2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

Attention-guided Deep Multi-instance Learning for Staging Retinopathy of Prematurity

*Shaobin Chen[1], Rugang Zhang[1], Guozhen Chen[1], Jinfeng Zhao[2], Tianfu Wang[1], Guoming Zhang[2], and Baiying Lei[*1]*
[1]National-Regional Key Technology Engineering Laboratory for MedicalUltrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China, 518060 (e-mail: leiby@szu.edu.cn)

[2]Shenzhen Eye Hospital, Shenzhen Key Ophthalmic Laboratory, The Second Affiliated Hospital of Jinan University, Shenzhen, China.

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9434012

ABSTRACT

Retinopathy of prematurity (ROP) is one of the commonest causes of acquired blindness in children. The stage of ROP is an important step to evaluate the ROP severity for disease control and management. However, there are still various challenges for ROP stage since the pattern of ROP is relatively obscure compared to the entire fundus image. Also, the dataset is small and the image quality is quite poor. To address these issues, we develop a multi-instance learning (MIL) network, which can extract the features of the images and these features can be enhanced by a fully convolutional network (FCN). The spatial score map (SSM) produced by the FCN is cropped into small patches and fed into the proposed MIL for further feature learning. An attention mechanism is leveraged to guide the MIL pooling, which can focus on the ROP features of different stages and improve the staging results. The proposed network is evaluated on an in-house ROP dataset and experimental results demonstrate that our proposed method is promising for the stage of ROP.

Index Terms— Multi-instance learning, Retinopathy of prematurity staging, Fully convolutional network.

1. INTRODUCTION

Retinopathy of prematurity (ROP) [1] is a common retinal disease in low birth weight infants, which is the main cause of blindness in children. The International Classification of ROP (ICROP) [2], which is formulated in 1984 by 23 ophthalmologists from 11 countries, provides a clinical guideline for ROP classification. According to the appearance of retinal at the blood vessels and non-vascular vessels, we use five stages to characterize the degree of ROP. If there is a "division line" between the vascular and avascular retina, we define it as Stage 1. The dividing line is comparatively flat and has unusual vascular branches on it. In Stage 2, there is a "ridge" in the "division line" and both the height and width are increased. When the symptom of fibrovascular hyperplasia appears outside the neural tube at the back or at the ridge, we can define it as Stage 3. Stage 4 has the symptom of partial retinal detachment and Stage 5 occurs total retinal detachment. In clinic, Stage 4 and Stage 5 are relative obvious and can be observed by the binocular indirect ophthalmoscope, which is a gold standard method for ROP diagnosis [3]. However, the appearance of Stage 1, Stage 2 and Stage 3 are similar, thus experienced ophthalmologists and image records by wide-angle digital retinal imaging system such as RetCam3 are necessary for the diagnosis. The Normal, Stage 1, Stage 2, Stage 3, Stage 4 and Stage 5 ROP are shown in Fig. 1, respectively.

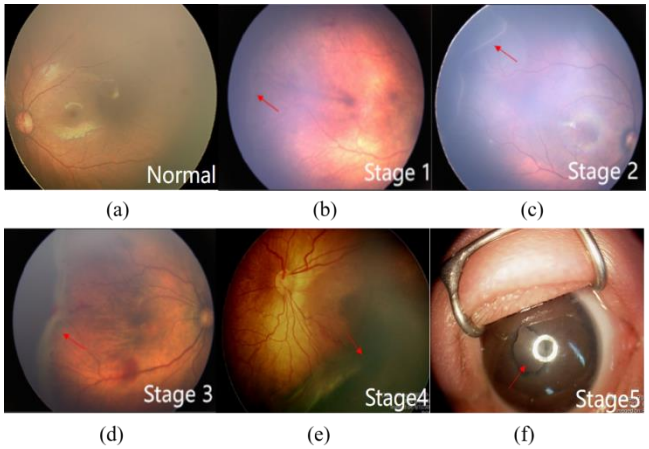

Fig. 1. Normal (a) and different stages of ROP. Stage 1 (b) ROP presents a slight division line (red arrow) at the junction between vascular and avascular. In Stage 2 (c) appears a ridge and the width and height of the division line increase. Stage3 (d) appears fibrovascular hyperplasia outside the neural tube at the back or at the ridge. Stage 4 (e) has partial retinal detachment and Stage 5 (f) arises total retinal detachment.

