A Good Closed-Set Classifier is all you need for the AIROGS Challenge

Jónathan Heras[1][0000−0003−4775−1306], Didac Royo[2], and Miguel Ángel Zapata[2,3]

[1] Department of Mathematics and Computer Science, University of La Rioja
[2] UPRetina
[3] Hospital Vall D'Hebron

Abstract. Glaucoma is an optic disease that leads to blindness, but this might be avoided with an early diagnosis thanks to a screening test. The AIROGS Challenge was organised to develop solutions for glaucoma screening from retinal fundus images that are robust to real-world scenarios. In this work, we present our solution to this challenge based on ideas from open-set recognition tasks. Namely, we build a strong deep classification model for diagnosing referable glaucoma using the latest techniques from the image recognition literature, and use such a model for deciding whether an image is ungradable. As shown in the challenge leaderboard, this approach provides good results for glaucoma screening performance and robustness.

Keywords: AIROGS · Glaucoma · Out-of-Distribution · Retinal Fundus · Image Classification

1 Introduction

Glaucoma is a progressive optic neuropathy, which is the leading cause of blindness in industrialised countries [4]. General screening for glaucoma in the population is ideal, as an early detection of glaucoma can avoid visual impairment. However, manual screenings are not possible when the population is large, or in low- and middle-income countries. Therefore, it is necessary the development of computer-based tools that automatise the screening process.

Artificial intelligence and, more specifically, deep learning methods, can be helpful for glaucoma screening from colour fundus images, by reducing the need for manual labour. However, the performance of deep learning models drops when applied to real-world settings due to unexpected out-of-distribution data and bad quality images. In this context, the AIROGS challenge was organised to develop solutions for glaucoma screening from retinal fundus images that are robust to real-world scenarios [9].

In this paper, we present our approach to the AIROGS challenge based on ideas from open-set recognition; namely, we adapt the method presented by Vaze et al. [8]. All the code and models developed for this work are available at https://github.com/joheras/airogs

2 The AIROGS challenge

We provide a brief overview of the AIROGS challenge, a complete description can be found in the challenge webpage. The AIROGS dataset contains 113,893 color fundus images that were labeled by experts as referable glaucoma, no referable glaucoma, or ungradable. The training set contains approximately 102,000 gradable images (98,172 non referable images, and 3270 referable images), and the test set contains about 11,000 gradable and ungradable images.

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), and a non-thresholded scalar value that is positively correlated with the likelihood for ungradability (O4). Such an output is used to evaluate both screening performance and robustness. The screening performance is evaluated using the partial area under the receiver operator characteristic curve (90-100% specificity) for referable glaucoma and sensitivity at 95% specificity. To measure robustness, Cohen's kappa score is used to calculate the agreement between the reference and the decisions provided by the challenge participants on image gradability. Furthermore, the area under the receiver operator characteristic curve is determined using the human reference for ungradability as the true labels and the ungradability scalar values provided by the participants as the target scores.

3 A Good Close-Set Classifier

Following the approach presented in [8], we used several techniques from the image recognition literature to produce image classification models for glaucoma screening that can also be used for deciding whether an image is ungradable. Our experiments have been implemented in Python and conducted thanks to the functionality of the libraries FastAI [3] and Timm [10] using a GPU Nvidia RTX 2080 Ti.

As a first step, we split the training dataset into three stratified sets using a 70% of the images for training (70,683 non referable images and 2,354 referable), a 10% for validation (7,854 non referable images and 262 referable), and a 20% for testing (19,635 non referable images and 654 referable). Subsequently, we preprocessed the images from the three sets with the algorithm introduced in [11]. This algorithm removes the black wrapper around the colour images, standardise their colour, and resize them to size 512×512. Subsequently, those images were used for training several classification models.

We have taken two deep learning architectures (ResnetRS50 [1] and EfficientNet v2 [7]) and trained them for 400 epochs — the best weights obtained in the training process regarding the validation set were stored. In order to speed up the training process and deal with the imbalance dataset, in each epoch we took the 2,354 referable images from the training set and 2,354 random non referable images from the same set. For training, we used the transfer-learning method presented in [3], and used progressive resizing [3] — we trained the models for

100 epochs with images of size 128 × 128, then 100 epochs with images of size 224 × 224, subsequently 100 epochs with images of size 384 × 384, and finally 100 epochs with images of size 512 × 512). In addition, we replaced the traditional cross entropy loss with the focal loss [5] and used Randaugment [2] as data augmentation method — the magnitude of the transformations was increased with the size of the images during the training process.

At inference time, we apply test-time augmentation [6] to obtain 5 predictions that are averaged to produce the output for the challenge. From a given image, the output of test-time augmentation is a likelihood for the image to be referable, l_r, and a likelihood for non referable, l_{nr} — these likelihoods adds to 1. Using those likelihood values, we produce the output for the challenge as:

- O1. We return the likelihood score for referable glaucoma is l_r.
- O2. We return True if $l_r > 0.5$, and False otherwise.
- O3. We return True if either $l_r > 0.8$ or $l_{nr} > 0.8$, and False otherwise.
- O4. We return the maximum value between l_r and l_{nr}.

4 Results and discussion

The results obtained by our models in the preliminary test phase 2 are provided in Table 1. The best results for all the metrics are obtained using the ResnetRS50 architecture, and this is the model submitted to the final test phase.

Table 1: Results of the trained models in Test Phase 2

Method	Ungradability AUC	Ungradability κ	Sens@95Spec	AUC-90Spec
ResnetRS50	0.8910	0.4598	0.7937	0.8526
EfficientNet v2	0.8566	0.3687	0.7875	0.8469

In addition to the models submitted to the challenge, we also tested other architectures like ConvNext, and the ensemble of models and test-time augmentation prediction that obtain good results in our test set. However, there is a time-limit for the submissions to the challenge and those methods were too slow.

References

1. Bello, I., Fedus, W., Du, X., Cubuk, E.D., Srinivas, A., Lin, T.Y., Shlens, J., Zoph, B.: Revisiting resnets: Improved training and scaling strategies. arXiv preprint arXiv:2103.07579 (2021)
2. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020)

3. Howard, J., Gugger, S.: Fastai: A layered api for deep learning. Information **11**, 108 (02 2020)
4. Leggio, G.M., Bucolo, C., Platania, C.B.M., Salomone, S., Drago, F.: Current drug treatments targeting dopamine d3 receptor. Pharmacology & therapeutics **165**, 164–177 (2016)
5. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
7. Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training. arXiv preprint arXiv:2104.00298 (2021)
8. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need. In: International Conference on Learning Representations (2021)
9. de Vente, C., Vermeer, K.A., Jaccard, N., van Ginneken, B., Lemij, H.G., Sánchez, C.I.: Rotterdam eyepacs airogs train set (Dec 2021). https://doi.org/10.5281/zenodo.5793241, https://doi.org/10.5281/zenodo.5793241
10. Wightman, R., et al.: Pytorch image models (2021), https://github.com/rwightman/pytorch-image-models
11. Zapata, M.A., Royo-Fibla, D., Font, O., Vela, J.I., Marcantonio, I., Moya-Sánchez, E.U., Sánchez-Pérez, A., Garcia-Gasulla, D., Cortés, U., Ayguadé, E., et al.: Artificial intelligence to identify retinal fundus images, quality validation, laterality evaluation, macular degeneration, and suspected glaucoma. Clinical Ophthalmology (Auckland, NZ) **14**, 419 (2020)