The Solution of Team PingAn Smart Health for REFUGE2 Challenge

Ge Li, Yang Liu, Chengfen Zhang and Dongyi Fan
PingAn Technology (shenzhen), Co, Ltd, China
Lige676@pingan.com.cn

Abstract. The REFUGE2 challenge focuses on the investigation and development of algorithms associated with the diagnosis of glaucoma in fundus photos. The goal of the challenge is to evaluate and compare automated algorithms for glaucoma detection and optic disc/cup segmentation on a standard dataset of retinal fundus images. The challenge includes three tasks: Classification of clinical Glaucoma, Segmentation of Optic Disc and Cup, and Localization of Fovea (macular center). The technical report describes our solution for each task in details.

Task 1 Classification

In the task of classification, there are two kinds of labels, namely nonglaucoma fundus and glaucoma fundus. We need to predict the risk of glaucoma for each image. So, we'll look at data preparation, network structure, and results output.

1.1 Data Preparing

Since only 1200 images are released in the training dataset, the amount of data is too little for the classification network, so it is easy to cause the over-fitting of the network results. Therefore, the training dataset is amplified online by image rollover, color transformation and small angle image rotation, and the image is rescaled before the image is inputted into the network.

After dozens of epoch training, we use the images on the validation dataset to validate the performance of the model. In the verification process, we found that there is a type of glaucoma fundus image with the position of optic disc close to the image edge, which can be easily divided into normal fundus images, resulting in false negative. Therefore, we selected 290 of these type of fundus images from ODIR exposure datasets. After cropping black edges and changing colors, these fundus images are put into the training dataset to help reduce the generation of such false negative images. In addition, each validation image is transformed by flipping, lighting and so on before inference and the final prediction is obtained by fusing the output results of the model. After these data preparation, the false negative images are eliminated completely, and the performance of the model is improved obviously.

1.2 Network structure

We used three lightweight networks for classification tasks. At first, we trained each of the networks separately. The networks were trained using binary cross entropy loss. And to maximize the contribution of all images, we obtained different data groups through cross-validation, and then conduct multiple model training in stage 1. After the training completed, the last convolution block of each network is concatenated. Finally, we fixed the other layers of the nine networks except the last convolution layers and trained a new classification network.

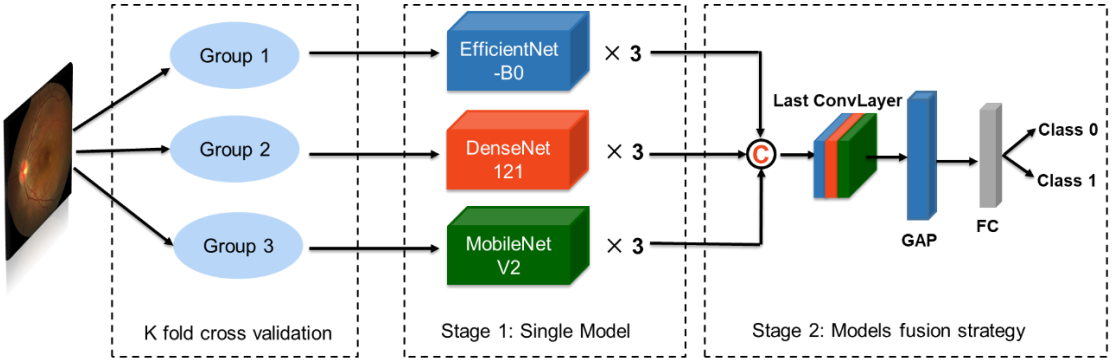

Fig. 1. Classification fusion network structure.

1.3 Results

In this report, we used multiple classification models to validate the results of the validation dataset. The AUC of the basic model is 0.950 and experiments show that when external datasets and models fusion are used, the results are further improved. In addition to model fusion, each validation image is transformed by flipping, contrast adjust and so on. Then the transformed images are fed into the fusion model for prediction, and the final result is the average value of multiple sets of results. This trick can reduce the impact of optic disc position and color style on results. We organize the results as shown in Table 1.

Table 1. Classification results on the test set.

