Potential of Attention Mechanism for Classification of Optical Coherence Tomography Images

Zhihua Shang, Zilong Fu, Chuanbin Liu, Hongtao Xie,Yongdong Zhang
School of Information Science and Technology, University of Science and Technology of China, Hefei, China
{ shangzh , JeromeF , lcb592}@mail.ustc.edu.cn,{ htxie , zhyd73}@ ustc.edu.cn

***Abstract*—Deep neural network (DNN) can extract high-dimensional feature of images for computer vision tasks including Optical Coherence Tomography (OCT) images classification. However, OCT images are usually processed by DNN just like natural images, thus the performance of DNN is not satisfactory. We present an end-to-end DNN targeting OCT images classification. Considering the characteristic of OCT images, we introduce attention mechanism into classifier to extract more specific feature of OCT images. Our network demonstrates its capacity to enhance the features that represent the disease region. Our method achieves the state-of-the-art performance with average accuracy of 99.5% and F1-score of 0.995 on the OCT images dataset.**

Keywords—Deep Learning, Attention Mechanism, OCT Images, Multi-Classification, Fine-Grained

I. Introduction

Deep learning has been widely used in various image classification tasks, including Optical Coherence Tomography (OCT) diagnosing [1][19][24]. However, the performance of deep neural networks for OCT image classification is not so satisfying, whose performance is close to traditional image processing algorithmics [2]. There is much dissimilitude between OCT image classification and the general image classification. General images usually have obvious differences between classes, while the differences in OCT image are fairly tiny. The characteristics of OCT images are analogous—with similar structures, similar shapes, similar distributions. For the similarity between OCT images, we reckon OCT image classification as a fine-grained image recognition task.

Recognizing fine-grained categories by computer vision techniques is very challenging as some fine-grained categories can only be recognized by domain experts. For fine-grained classification, many related methods have been studies extensively in the previous literatures [2][4][5]. As other fine-grained image recognition tasks, OCT image classificater should have capacity to localize the significative regions and represent the visual differences. Commonly, the difference between OCT images locates on a small region of image. Usually, the ophthalmologists pay little attention to those similarities and meticulously distinguish the differences of abnormal images from the normal ones. To imitate this pattern, we add a soft attention mask branch to our neural network..

Attention mechanism [6][7][9][10][20][21][23] is the emphasis of our paper. Many studies have been done on attention mechanism, which can be generally divided into hard-attention [10] [20] [21] and soft-attention[9]. A represent of hard attention is Region Proposal Network(RPN), which can make neural network focus on only some regions of image and overlook the rest. RPN is used in image detection task and achieves state-of-the-art performance for a time. Nevertheless, in image classification task, datasets do not have enough supervised information to affirm which region needs more attention.

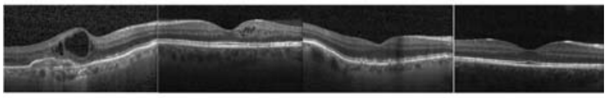

Fig.1 (A) belongs to CNV, (B) belongs to DME, (C) belongs to drusen and (D) is normal. These images are picked for the reason that our model (not our best model) classifies them mistakenly.

Soft-attention mechanism can both increase the weights of related features and retain global information without extra labelling information. In this paper, we design a deep neural network combined with a soft-attention module—Residual Attention (RA) module[9], which is based on the Inception Architecture[8]. RA represents its advantage in classification task and achieves state-of-the-art object recognition. RA takes benefit of residual learning to make its architecture deep enough. Moreover, RA's another characteristic is bottom-up and top-down feed forward structure, which enables network to be trained end to end. Inception is a classifier for general computer vision tasks. The details of how to combine RA and Inception is described in III.

