Deep Learning based for Fundus Analysis

Benjian Zhao, Weixin Liu, Rugang Zhang and Baiying Lei

Shenzhen University, Shenzhen, China

Abstract. Glaucoma is an irreversible eye disease and it is considered the second leading cause of blindness globally. Its early diagnosis is very important to prevent glaucoma. Great progress has been made in the automatic diagnosis and ONH analysis of fundus images based on deep learning. However, the deep learning model relies on a large number of labeled data. Although more and more datasets have been proposed for deep learning research, the differences between different datasets hinder the performance of deep learning model. In this paper, in order to improve the classification performance and generalization performance of glaucoma, we first use five different scale images and models to learn robust glaucoma classification, and then use ensemble learning to merge their results. For the joint segmentation of optic disc and cup, we use CENet network to improve the segmentation ability of the model. This network uses ResNet34 as the encode path, and proposes a context encoder module to extract deeper features and retain more spatial information. At the same time, residual connection is used to reduce over fitting. For the fovea localization task, we use a pixel by pixel regression method to transform the localization task into a regression task.

1 Glaucoma Classification

Automatic fundus image process based on deep learning has made progress, and many effective models have been proposed to solve the task of fundus image diagnosis and segmentation [1]. However, deep learning model depends on a large number of labeled data, and training in a small data set may lead to the problem of model over fitting [2]. Although more and more fundus image data sets have been proposed for deep learning research, there are differences between these data sets due to the different cameras and different imaging protocols. The generalization performance of the model trained on one dataset is greatly reduced in other datasets.

In order to improve the generalization of the model on the new dataset, we use several models of different scales to train the glaucoma classification, and use ensemble learning to merge their results. Specifically, we use five different methods, as shown in Figure 1. Our input image includes a full image and a crop image of the disc area. We first train ResNet50 [3] model to extract global and local features for the whole image and the crop image of disc area respectively. However, considering that the model is easy to over fit and has poor generalization, we modify the ResNet50 model and fuse the multi-scale features in the model for classification. We replace the fully connection layer with global average pooling (GAP) [4], where 2GAP and 3GAP

represent the fusion of the last two or three scales of the model, respectively, to improve the robustness of classification [5, 6]. Finally, we use 1×1 convolutional compression feature channel (CC) to reduce over fitting and improve channel robustness.

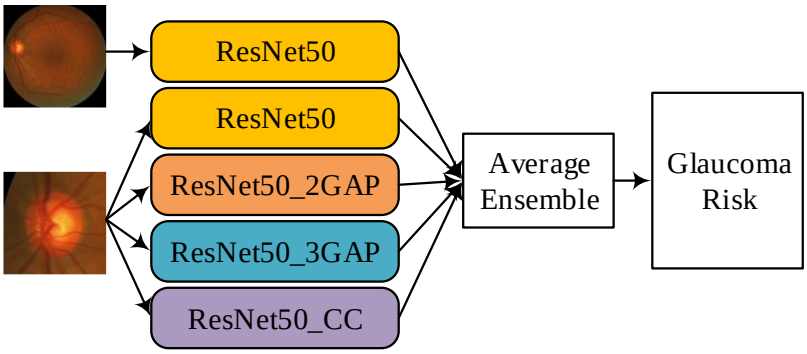

Figure 1. Ensemble learning models of glaucoma classification.

We use cross entropy loss to train each model. The purpose of training these models is to gradually improve the robustness of classification. Finally, experiments show that integrating all the results can improve the classification performance.