The stage of ROP is complicated because of the following reasons: 1) for early stage of ROP (Stage 1 and Stage 2), the appearance of the ROP lesion is unclear and of small region in contrast to the whole fundus images (see Fig. 1); 2) the appearance of different stage of ROP is similar; 3) the clinical evaluation of ROP staging primarily relies on the subjective interpretation of the symptoms by ophthalmologists, which requires professional knowledge and takes a long time. To address these challenges, a method to extract high-level features and robust for small dataset is required to complete the ROP stage task. In view of this, we design a fully convolutional network (FCN) [4] to extract the ROP features and generate a spatial score map (SSM), which provides a pixel level probability of the ROP lesion. Since the ROP dataset is small, to obtain a robust staging performance, the SSM is cropped into small patches to increase the dataset and fed into a multiple-instance learning (MIL) [5] network for training. However, although different MIL poling methods

Authorized licensed use limited to: UNIVERSITY OF ABERDEEN. Downloaded on September 25,2023 at 13:24:07 UTC from IEEE Xplore. Restrictions apply.

had been applied to improve the network performance, the results are still not satisfactory. Besides, the SSM and the patch size are not well studied. In this paper, an attention module is employed to merge the extracted features based on bag-level pooling and get the final categories of the ROP stage. Different patch sizes are compared to optimize the network performance. Furthermore, to enlarge the dataset and increase robustness of the network, we collect the data from multi-hospital and annotate the data by three ophthalmologists. Experimental result shows that our presented method has achieved the best performance comparing the traditional module in detecting the stage of ROP. In general, our contributions are as follows:

1)The proposed framework includes two main modules, a FCN module to complete end-to-end segmentation tasks and a MIL module to generate a SSM, which provides a pixel-wise probability of the ROP lesion.

2) The SSM is cropped into different size of small patches to fit the MIL setting, which effectively expands the size of the dataset.

3) An attention module is employed to merge the features based on bag-level pooling and get the final categories of the ROP stage.

2. METHODOLOGY

2.1. The fully convolutional network module

The architecture of our proposed model is shown in Fig. 2. We introduce FCN to get the local features from the fundus image and produce the SSM. To acquire a pixel level output (i.e., SSM), the fully connected (FC) layer of the convolutional neural network (CNN) is removed from our model and then take up-sampling. This method has been proved to be an effective end-to-end pixel level segmentation method. We use the SSM as the MIL input instead of the original image.

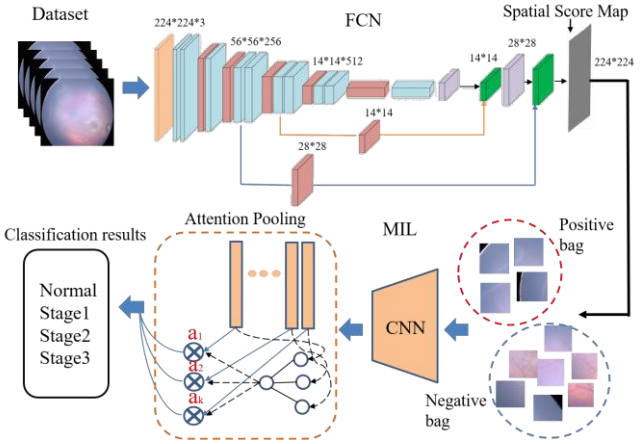

Fig. 2. Architecture of our model. The framework includes two modules, a FCN and MIL module.

In this paper, the original size of the retinal image is first downed sample to 224×224×3 and goes through four pairs of convolution layers and pooling layers. At the fully FC layers, the channels are amended from {4096; 4096; 1024} to {1024; 1024; C}, where C is the number of categories in our task, so the size of weight matrix for first FC layer is 1024×1024×14. For convolution representation, the weights of the first FC layer are reshaped into a 4-dimensional tensor with size of 1024×512×7×7. An up-sampling is performed after the FC layer and size of the feature map is restored to 224×224×C (i.e., SSM).

2.2. Multiple Instance Learning Network

Deep learning for medical image classification faces two common challenges: 1) regions of interest (ROIs) are relatively small with unclear boundaries in the whole medical images; 2) the size of the dataset is small and unbalance. For the first challenge, the impact of the ROIs on the global feature vector is limited, which seriously impacts the performance of the resulting classifiers. MIL network, which firstly introduced with weakly labelled to tackle the problem of drug activity prediction [6], has been successfully used in detecting medical image classification problem. MIL is a natural learning scenario for medical image classification since it allows to automatically detect target patterns locally in images and to propose an automatic diagnosis for each image as a whole. MIL is an extension of supervised learning. It can train classifiers to use such weakly labeled data, which are in such bags and single instances in the bags have no labels. In MIL, if there is a positive instance, the bag is assumed to be positive, and if all instances in the bag are negative, the bag is assumed to be negative. Initially, MIL was formulated as a binary classification problem, but multi-class extensions were also proposed. Another advantage of MIL in medical imaging learning is that it can process a single image.