Model	AUC
Basic model	0.950
Basic model + Histogram matching	0.959
Basic model + Histogram matching + External datasets	0.967
Models fusion	0.984
Models fusion + validation image transformation	0.985

Task 2 Optic Disc and Cup Segmentation

In this task, we proposed a two-stage module for the optic disc and cup segmentation. At the first stage, the optic disc region will be detected from the fundus image by the existing method Mask R-CNN. At the second stage, the optic disc region is cropped by the detected bounding box and then fed into our designed DA-MobileNet for segmentation.

2.1 Data Augmentation

We performed data augmentation to increase the amount and richness of training data. In addition to the usual transformation such as random rotation, random flipping, and random color transformation, we also add random motion blur to rich the dataset for improving the generalization of the model. In order to make the hard sample suffering from the severe uneven color distribution can be detected and segmented, we add the data that using histogram matching to transform the image from a style to another style to the training set.

2.2 Network structure

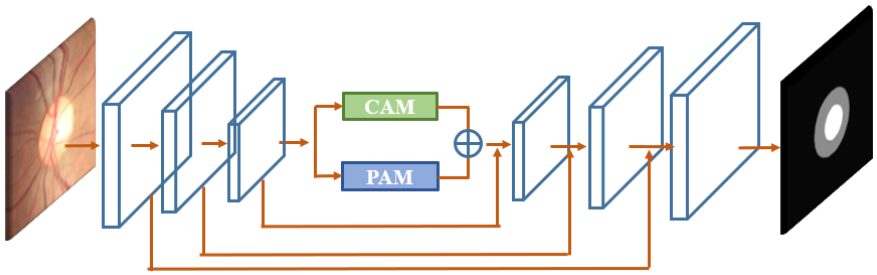

Fig. 2. DA-MobileNet for optic disc and cup segmentation.

Fig.2. is our proposed DA-MobileNet for optic disc and cup segmentation. Here we use the lightweight MobileNet-v2 as the backbone for feature extraction, which has a faster inference speed with a fine result. Connected after the backbone, the decoder part is devised with a channel attention module and a position attention module to capture the global feature dependency in the spatial. In addition, we also introduced the skip connection structure proposed in U-Net, which can fuse the low-level features and make the final output of the network have multi-scale features to obtain higher segmentation accuracy. Our DA-MobileNet is build with the Keras framework.

2.3 Training Detail

We use the optic disc as the center to crop three scale images from the original image, which are 600*600, 800*800, and 1000*1000 as the training set. The cropped image is resized to 256*256 as the network input. Our loss function for the segmentation task uses the weighted sum of focal loss and dice loss. Each mini-batch has 16 images per GPU. We train the model on 1 NVIDIA Tesla P100 GPU for 100 epochs and employ stochastic gradient descent (SGD) for optimizing the deep model. We use a gradually decreasing learning rate starting from 0.01 and a momentum of 0.9. We employ the piecewise constant learning rate policy where the learning rate is multiplied by 0.1 every 30 epoch.

2.4 Results

Table 2. is our segmentation results on the REFUGE2 testing dataset.

Table 2. Segmentation results on the test set.

Model	Dice Cup	Dice Disc	Mae CDR
Baseline	0.8523	0.9569	0.0421
Baseline+ histogram matching	0.8679	0.9615	0.0419

Task 3 Localization of Fovea (macular center)

In this task, 1200 color fundus images with unified manual pixel-wise annotations of the fovea (macular center) were given, varying in scan device and contrast. To tackle with such challenging task, we applied a two-stage solution to detect the fovea. At the first stage, the macular region would be extracted and cropped. And the fovea would be detected in the region for the next step.

3.1 Data preparation

We firstly performed preprocessing to the data sets to ensure the style consistency of training sets and validation set. The mean and variance of each channel among RGB of input image was adjusted to a standard value, in order to solve domain adaptive problem. In addition, we applied online random augmentation including rotation, contrast and brightness shift, blur and sharpen, in

order to increase the generalization of our model.