2 Optic disc and cup segmentation

2.1 Backbone network

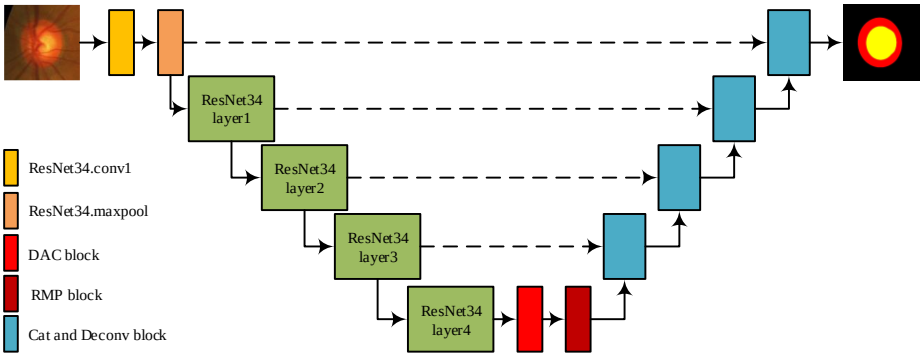

Figure 2. The architecture of the CENet. It is based on UNet model, and ResNet34 module replace the encode path. And a new context encoder module is proposed in CENet, which is composed of dense atrous convolution block and a residual multi-kernel pooling block.

In the task of segmentation of optic disc and cup, we use [7] proposed model CENet, as show in Figure 2. Specifically, the model is based on the structure of UNet [8], but ResNet34 [3] is used to replace the encode path in UNet to increase the ability of feature extract. At the same time, ImageNet pretrained parameters are used to speed up the model training. In order to alleviate the consecutive pooling and strided

convolutional operation led to reduce the feature resolution, loss of some spatial information, a context encoder module is proposed in CENet, which is composed of dense atrous convolution (DAC) block and a residual multi-kernel pooling (RMP) block. Finally, in the decode path, the decoder module first combines the features in the encode path by skip connection to obtain some spatial information, and then recovers the high-resolution features in the decoder by 1×1 convolution and 3×3 deconvolution respectively. The dice loss of the loss function is used for training, and we use ellipse fitting to smooth the segmentation boundary.

2.2 Context encoder module

In the encode path of UNet, in order to learn the abstract feature representation, the feature resolution is gradually reduced. This feature representation usually hinders the intensive prediction task which needs detailed spatial information. In order to capture more high-level features, and preserve more spatial information in the encoder to improve the performance of segmentation, CENet is proposed a context encoder module which composed of dense atrous convolution (DAC) block and a residual multi-kernel pooling (RMP) block.

DAC block are inspired by the Inception-Resnet [9] structure, as shown in the Figure 3. DAC block can capture more extensive and deeper semantic features by fusing four cascaded branches with multi-scale atrous convolution. In this module, residual connection is used to prevent the gradient from disappearing.

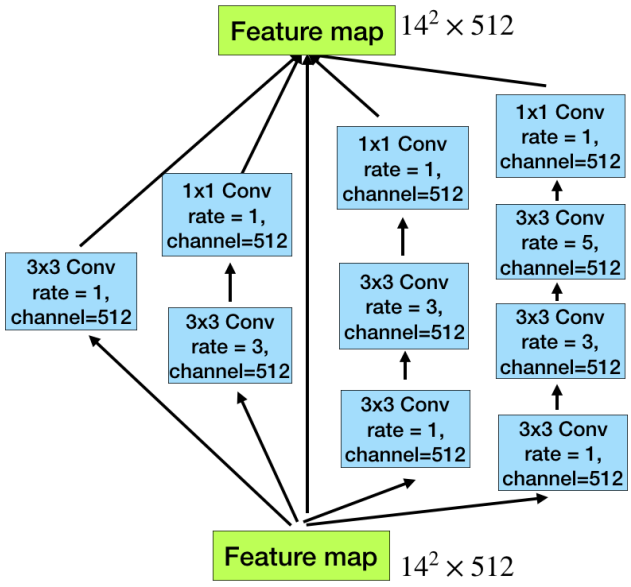

Figure 3. The dense atrous convolution (DAC) block architecture. It contains four cascade branches with the gradual increment of the number of atrous convolution. And it use a residual connection to prevent the gradient disappearing.

RMP block is used to address the problem of the variation of object size, its mainly relies on multiple effective field-of-views to detect objects at different sizes. The size

of receptive field roughly determines how much context information we can use. The general max pooling operation just employs a single pooling kernel, such as 2×2. As illustrated in Figure. 4, the proposed RMP encodes global context information with four different-size receptive fields: 2×2, 3×3, 5×5 and 6×6. The four-level outputs contain the feature maps with various sizes. To reduce the dimension of weights and computational cost, RMP block use a 1×1 convolution after each level of pooling. It reduces the dimension of the feature maps to the $1/N$ of original dimension, where N represents number of channels in original feature maps. Then upsample the low-dimension feature map to get the same size features as the original feature map via bilinear interpolation. Finally, concatenate the original features with upsampled feature maps.