To sustain the MIL assumption and better deal with the small dataset problem in the ROP stage task, the network is modified by cropping the SSM produced by the FCN to small non-overlapping patches with equal size and fed the patches into the MIL for training. In this way, the training images are enlarged dozens of times. The SSM is an instance package which we call it a bag, and each patch in it is also an instance. In this study, a total number of five patch sizes are used to evaluate the network performance.

2.3. Attentive MIL Pooling

In this study, we use MIL module to learn a function which can map the input dataset $\{X_1, \ldots, X_N\}$ to the corresponding label set $\{Y_1, \ldots, Y_N\}$, where $Y_i \in \{0,1\}$. Assuming X_k is a bag with many instances, where $k \in \{1, \ldots, N\}$. Then we define the instance in the bag as $\{X_{k1}, \ldots, X_{kn}\}$, where n is the total number of instances of the bag X_k. The symmetric function to solve the MIL problem is denoted as:

$$f(X) = \theta\left(\eta_{x \in X} \varphi(x)\right), \quad (1)$$

where φ is suitable transformations and η is the permutation invariance of the function. θ is an evaluation function for the instance in the bag. As an essential step in bridging instances to bags, different applications have different preferences for MIL pool methods. In this study, the prediction probability of X_{kj} is denoted as $p_{kj}^{c} =$

Authorized licensed use limited to: UNIVERSITY OF ABERDEEN. Downloaded on September 25,2023 at 13:24:07 UTC from IEEE Xplore. Restrictions apply.

$P(c_{kj} = c|X_{kj})$, where c is the category of our task and $c \in \{1,\ldots, C\}$. The polymerization function $\mathcal{F}$ of the bag X_k can be defined as follows:

$$p_k^c = P(c_k = c|X_k) = \mathcal{F}(p_{k1,}^c \ldots, p_{kn}^c). \quad (2)$$

This research will find the polymerization function $\mathcal{F}$ to achieve the best performance.

The purpose of the attention MIL pooling is to assign weights to instances by training neural networks. In our proposed method, it is used in the bag-level. Additionally, the weights must sum to 1 to be unaffected by the bag size. The weighted average meets the requirements that the weights together with the embedding are part of the f function. Let $H = \{h_1, \ldots, h_k\}$ be a bag of k embeddings, then we propose the following MIL pooling:

$$z = \sum_{k=1}^{K} a_k h_k, \quad (3)$$

$$a_k = \frac{exp\{W^T tanh(Vk_k^T)}{\sum_{j=1}^{K} exp\{W^T tanh(Vk_k^T)}, \quad (4)$$

Moreover, we use the nonlinearity of the hyperbolic tangent element to include both positive and negative values for appropriate gradient flow. In this study, max pooling and Softmax pooling are employed on every instance as the instance-level MIL pooling to compare the MIL classifier performance.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

Our dataset consists of 1558 ROP cases from 2015 to 2018 from Shenzhen Eye Hospital, Shenzhen Maternal and Child Health Hospital and Weifang Eye Hospital. At each examination, a standard 10-views photograph of the baby's eyes is taken. We just choose images which are consistent with the masks of all ophthalmologists to train, and discard the images with inconsistent marks. Finally, we get 6209 retinal images from 893 ROP inspections. In order to assess the performance of our method, two datasets are used to compare the network performance. At last, we divide the images into cross validation set, training set and testing set, respectively. Dataset I and Dataset II separate the images into four groups, which include Stage 1, Stage 2, Stage 3 and Normal. However, a data size in the dataset I which is similar to other class of the Normal class is selected for training while 2400 Normal images in Dataset II are used in the training procedure. In Dataset I, there are 1000 labeled as Normal and 360 labeled as Stage 1. In Dataset II, among all infants, there are 4000 labeled as Normal and only 360 labeled as Stage1. In addition, in Dataset I and Dataset II, there are 1064 labeled as Stage 2 and 785 labeled as Stage 3, respectively. Thus, the dataset is quite unbalanced. Dataset I and Dataset II are used to verify the effectiveness of the proposed imbalance data learning.

