MH) and serous macular detachment (SMD) from the healthy fundus OCT images.

The dataset for this study was obtained by adding a small dataset to account for the balance of each exception type, which may reduce the diversity of the data and affect the generalization ability of the model. To verify the effectiveness of the algorithm, future studies should involve more abnormal data sets to further optimize and make it suitable for clinical use.

References

[1] R.R.A. Bourne, J.B. Jonas, S.R. Flaxman, J. Keeffe, J. Leasher, K. Naidoo, M.B. Parodi, K. Pesudovs, H. Price, R.A. White, T.Y. Wong, S. Resnikoff, H.R. Taylor, Prevalence and causes of vision loss in high-income countries and in Eastern and Central Europe: 1990–2010, British Journal of Ophthalmology 98(5)(2014) 629-638.

[2] M.D. Abramoff, M.K. Garvin, M. Sonka, Retinal Imaging and Image Analysis, IEEE Reviews in Biomedical Engineering 3(2010) 169-208.

[3] L. Kagemann, G. Wojtkowski, M. Ishikawa, H. Ishikawa, K.A. Townsend, M.L. Gabriele, V.J. Fujimoto, J.G. Fujimoto, J.S. Schuman, Spectral oximetry assessed with high-speed ultra-high-resolution optical coherence tomography, Journal of Biomedical Optics 12(4)(2007) 041212.

[4] S.P.K. Karri, D. Chakraborty, 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)(2017) 579-592.

[5] F. Li, H. Chen, Z. Liu, X.D. Zhang, Z.Z. Wu, Fully automated detection of retinal disorders by image-based deep learning, Graefes Archive for Clinical and Experimental Ophthalmology 257(3)(2019) 495-505.

[6] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. 3rd International Conference on Learning Representations, 2015.

[7] J.J. Gomez-Valverde, A. Anton, G. Fatti, B. Liefers, A. Herranz, A. Santos, C.I. Sanchez, M.J. Ledesma-Carbay, Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning, Biomedical Optics Express 10(2)(2019) 892-913.

[8] S.L. Cheng, L.J. Wang, A.Y. Du, Histopathological Image Retrieval Based on Asymmetric Residual Hash and DNA Coding, IEEE Access 7(2019) 101388-101400.

[9] P.P. Srinivasan, L.A. Kim, P.S. Mettu, S.W. Cousins, G.M. Comer, J.A. Lzatt, Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images, Biomedical Optics Express 5(10)(2014) 3568-3577.

[10] Y. Wang, Y.N. Zhang, Z.M. Yao, R.X. Zhao, F.F. Zhou, Machine learning based detection of age-related macular degeneration (AMD) and diabetic macular edema (DME) from optical coherence tomography (OCT) images, Biomedical Optics Express 7(12)(2016) 4928-4940.

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

[12] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86(11)(1998) 2278-2324.

[13] Y. Liu, L.W.L. Yip, Y.J. Zheng, L.P. Wang, Glaucoma screening using an attention-guided stereo ensemble network, Methods 202(2022) 14-21.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception architecture for computer vision, in: Proc. 2016 IEEE Conference on computer vision and Pattern Recognition (CVPR), 2016.

[15] Y.C. Gong, Y.Q. Jia, T.K. Leung, A. Toshev, S. Loffe, Deep convolutional ranking for multilabel image annotation, in: Proc. 2nd International Conference on Learning Representations, 2014.

[16] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, in: Proc. 10th European conference on computer vision, 2008.

[17] Y.C Wei, W. Xia, M. Lin, J.S Huang, B.B. Ni, J. Dong, Y.Y. Zhao, C. Shui, HCP: A Flexible cnn Framework for multi-label Image Classification, IEEE Transactions On Pattern Analysis and Machine Intelligence 38(9)(2016) 1901-1907.

[18] Z.D. Li, K. Cheng, P.W. Qin, Y.H. Dong, C.M. Yang, X.F. Jiang, Retinal OCT Image Classification Based on Domain Adaptation Convolutional Neural Networks, in: Proc.14th International Congress on image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2021.

[19] J. Wang, Y. Yang, J.H. Mao, C. Huang, W. Xu, CNN-RNN: A unified framework for multi-label image classification, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.