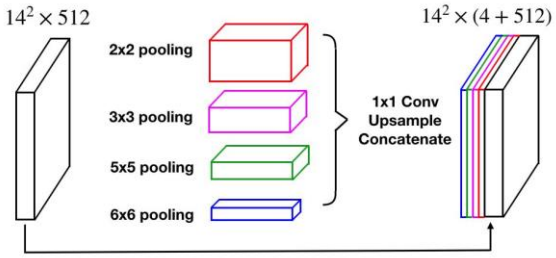

Figure 4. The residual multi-kernel pooling (RMP) block architecture. The RMP gather context information with four different-size pooling kernels. And use 1×1 convolution to reduce the dimension of feature maps.

3 Fovea localization

Fovea is the darkest region in the retinal, which has no blood vessels and lacks features, which poses a great challenge to the localization task. We use the [10] proposed method, it is construct the fovea location task as a pixel by pixel regression task. The regression quantity includes the distance from the nearest landmark of interest. After optimization, the fully convolution neural network can predict the distance of each image position, thus the problem is implicitly introduced into the multi-task learning method of each pixel, so as to understand the global uniform distribution of the distance in the whole image. According to the method in [10], the joint learning of each pixel position related to the optic disc and fovea helps to automatically understand the overall anatomical distribution. This idea is shown in the Figure 5. A fully convolution neural network is used to solve the regression problem, and the information of each pixel is helpful to generate a globally consistent prediction graph by using the method multi-task learning.

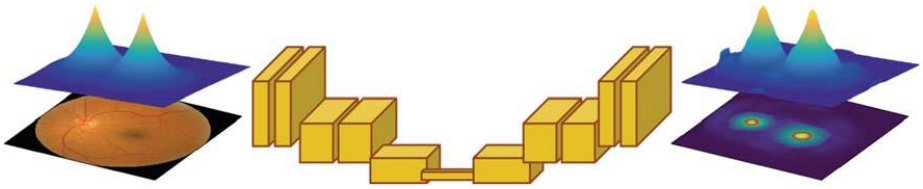

Figure 5. The method for joint fovea and OD localization via regressing a distance map.

In the implementation, we use the UNet to solve the regression task. For this, we first define the Bi-Distance Map $\mathcal{B}(x,y)$ for each pixel location $(x,y) \in \Omega, \Omega \in \mathbb{R}^2$, the image domain on which a retinal image $I(x,y)$ is defined. Given the location of the OD (x_{od}, y_{od}) and the fovea (x_{fov}, y_{fov}), $\mathcal{B}(x,y)$ is defined as follows:

$$\mathcal{B}(x,y) = \min\left(\sqrt{(x-x_{od})^2 + (y-y_{od})^2}, \sqrt{\left(x-x_{fov}\right)^2 + \left(y-y_{fov}\right)^2}\right), \quad (1)$$

From the Bi-Distance Map definition, a normalized form, bounded in [0, 1], can be easily built:

$$\mathcal{B}^N(x,y) = \left(1 - \frac{\mathcal{B}(x,y)}{\max_{\Omega} \mathcal{B}(x,y)}\right)^{\gamma}, \quad (2)$$

where γ is a decay parameter governing the spread of $\mathcal{B}^N$ across the image domain, in our experiment, we set $\gamma = 7$.

The output of the above model is a smooth prediction of the distance to two interesting landmarks. For this purpose, Laplacian-of-Gaussian operator is applied to extract the two most prominent maxima.

4 Experimental

4.1 Datasets

We used five related datasets, namely REFUGE [11], ORIGA [12], Drishti-GS1 [13], RIMONE_r3 [14], ACRIMA [15], details are shown in Table 1. In these datasets, we alternately select one of the data sets as the validation set, and the rest of the data sets are mixed for training to find the model with the best generalization performance.