3.2. Experimental Setup

We use the Pytorch framework on NVIDIA TITAN XP GPU with 12 GB RAM, which can accelerate the speed of our work. In order to save the computation resources, we also resize the training dataset. By rotating each adjusted image by 90 degrees, we augment the datasets to four times. In our task, we set the maximum epoch to 100 and set the batch size to 10. The initial learning rate is set to 0.0001, and then reduced by 0.9 attenuation when the train loss converges.

In order to assess our proposed method, we use many common evaluation criteria, such as specificity (SPEC), area under the curve (AUC), sensitivity (SEN), accuracy (ACC) and F1 score (F1). For different models, we evaluate the accuracy of total classification and the test accuracy for each stage of ROP.

3.3. Overall Staging Performance

To emphasize the effectiveness of the FCN network, we use three MIL pooling [7] algorithms including Max Pooling (MP), Softmax Pooling (SP) and Attention Pooling (AP). We evaluate the performance of different MIL pooling algorithms compared to the proposed MIL network with and without the FCN for feature exaction and different MIL pooling in the aggregated procedure. MIL without FCN for feature exaction means the fundus images are resized and directly cropped into small patches and fed into the aggregated procedure. As seen in Table 1, the classification accuracies (%) for Stage 1(S1), Stage 2(S2), Stage 3(S3), Normal (NC) and Total(T) of the proposed model in dataset I with three MIL pooling outperforms that without FCN by a large margin.

The results indicate that the segmentation by the FCN has a supper performance in exacting high-level features of the fundus images and the classification performance of MIL is greatly improved. Our proposed attention mechanism successfully enables neural networks to focus more on the most likely positive instances. In this way, the model can be more interpretable, and the key information can be detected from the fundus image. To verify the imbalance learning algorithm performance, we randomly select 1000 normal fundus image from dataset II and construct the new dataset I. The model performance in Table 1 using dataset II outperforms that of dataset I with a substantial margin.

3.4. Staging Performance Comparison

In this study, we use some common models such as AlexNet [8], VGG [9], ResNet50 [10], ResNet101 and Inceptionv4 [11] to compare the performance of our models. Experimental results in the dataset II are shown in Table 2. We can see that there is an obvious difference of staging performance among different models. Our method, FCN+MIL, had the best performance in terms of all evaluation metrics. In fact, all the evaluation metrics of our method is much better than all the other models. The receiver operating characteristic (ROC) curve of our method using FCN+MIL framework and other networks are shown in Fig. 4.

Authorized licensed use limited to: UNIVERSITY OF ABERDEEN. Downloaded on September 25,2023 at 13:24:07 UTC from IEEE Xplore. Restrictions apply.

Table 1. Comparison of our model and MIL network without FCN(%).

	P	S	MIL without FCN			FCN+MIL		
			ACC	AUC	F1	ACC	AUC	F1
Dataset I	MP	S1	80.3	88.4	80.1	87.53	91.1	86.6
		S2	87.1	91.2	86.2	89.27	94.3	88.3
		S3	87.2	91.5	86.7	89.3	94.6	88.4
		N	90.7	94.1	89.5	92.7	96.7	91.1
		T	88.3	93.5	87.4	91.5	95.8	90.1
	SP	S1	78.5	87.6	78.1	86.3	90.5	85.55
		S2	83.3	89.5	82.9	89.1	94.2	88.28
		S3	85.5	90.5	85.0	88.9	93.7	88.1
		N	90.0	93.8	89.3	92.2	96.3	91.7
		T	88.1	93.1	87.6	90.9	95.1	89.8
	AP	S1	81.4	89.2	81.9	89.3	93.8	88.63
		S2	85.2	90.7	84.7	90.1	94.7	89.24
		S3	86.3	92.1	85.6	91.3	94.7	90.1
		N	91.9	95.3	91.0	93.5	96.8	92.6
		T	89.5	93.8	88.3	92.4	95.9	91.5
Dataset II	MP	S1	83.5	90.2	82.7	88.7	93.4	87.86
		S2	89.4	92.8	88.8	92.5	95.5	91.53
		S3	90.4	93.1	89.6	92.7	95.7	91.6
		N	92.5	96.3	91.7	95.5	97.2	94.3
		T	89.3	93.5	88.4	93.2	96.3	92.1
	SP	S1	82.8	89.5	82.2	87.8	92.3	87.12
		S2	88.9	93.2	88.1	91.5	94.5	89.97
		S3	89.7	93.7	87.8	91.2	95.0	89.7
		N	92.2	95.1	91.8	94.7	96.6	93.2
		T	88.4	93.1	87.6	92.8	95.8	91.3
	AP	S1	85.3	90.7	84.5	90.2	94.7	89.2
		S2	90.3	95.2	89.5	94.1	96.8	93.0
		S3	91.1	96.1	90.4	93.8	96.3	91.9
		N	95.4	98.3	94.2	96.7	98.5	95.2
		T	90.2	95.1	89.0	94.4	97.2	92.2

Table 2. Performance for the proposed model and pure MIL network (%).

