Optic Disc/Cup Segmentation and Glaucoma Classification from Fundus Images with Fully Convolutional Networks

Zifeng Wu, Chunhua Shen, Anton van den Hengel and Jianpeng Zhang
University of Adelaide
Adelaide, SA 5005, Australia
firstname.lastname@adelaide.edu.au

Abstract

We evaluate the performances of general approaches to image classification and semantic image segmentation on the provided retinal fundus image data. To segment the optic disc and cup from the background in an image, we combine multiple carefully tailored fully-convolution networks. The pipeline consists of the coarse and fine stages, where we first efficiently locate the disc/cup per image using a single network, and then finely segment them by combining a group of more effective networks. To identify glaucoma, we combine two kinds of deep convolution neural networks (CNNs). One is based on the whole images, while the other is based on the local regions around optic discs/cups. Experimental results show that our solution to optic disc/cup segmentation performs promisingly on the first part of testing set, even if images in this set are filmed using a quite different device compared to those in the training set.

1. Optic Disc/Cup Segmentation

In this year's REFUGE (Retinal Fundus Glaucoma Challenge), the segmentation task amounts to identifying all pixels of the optic disc and cup in a fundus image, as shown in Fig. 1. Officially, the quality of segmentation is evaluated via the mean optic cup dice score, the mean optic disc dice score and the mean absolute error of the cup to disc ratio on a test set.

The provided training set is composed of 400 images, taken using a Zeiss Visucam 500 (2124×2056 pixels) camera. Besides, there are 400 images as the validation set, as well as another 400 ones being the test set, which are however taken using a Canon CR-2 (1634×1634 pixels) camera. Within each of the three sets, there are 10% (40) images wherein glaucoma is present. They were first manually labelled by seven independent glaucoma specialists. Thus, there were seven annotations per image, which were merged into a single one by another senior glaucoma specialist.

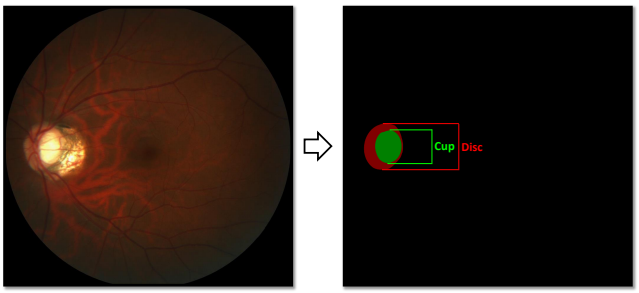

Figure 1. Optic Disc/Cup Segmentation from a fundus image.

1.1. Method

The inference pipeline of our proposed method is illustrated in Fig. 2. The notable points here are threefold. First, we use Model A to efficiently find the coarse locations of the optic disc and cup in an image. This reduces the computational cost by skipping a large background part of the image for costly Models B1 and B2, and applies a sanity check by suppressing all the false positive (disc/cup) pixels in that background part. Second, we apply the multi-scale testing technique, which has shown to be useful in the literature [3]. Last, we combine the locally-normalized input with the original one, in order to relieve the impact of cross-domain testing, considering that the used cameras are different among the training and test sets.

The structures of our used fully convolutional networks (FCNs) in this work is shown in Table 1, wherein ResNet-50 plays the role of a backbone network. In this work, we also use ResNet-101 [1], ResNet-152 [1] and wide ResNet-38 [3] as backbones. However, we omit the detailed structures of these networks, since we can trivially derive them from Table 1.

1.2. Implementation Details

During the competition, annotations on the validation and test sets are not accessible. We thus split the training set into the *train* and *val* sets. Here, the *train* set contains