Table 1. List of used publicly available datasets. GT stands for Ground Truth.

Dataset	Glaucoma	Non-Glaucoma	OD/OC GT	Fovea GT
REFUGE	120	1080	Yes	Yes
ORIGA	168	482	Yes	No
Drishti-GS1	70	31	Yes	No
RIMONE_r3	39	85	Yes	No
ACRIMA	396	309	No	No

4.2 Implementation

Our network implementation uses Python based on Pytorch [16]. We used an NVIDIA Titan XP GPU with 11GB of memory to speed up model training and testing. Due to the limitation of GPU memory, we resize the whole image to 512×512 to train classification and fovea localization. Because the size of the optic disc in different data sets is different, we locate the optic disc area, and then take it as the crop size about twice the diameter of the optic disc, and then resize it to 512×512. During the training period, we use RAdam [17] to optimize the network model. The advantage of RAdam is that it combines the fast convergence speed of Adam and the high precision of SGD optimization, avoiding the problem that Adam may fall into the local optimum. The batch size is 16, and a total of 200 epoch is trained. The initial learning rate is 0.001, and the learning rate decay strategy is used, which is multiplied by 0.1 every 30epoch.

References

1 A.R. Ran, C.Y. Cheung, X. Wang, H. Chen, L.-y. Luo, P.P. Chan, M.O. Wong, R.T. Chang, S.S. Mannil, A.L. Young: Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis. The Lancet Digital Health, vol. 1, pp. e172-e182 (2019).

2 G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez: A survey on deep learning in medical image analysis. Medical Image Analysis, vol. 42, pp. 60-88 (2017).

3 K. He, X. Zhang, S. Ren, J. Sun: Deep residual learning for image recognition. CVPR, pp. 770-778, (2016).

4 B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba: Learning deep features for discriminative localization. CVPR, pp. 2921-2929, (2016).

5 R. Zhao, W. Liao, B. Zou, Z. Chen, S. Li: Weakly-supervised simultaneous evidence identification and segmentation for automated glaucoma diagnosis. AAAI, pp. 809-816, (2019).

6 W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, M. Zhou: Clinical Interpretable Deep Learning Model for Glaucoma Diagnosis. IEEE journal of biomedical health informatics, vol. 24, pp. 1405-1412 (2019).

7 Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu: Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging, vol. 38, pp. 2281-2292 (2019).

8 O. Ronneberger, P. Fischer, T. Brox: U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, pp. 234-241, (2015).

9 C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi: Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI, (2017).

10 M.I. Meyer, A. Galdran, A.M. Mendonça, A. Campilho: A pixel-wise distance regression approach for joint retinal optical disc and fovea detection. International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 39-47, (2018).

11 J.I. Orlando, H. Fu, J.B. Breda, K. van Keer, D.R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee: REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis, vol. 59, pp. 101570 (2020).
12 Z. Zhang, F.S. Yin, J. Liu, W.K. Wong, N.M. Tan, B.H. Lee, J. Cheng, T.Y. Wong: Origa-light: An online retinal fundus image database for glaucoma analysis and research. 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 3065-3068, (2010).
13 J. Sivaswamy, S. Krishnadas, G.D. Joshi, M. Jain, A.U.S. Tabish: Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. 2014 IEEE 11th international symposium on biomedical imaging (ISBI), pp. 53-56, (2014).
14 E. Medina-Mesa, M. Gonzalez-Hernandez, J. Sigut, F. Fumero-Batista, C. Pena-Betancor, S. Alayon, M. Gonzalez de la Rosa: Estimating the amount of hemoglobin in the neuroretinal rim using color images and OCT. Current Eye Research, vol. 41, pp. 798-805 (2016).
15 A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J.M. Mossi, A. Navea: CNNs for automatic glaucoma assessment using fundus images: an extensive validation. Biomedical engineering online, vol. 18, pp. 29 (2019).
16 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, pp. 8026-8037, (2019).
17 L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han: On the Variance of the Adaptive Learning Rate and Beyond. International Conference on Learning Representations, (2019).