Overall, the experimental comparison is conducted on the dataset in [1], which is a relatively large dataset of OCT images, and our method achieves the state-of-the-art performance. Our network can classify OCT images of normal images and three kinds of disease images with average accuracy of 99.5 in test set. Besides we get a high accuracy, the motivation of this paper is that attention mechanism is very effective for OCT images classification. We demonstrate that classifier with attention mechanism converge faster and achieve higher accuracy. Furthermore, by observing the output of attention module, it is proved that attention module indeed learns the knowledge about the disease region.

II. Related Work

978-1-5386-4458-4/18/$31.00 ©2018 IEEE

Convolutional Neural Network (CNN) is proven to be an effective architecture for computer vision tasks, whether classification or other (e.g., segmentation [11], detection [12][13][22], caption [14], etc.). The performance of neural network would be improved as increasing its depth [15][16][17]. Inception-v4 [8] is a well-known deep convolutional network, which has an outstanding performance on ImageNet classification challenge, benefitting from both its depth and width. The depth can make its nonlinear-representation capacity more excellent, and the width enables it to extract features of images in multiple scales and in wider receptive fields. These advantages play a key role to realize many kinds of computer vision tasks outstandingly. Hence, we design our neural network basing on Inception-v4.

Although deep neural network has been applied in varieties of medical image processing fields, deep neural networks for OCT image classification have not gained satisfactory answer, [2]. Although Kermany etc.[1] achieve astonishing performance, the dataset is carefully picked, thus the claimed results are not trustworthy. Most images in that dataset have distinct characteristics in contrast to the images of other categories, which can be classified easily even by non-experts. Therefore, the classifier based on deep learning for OCT images has a huge room for improvement.

III. APPROACH

A. Architectural

The backbone of our neural network is modified from Inception-v4 [8]. We replace the last fully connect（FC） layer with a new FC layer whose output is a 1×4 tensor corresponding to four categories of OCT images dataset. The structure of the stem of Inception-v4 is complex to mix attention modules—there are three blocks, each of which has several different filters, also, input size, output size and channels of blocks are diverse. Due to this reason, we mix three attention modules in middle layers, as shown in Fig.2. Residual attention module [9] has hourglass structure [18]. The top-down part will down-sample input feature map and the bottom-up part can up-sample. In our neural network, there are two types of attention modules—the first order and the second order, as shown in Fig.3, which are similar in structure. The dissimilitude is that the second order module is deeper than the first order. Besides, we don’t use more deeper modules. The reason is that in the middle layers of Inception-v4 [8] like Inception-A block, the size of feature map is 35×35, and after twice down-sampling, it would be 7×7. It is too small to down-sample. In the input of an attention block, attention modules layers generate a low resolution soft mask by convolutional layers. To the output, it up-samples the masks to the size of input by linear mapping and output the high resolution soft mask. Our experiments demonstrate that the attention mechanism can achieve great success in multi-classification of OCT images. More details are shown in section Ⅳ.

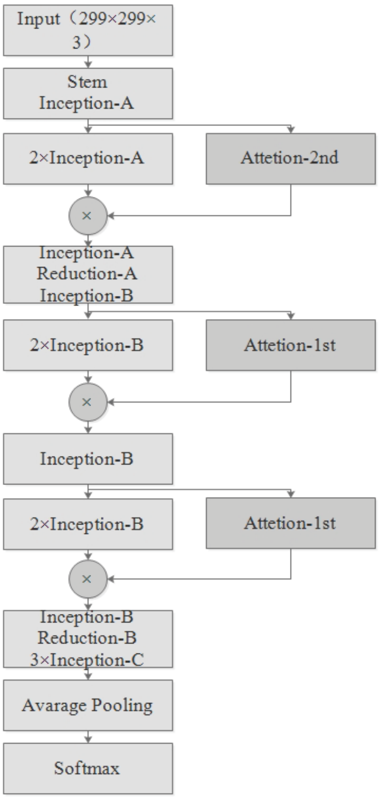

Fig.2 The schema for our neural network with attention modules.