Method	ACC	SEN	SPEC	AUC	F1
VGG16	86.53	84.22	85.33	86.4	86.31
AlexNet	87.21	88.33	84.54	84.5	83.64
ResNet50	79.46	76.87	80.27	81.5	78.75
InceptionV4	87.23	88.35	84.52	84.3	83.62
MobileNet	89.42	88.32	88.91	89.8	87.53
OURS	**94.43**	**92.71**	**94.23**	**97.2**	**92.27**

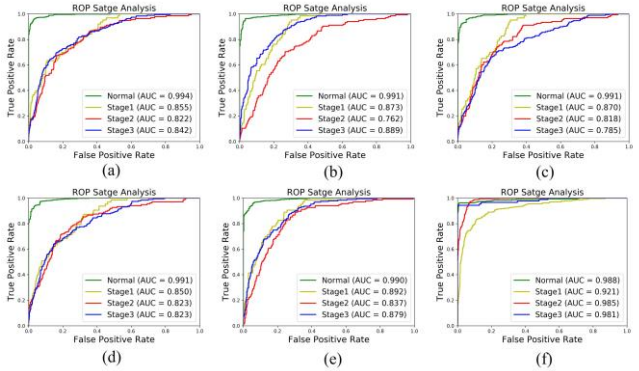

Fig. 4. ROC curves of each class using the proposed method with FCN+ MIL framework and the compared networks.

4. CONCLUSION

In this paper, we propose an attention-guided FCN+MIL network for ROP staging, where the FCN is responsible to extract the high-level features of the fundus images and generate an SSM. The SSM, which is a pixel level probability of the ROP lesion in the same spatial location with the original image, is cropped into different sizes of small patches and fed into the MIL network for learning. To further address the critical issues of the slight pattern in the ROP stage task, an attention mechanism is leveraged to guide the MIL pooling, which can focus on the ROP features of different stage and improve the staging results. The experimental results demonstrate that our proposed method is effective and the results are promising to stage ROP.

5. COMPLIANCE WITH ETHICAL STANDARDS

Conflict of interest The authors declared that they have no conflict of interest to this work.

6. ACKNOWLEDGMENT

This work was supported partly by National Natural Science Foundation of China (This work was supported partly by National Natural Science Foundation of China (Nos.61871274, and U1909209), Key Laboratory of Medical Image Processing of Guangdong Province (No. K217300003), Guangdong Pearl River Talents Plan (2016ZT06S220), Shenzhen Peacock Plan (Nos. KQTD2016 053112051497 and KQTD2015033016104926), and Shenzhen Key Basic Research Project (Nos. JCYJ20180507184647636 and JCYJ20170818094109846).

7. REFERENCES

[1] A. Hellström, L. E. Smith, and O. J. T. l. Dammann, "Retinopathy of prematurity," vol. 382, no. 9902, pp. 1445-1457, 2013.

[2] I. C. f. t. C. o. R. o. P. J. A. o. ophthalmology, "The international classification of retinopathy of prematurity revisited," vol. 123, no. 7, p. 991, 2005.

[3] J. M. Brown, J. P. Campbell, A. Beers, K. Chang, S. Ostmo, R. P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, and J. J. J. o. Kalpathy-Cramer, "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks," vol. 136, no. 7, pp. 803-810, 2018.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[5] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 603-611: Springer.

[6] T. G. Dietterich, R. H. Lathrop, and T. J. A. i. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," vol. 89, no. 1-2, pp. 31-71, 1997.

[7] H. Liu, D. W. Wong, H. Fu, Y. Xu, and J. Liu, "DeepAMD: detect early age-related macular degeneration by applying deep learning in a multiple instance learning framework," in *Asian Conference on Computer Vision*, 2018, pp. 625-640: Springer.

[8] A. Krizhevsky, I. Sutskever, and G. E. J. C. o. t. A. Hinton, "Imagenet classification with deep convolutional neural networks," vol. 60, no. 6, pp. 84-90, 2017.

[9] K. Simonyan and A. J. a. p. a. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

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

[11] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31, no. 1.

Authorized licensed use limited to: UNIVERSITY OF ABERDEEN. Downloaded on September 25,2023 at 13:24:07 UTC from IEEE Xplore. Restrictions apply.