A workflow for computer-aided diagnosis of glaucoma

He Wang[1] **Shuo Li**[1] **Yi Fang**[2]

[1] Peking Union Medical College Hospital
[2]The Fuzong Clinical Medical College of Fujian Medical University

Abstract

Glaucoma is an optic disease with visual impairment, which even leads to blindness. The timely diagnosis of glaucoma can reserve the blinding process with a good prognosis. In this paper, we present a workflow to automatically diagnose glaucoma and distinguished fundus images with low qualities. The performance in the hidden test set shows the promising clinical usefulness against experts.

Correspondence: wanghe_15@126.com

Introduction

Glaucoma is a group of eye diseases often characterized by elevated intraocular pressure. Optic nerve injury results in progressive loss of retinal ganglion cell axons, which initially manifests as visual field defect and, if untreated, eventually progresses to irreversible blindness.[1]

To facilitate clinicians with rapid and accurate computational aided diagnosis, the AIROGS challenge was held to exploit solutions based on the color fundus photograph.

The AIROGS challenge

The Rotterdam EyePACS AIROGS dataset [2] (in full, so including train and test) contains 113,893 color fundus images from 60,357 subjects and approximately 500 different sites with a heterogeneous ethnicity. All images were assigned by human experts with the labels referable glaucoma, no referable glaucoma, or ungradable.

The provided training set contains approximately 102,000 gradable images. The hidden test set contains about 11,000 gradable and ungradable images (both gradable and ungradable).

For each input image during evaluation, the desired output is a likelihood score for referable glaucoma (O1), a binary decision on referable glaucoma presence (O2), a binary decision on whether an image is ungradable (O3, true if ungradable, false if gradable), and a non-thresholded scalar value that is positively correlated with the likelihood for ungradability (e.g. the entropy of a probability vector produced by a machine learning model or the variance of an ensemble) (O4).

The screening performance will be evaluated using the partial area under the receiver operator characteristic curve (90-100% specificity) for referable glaucoma (α) and sensitivity at 95% specificity (β). The screening performance metrics are based on these specificity ranges, since a high specificity is generally desired in screening settings. To calculate α and β, we compare output O1 to the referable glaucoma reference provided by human experts.

Using Cohen's kappa score, the agreement between the reference and the decisions provided by the challenge participants on image gradability, O3, is calculated (γ). Furthermore, the area under the receiver operator characteristic curve will be determined using the human reference for ungradability as the true labels and the ungradability scalar values provided by the participants, O4, as the target scores (δ).

Diagnosis workflow of glaucoma

Manual annotations

From clinical knowledge, glaucoma can be diagnosed by the ratio of optic disk/cup. To force the models' attention to the optic disk area, we manually annotated several images to train a not-so-smart segmentation model.

- We annotated optic disks on 40 images.
- For the convenience of annotation of retina vessel, we centercropped and resized 40 images to 512x512. The rough annotations were made on those images.
- For the task of detection of ungradable images, we manually selected 100 images with relatively poor image qualities from the provided training dataset.

Basing model training

- Optic disk segmentation model (M_{disk}), we used resnet101-upernet[3] as the backbone. The batch size was set to 4 with 100 iterations in each epoch. A total of 100 epochs with an exponential learning rate scheduler was applied to the training process.
- Retina Vessel segmentation model (M_{vessel}), we used UNet[4] as the backbone. We trained the model with a total of 10000 iterations with a batch size of 4.
- Both segmentation models above were based on only 40 images (covered only 0.4% of the provided training dataset), which led to a not-so-satisfying performance. And it may fail to detect the disk caused by the variety of image qualities. Therefore, we did not directly apply the segmentation result to the diagnosis model.

Crop

Case 1: the disk was successfully detected

- We used the output of M_{disk} to roughly compute the location and diameter of the optic disk.
- We cropped the image in the disk location with the image size of 3 times of disk diameters.
- We resized the image to 384 x 384 as the input (I_{disk}) of the diagnosis model.

Case 2: the disk was not detected

- We centercropped the image to remove its black edge and resized the image to 384 x 384 as the input (I_{center}).

Establishment of the diagnosis model

We adopt the vision transformer (Vit-base with the image size of 384) as the classifier backbone.

The training methods were using a pretraining method described as mae[5] (masked auto-encoder).

Case 1: the disk was successfully detected

- We toke all I_{disk} as input to train the VIT model named V_{disk}

Case 2: the disk was not detected

- As the provided image dataset was of high quality, only 200 images failed to be detected. Therefore, we centercropped and resized all images as the input to train the VIT model named V_{center}

Establishment of the ungradable model

Here, we defined the image as gradable when the disk could be seen in the image with a clear view of the retina vessel.

As we have trained two models (M_{disk} and M_{vessel}) to segment the disk and vessel, the only thing need to do was to judge whether the retina vessel could be clearly seen. So we trained a resnet model, named R_{vessel} based on the segmentation maps.

The inputs were the vessel segmentation of the first 500 images in the trained dataset as well as manually selecting 100 ungradable images. The output of R_{vessel} was O_{vessel}.

The ungradable likelihood was calculated:

$$O_4 = O_{vessel} + 0.75 \times O_{disk}$$

$$O_{disk} = \begin{cases} 1, & if\ optic\ disk\ not\ be\ detected \\ 0, & if\ optic\ disk\ was\ detected \end{cases}$$

$$O_3 = \begin{cases} 1, & O_4 > 0.95 \\ 0, & O_4 \leq 0.95 \end{cases}$$

Test workflow

- An image was first run through the M_{disk} and M_{vessel} and the results is X_{disk} and X_{vessel}.
- Try to automatically calculate the diameter and location of the optic disk. If error returns O_{disk} as 1, else returns O_{disk} as 0.
- If Disk was successfully detected. The original Image was cropped to the disk neighborhood and fed to the model V_{disk} and output O_1
- If Disk was not detected. The original image was centercropped and fed to the model I_{center} and Output O_1.
- X_{vessel} was put into the R_{vessel} and Output O_{vessel}.
- Calculate O_3 and O_4.

Results

All training processes split 80% data as training dataset and 20% data as validation dataset.

Table 1. performance on the preliminary test phase

	pAUC	TPR@95	κ	gAUC
Workflow	0.8948	0.8688	0.6675	0.9589

Our method showed high performance on the clinical task as well as promising distinguishing abilities of images of poor qualities.

Reference

[1] Weinreb, R.N., et al., *Primary open-angle glaucoma.* 2004. **363**(9422): p. 1711

[2] "Rotterdam eyepacs airogs train set." [Online]. Available:https://doi.org/10.5281/zenodo.5745363

[3] Xiao T, Liu Y, Zhou B, et al. Unified perceptual parsing for scene understanding[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 418-434.

[4]Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[5] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[J]. arXiv preprint arXiv:2111.06377, 2021.