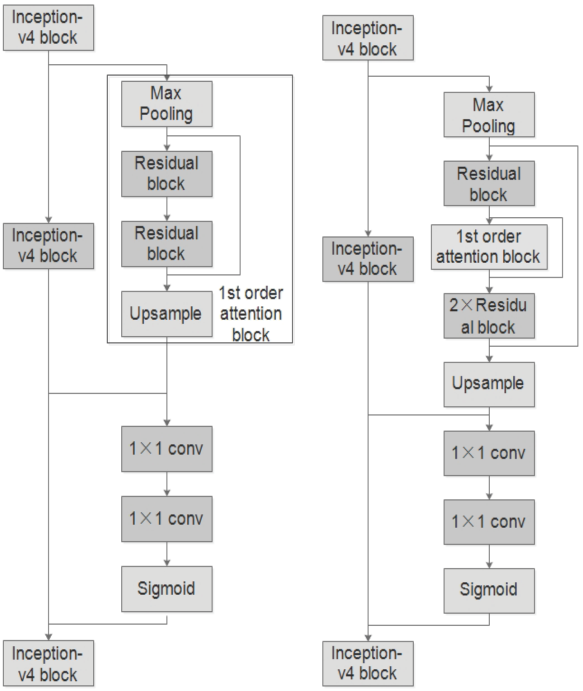

Fig.3 The schema for our 1st order attention module(left) and 2nd order attention module(right). The 1st order attention block in right refer in particular to the one in left.

B. Image Pre-processing

Since there is an uneven distribution of the number of images in training set—there are 37,205 CNV images, 26,315 normal images while only 11,348 DME images and 8,616 DRUSEN images. This unevenness makes the gap of performance of our neural network between categories huge. We try to triple the images belong to DME and DRUSEN by rotating 10° (both left and right) and flipping horizontal. Compared to the neural network trained with original dataset, the performance of the network trained with enriched dataset has an obvious improvement—both precision and F1-score are improved 1% approximately. Another defect of this dataset is that the sizes of images are different. At first, we crop input images randomly and resize them to 299×299, but when we observe the image classified incorrectly, we find that many images is cropped badly with almost background left. Therefore, we directly resize input images to 299×299. It is very effective—the precision as well as the F1-score is improved by about 2%.

C. Training

We utilize transfer learning to train our neural network. Our neural network is initialized with the pre-trained model of Inception-v4 network that is trained by Image-Net. For the backbone of Inception-v4, we only train the last three Inception-C blocks with learning rate of 0.001. And for the layers modified (last FC layer and layers of attention modules), we set the learning rate to 0.1 and train the models for 20 epochs (an epoch means that network is trained with the whole dataset). In addition, we use a weight decay of 0.0001. The Inception-v4 without attention modules is used as baseline method. To conduct fair comparison, we train its last FC layer and last three Inception-C blocks of Inception-v4 with the same hyper parameters. We also train models using different strategies to achieve lower loss and we find the models over-fitting slightly.

As complementary, we once train the last one Inception-C block instead of the last three. But it is proved not enough to fitting this task by experiments.

IV. Result

We compare our neural network with attention module with the network in [1] and the Inception-v4 [8] modified. The network in [1] uses original Inception-v3 with fine-turning. The results are shown in the TABLE I and the TABLE II. Our network outperforms other networks on both precision rate and recall rate. We achieve the best accuracy—better than Inception-v4 with 0.5% and 0.4% which seems to be not much. This is because the dataset is selective, as we discuss in Approach section, and both attention model and original Inception model achieve a very high accuracy. Despite all this, before we change pre-processing for the unbalance of dataset, the accuracy of attention model is always higher than the original Inception model 2% approximately, in TABLE III. And after the change of pre-processing, the accuracy of both two model risen close to 100% and the gap shrink.