2.2 Implementation details

The pipeline of our proposed method is shown in Fig 3. YOLOV4 was implemented as our detection model for this task. We compared performance of this detection baseline with different backbones, including CSPDarknet53，EfficientNetB0 and EfficientNetB1, to explore the precision of the model with different feature encoders as well as avoiding deep backbone leading to overfitting because of lacking data. We added smooth L1 loss to fit coordinate offset of bounding box center and the scaling change of width and height together with CIOU loss and thus enabled the model to converge and focus more on location precision of small objects simultaneously.

At first, we resized the input image to 512*512. Considering the fovea as the center coordinate, we cropped a square with side of 100 pixels at current resolution as the ROI of macular region, and set a square of 20 pixel as input bounding box of detection. We resized the macular region ROI to 512*512 as well. We trained a single detection model A to locate macular region. And then trained the second model B on cropped macular region ROI as a refined model to obtain final fovea location. On the validation set, we used model A to acquire macular region, and subsequently located the fovea on the predicted ROI.

During the training procedure, we trained the detection head of the model with the layers of backbone frozen for 50 epochs and then fine-tuned for 50 epochs afterwards with all the layers unfrozen. The learning rate was initialized at 1e-3 and drop 10× at 50 and 80 epochs. The weights of backbone EfficentNetB1 were initialized with ImageNet pretrain and other layers are randomly initialized. Adam was adopted as the optimizer with batch size of 8. Our Model was implemented in Keras API with Tensorflow backend.

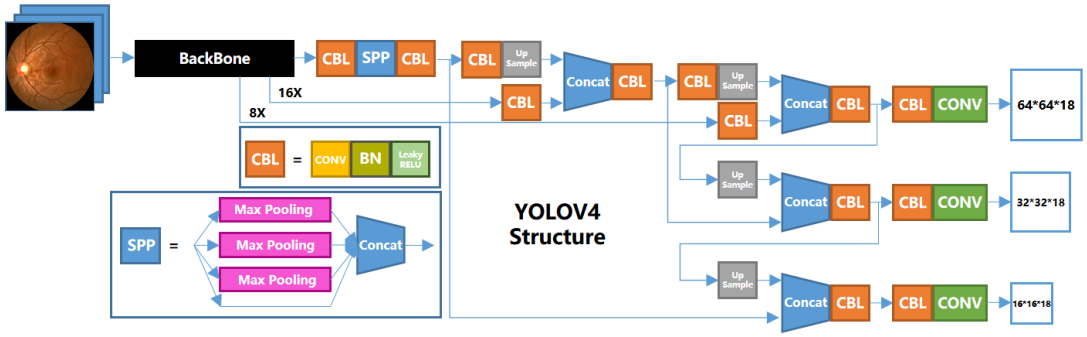

Fig. 3. YOLOV4 for fovea detection.

On the validation set, for missed detection results in Fig, we solved the problems by joint location. i) When fovea missed, but macular region was available, we assumed that the macular ROI is relatively referable, namely regarded the center of macular ROI as the result; ii) If macular region missed, we obtained the location of disc according to the segmentation results of disc. And we estimate the approximate location of macular ROI base on statics counted on training set of their relative position relationship, so as to accomplish cropping and detection; iii) if both macular region and fovea missed, we regarded the estimation of macular ROI center as the final result. Here we applied 0.3 as the threshold to verify whether the prediction is valid. If more than one bounding box which was preserved, localization of several results would be weighted averaged according to corresponding score to obtain better positional precision.

3.3 Results

In this report, we adopted YOLOV4 to locate macular region and the fovea position with 2 stages

on the validation dataset. Experiments showed that the best performance of a single model is YOLOV4 with EfficientNetB0, and the results were further improved by fusing the results of different models using max score. We organized the results as shown in Table 3.

Table 3. Localization results on the test set.

Model	Backbone	Euclidean Distance
YOLOV4	CSPDarknet	10.25
YOLOV4 + style consistency	CSPDarknet	9.44
YOLOV4 + style consistency+smooth L1	CSPDarknet	8.96
YOLOV4 + style consistency+smooth L1	EfficientNet B0	8.76
YOLOV4 + style consistency+smooth L1	EfficientNet B1	8.78
YOLOV4 + style consistency+smooth L1	Fusion	8.63