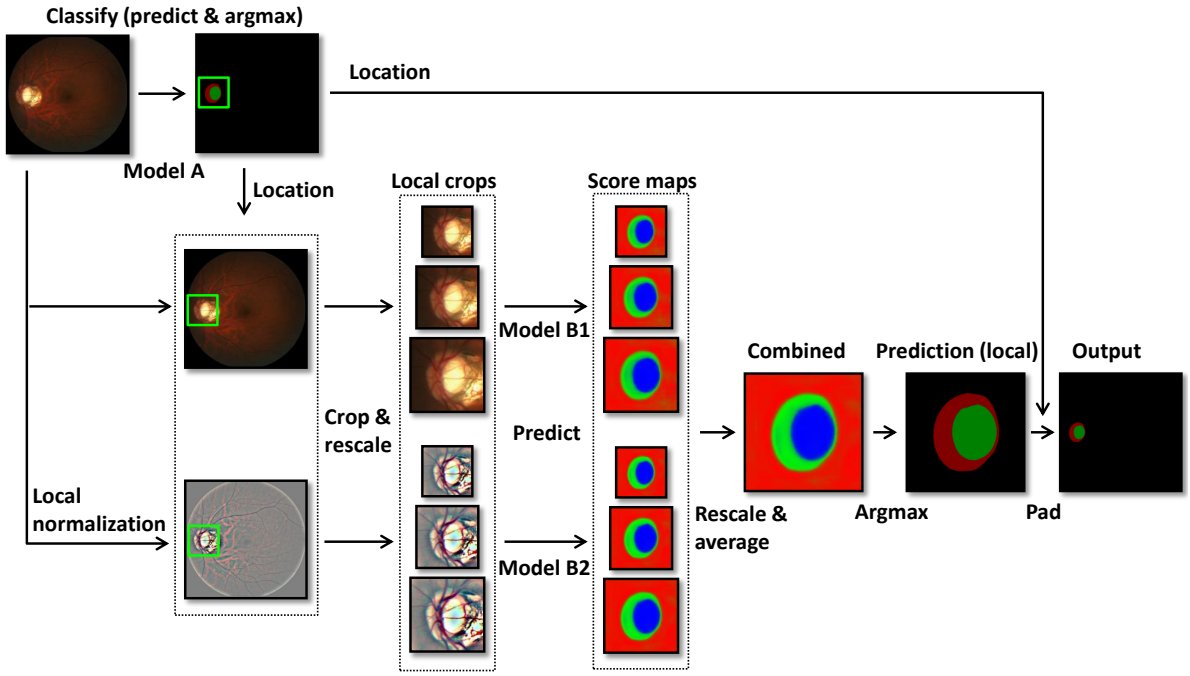

Figure 2. Inference pipeline of our proposed method. We locate the disc and cup using Model A; crop the local region around them; rescale the crop by different factors (0.9, 1.0 and 1.1) to build a pyramid; compute the score maps using Model B1; do the same again for a locally-normalized input using Model B2 to compute another three groups of score maps; combine all the obtained score maps; calculate the prediction (pixel-level category annotations); and finally pad it with background pixels to get the image-level prediction.

Stage	Operator	Channels	Stride	Group	Dilation	#	Connection	Spatial shape
1	Conv: 7	64	2	–	–	1	–	1640
	MaxPool: 3	64	2	–	–	1	–	820
2	Bottleneck	64	1	–	–	1	–	410
	Bottleneck	64	1	–	–	2	–	410
3	Bottleneck	128	2	–	–	1	–	410
	Bottleneck	128	1	–	–	3	–	205
4	Bottleneck	256	1	–	2	1	–	205
	Bottleneck	256	1	–	2	5	–	205
5	Bottleneck	512	1	–	4	1	–	205
	Bottleneck	512	1	–	4	2	–	205
Newly-added layers								
	DilatedConv: 3	128	1	–	6	1	–	205
	Upsampling: 2	128	–	–	–	1	I: 1	205
	Conv: 3	128	1	–	12	1	–	410
	Conv: 1	3	1	–	–	1	–	410

Table 1. Structure of our used FCN, with ResNet-50 as the backbone network. The spatial shapes here are given for Model A, which is applied on a whole image. 'Conv: 7' is a convolution layer with 7×7 kernels. 'MaxPool: 3' is a 3×3 max pooling layer. 'Bottleneck' denotes a classic residual unit [1] composed of three convolution layers, respectively with U 1×1, $4U$ 3×3 and U 1×1 kernels, supposing that U is the number of channels for that bottleneck residual unit. Note that the stride of a bottleneck unit only applies to the first convolution layer. By 'I: 1', we mean adding an in-column connection to the output of this layer, and that the in-column connection is implemented as a 1×1 convolution layer.

32 glaucoma and 288 negative images, while the *val* set has 8 and 72 respectively. We can thus evaluate the hyper-parameters of our method using the *val* set, including learning rates, mini-batch sizes, training epochs and so on.

We initialize all networks with the weights pre-trained using the ImageNet [2] classification data, and tune them using the stochastic gradient decent (SGD) algorithm. For all networks, the learning rate is 0.016, and the mini-batch size is 16. For each image, we rescale it by a factor sampled from $[0.7, 1.3]$, rotate it by an angle sampled from $[-45, 45]$, and then crop a 512×512 region at some random location of the image.

2. Glaucoma Classification

The classification task amounts to identifying the presence of glaucoma given a fundus image. We combine two kinds of classification models in this task. One is trained using the whole fundus images, while the other is trained using the local regions around the discs and cups. We locate the disc and cup per image using Model A as described in the previous section. Here, we fine-tune ResNet-50 [1], ResNet-101 [1], ResNet-152 [1] and wide ResNet-38 [3] using the provided data, and combine them by averaging their predictions.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vision*, 115(3):211–252, 2015.

[3] Z. Wu, C. Shen, and A. van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. arXiv:1611.10080, 2016.