TABLE I
Precision Rate of Networks

Network	Mean	CNV	DME	DRUSEN	Normal
Inception-v3	96.1%	93.8%	97.1%	96.7%	96.9%
Inception-v4	99.0%	98.4%	100%	98.4%	99.2%
Attention	99.5%	98.8%	100%	99.6%	99.6%

TABLE II
Recall Rate of Networks

Network	Mean	CNV	DME	DRUSEN	Normal
Inception-v3	96.6%	96.8%	94.8%	94.8%	98.4%
Inception-v4	99.0%	98.7%	98.4%	99.1%	99.5%
Attention	99.4%	99.3%	99.0%	99.7%	99.7%

TABLE III
Precision Rate of Networks With Original pre-Processing

Network	Mean	CNV	DME	DRUSEN	Normal
Inception-v4	93.6%	99.2%	88.4%	88.0%	98.8%
Attention	95.6%	99.2%	90.8%	92.8%	99.6%

(original pre-processing means random crop and no data-augmentation)

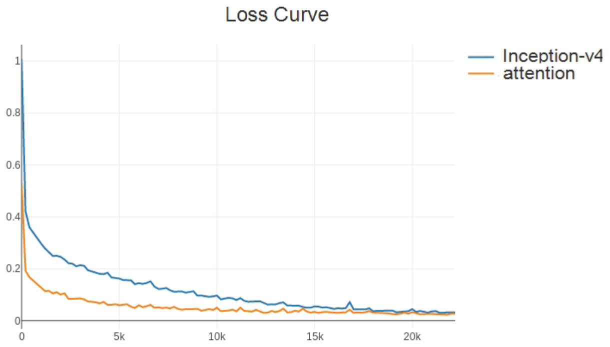

Fig. 4 Loss on training set with attention network and original network. Loss of attention network decrease faster and always lower than original network

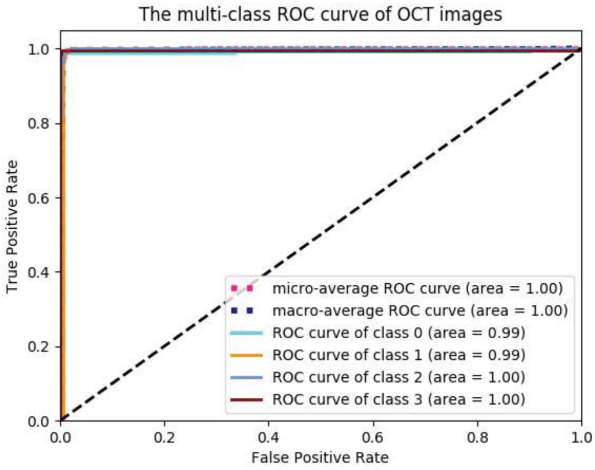

Fig.5 Receiver operating characteristic (ROC) curves for our attention network. The areas under the ROC curves are 1.00 or 0.99.

Mixing attention mechanism into Inception-v4 doesn’t make the process of training more difficult, however it can make the training faster. Fig.4 shows the loss of our network and Inception-v4 during training. Obviously, the loss of our network descends faster and achieves a lower level. It is noteworthy that our network converges faster, after one epoch it achieves the accuracy of 99.1%, while Inception-v4 achieves 90.9%.

We draw the receiver operating characteristic curves (ROCs) for four categories, micro-averaging and macro-averaging. The ROCs are plotted in Fig.5. The areas under curve (AUC) respectively are all 1.00 or 0.99.

To demonstrate attention modules has learned location information of sensitive features, the outputs of soft mask are shown in Fig.6. In B, the white regions present enhance enhancement and black regions present attenuation. Obviously, the mask form is related to the retina. Some masks enhance all features of retina region and some enhance partly. It is similar to ophthalmologist—search the whole retina and discriminately observe detail in retina region.

