Deep Learning For Referable Glaucoma Screening and Out-of-Distribution Detection

Zekang Yang[1], **Hong Liu**[1], **and Zihao Shang**[1]

[1]Institute of Computing Technology

Abstract

This paper presents a solution for Artificial Intelligence for RObust Glaucoma Screening (AIROGS) Challenge. We utilized ResNet50 to train a model to diagnose the referable glaucoma using color fundus images, and also give the ungradable probability when the images' quality is poor.

Introduction

Glaucoma is the second leading cause of blindness in the world. Early detection of glaucoma can lead to timely treatment. Glaucoma patients are mainly diagnosed by specialist looking at the fundus images. Once the fundus examination increases, the workload of specialists also increases, which may result in misdiagnosis or missed diagnosis. Automatic diagnosis of glaucoma by computer aided technologies can reduce the workload of specialists, which has great significance to promote largescale fundus screening. Recent years, artificial intelligence technology has been applied to aid in the diagnosis of glaucoma and achieves good performance at-the-lab. However, in real-world application, the performance of artificial intelligence will deteriorate due to the existence of out-of-distribution data, such as bad quality images. We construct an AI model to diagnose glaucoma and also give the confidence of the diagnosis.

The AIROS challenge

In this section, we briefly describe the AIROS challenge [1] including dataset, task and evaluation metrics. The Rotterdam EyePACS AIROGS datasets contains 113,893 color fundus images from 60,357 subjects. The training data set contains approximately 102,000 gradable images (3,270 referable glaucoma, and 98,172 no referable glaucoma). The test data set contains about 11,000 gradable and ungradable images.For each input image during testing and evaluation, the desired outputs include 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 (O4).

The evaluation will be based on two aspects including screening performance and the robustness. 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). 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).Finally, all participants will be ranked on the individual metrics α, β, γ and δ, resulting in rankings Rα, Rβ, Rγ and Rδ, respectively. The final score will be calculated as follows:

$$S_{final} = \frac{R_\alpha + R_\beta + R_\gamma + R_\delta}{4} \qquad \textbf{(1)}$$

The final ranking is subsequently based on Sfinal, where a lower value for Sfinal will result in a higher ranking.

Solution

Our solution contains four components: data preprocessing, data augmentation, screening referable glaucoma and out-of-distribution detection.

Data preprocessing. Because the image resolution is large and contains redundant information, we preprocess the input image first. We discard the black edges to crop out the eye position and reduce the width to 512 in proportion to the length and width.

Data Augmentation. In the training stage, we use RandomAugment[2] for data augmentation. RandomAugment is a out-of-box data augmentation measure with two hyperparameters, and applies a sequence of random image transformations.We set the number of transformations per image to be 2 and the magnitude of transformations to be 10. During the inference phase, we random crop five 512x512 crops from input image, then input all crops separately into the model and get five scores. If the maximum of five scores is greater than 0.9, we let it be the output of the model, otherwise we take the mean of the five scores as the output of the model.

Screening referable glaucoma. We trained a model utilizing ResNet50[3] to classify glaucoma and non-glaucoma.We use Adam[4] as optimizer. We use StepLR as the learning rate decay strategy, and set the initial learning rate of 1e-4 and gamma of 0.98. We also use weight decay of 5e-4 as the regularization method. In the inference stage, the model will give a score ∈ [0,1] for the probability of glaucoma. When

Table 1. Performance of the our method on our test data and the online test data.

Test Data	pAUC	TPR@95	kappa	gAUC
our test data	0.9043	0.8628	-	-
online test data	0.8542	0.7875	0.5073	0.8994

the score is greater than 0.5, we diagnose glaucoma, otherwise non-glaucoma.

Out-of-distribution detection. Due to the existence of out-of-distribution data is difficult to diagnose glaucoma or non-glaucoma. The score of out-of-distribution data is far from 0 and 1. We use the difference between the score and 0 (when the score is lower than 0.5) or 1 (when the score is greater than 0.5) to represent the probability of ungradable. And when the probability of ungradable is greater than 0.1, we think it is ungradable, otherwise gradable.

Experimental Results

We divided all training data into training set, validation set and test set according to 7:1:2. We trained our model on the training set, selected the model with the lowest loss value in the validation set, and finally tested the generalization ability of our model on the test set. The results of our model is summarized in the Table 1 .

Conclusion

For AIROGS challenge, we propose a solution that uses fundus color images to give the probability of referable glaucoma and ungradable. At present, we only use a simple convolutional neural network, and we hope to introduce attention mechanism or contrastive learning to improve feature expression ability in the future.

References

1. Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, Bram van Ginneken, Hans G. Lemij, and Clara I. Sánchez. Rotterdam eyepacs airogs train set, December 2021. The previous version was split into two records. This new version contains all data and the second record is deprecated.
2. Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.
3. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
4. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.