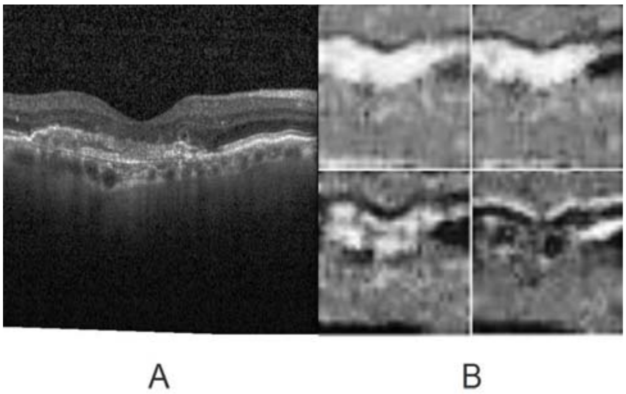

Fig.6 A is original input image. B is output of 1^{st} mask module.

V. Conclusions

In this paper, we introduce attention mechanism into OCT image classification. We design an end to end neural network based on Inception and combined with residual attention model. Our experiments demonstrate the effectiveness of attention mechanism for OCT image classification. The proposed network can learn the knowledge about sensitive information in OCT image without extra labeling information. Our model can enhance the distinction between OCT image categories, achieving the state-of-the-art accuracy of 99.5% on benchmark dataset. In the future, we will conduct the research on two directions. First, experiments on more complex and actual datasets will be conducted to evaluate the effectiveness of attention mechanism. Second, mix attention modules into different layers and observe the influence of the location of attention modules.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (2017YFC0820600), National Defense Science and Technology Fund for Distinguished Young Scholars (2017-JCJQ-ZQ-022), the National Nature Science Foundation of China (61525206, 61771468), the Youth Innovation Promotion Association Chinese Academy of Sciences (CAS, 2017209).

References

[1] Kermany, Daniel S., et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." Cell 172.5 (2018): 1122-1131.

[2] Gholami, Peyman, et al. "Classification of optical coherence tomography images for diagnosing different ocular diseases." Multimodal Biomedical Imaging XIII. Vol. 10487. International Society for Optics and Photonics, 2018.

[3] S. Branson, G. V. Horn, S. J. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. In BMVC, 2014.

[4] O. M. Parkhi, A. Vedaldi, C. Jawajar, and A. Zisserman. The truth about cats and dogs. In ICCV, pages 1427–1434, 2011.

[5] J. Krause, H. Jin, J. Yang, and F.-F. Li. Fine-grained recognition without part annotations. In CVPR, pages 5546–5555, 2015.

[6] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. arXiv preprint arXiv:1511.03339, 2015.

[7] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In NIPS, 2015. “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.

[8] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." AAAI. Vol. 4. 2017.

[9] Wang, Fei, et al. "Residual attention network for image classification." arXiv preprint arXiv:1704.06904 (2017).

[10] Girshick, Ross. "Fast r-cnn." arXiv preprint arXiv:1504.08083 (2015).

[11] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[12] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[13] He, Kaiming, et al. "Mask r-cnn." Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.

[14] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015, June). Show and tell: A neural image caption generator. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on (pp. 3156-3164). IEEE.

[15] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[16] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[17] Szegedy, Christian, et al. "Going deeper with convolutions." Cvpr, 2015.

[18] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. arXiv preprint arXiv:1603.06937, 2016.

[19] Awais, M., Muller, H., & Meriaudeau, F. Classification of SD-OCT images using Deep learning approach.

[20] Fu, Jianlong, Heliang Zheng, and Tao Mei. "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition." Conf. on Computer Vision and Pattern Recognition. 2017.

[21] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[22] Hongtao Xie, Dongbao Yang, Nannan Sun, Zhineng Chen, Yongdong Zhang. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. Pattern Recognition, https://doi.org/10.1016/j.patcog.2018.07.031

[23] Hongtao Xie, Shancheng Fang, Zheng-Jun Zha, Yating Yang, Yan Li and Yongdong Zhang. Convolutional Attention Networks for Scene Text Recognition. ACM Trans. Multimedia Comput. Commun. Appl.,2018..

[24] Hu, K., Zhang, Z., Niu, X., Zhang, Y., Cao, C., Xiao, F., & Gao, X. (2018). Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function. Neurocomputing.