Open-Set Glaucoma Screening from Eye Fundus Images: Domain Knowledge to the Rescue

Adrian Galdran[1,2]*, Gustavo Carneiro[1], Miguel A. Gonzlez Ballester[1]

[1] Universitat Pompeu Fabra, Barcelona, Spain [2]University of Adelaide, Adelaide, Australia
[3]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

Abstract

Detecting early signs of glaucoma can avoid visual impairment in the general population, and this goal could be approached through the examination of routinely acquired retinal color fundus in general screening programs. In a screening scenario, the amount of data that needs to be reviewed manually by ophthalmic experts is massive, and effective machine learning tools for effective glaucoma detection would provide great clinical value, enhancing the cost-effectiveness of glaucoma screening, by decreasing the amount of manual labor required. Unfortunately, the unpredictable behavior of modern neural networks on samples that do not come from the same distribution as the training data can result in relevant and unforeseen performance deterioration. Such out-of-distribution/open-set data, e.g. ungradable or bad quality images, needs to be flagged in test time, but it might not be available during training. This short manuscript provides a description of our solution to this challenge in the context of the AIROGS: Artificial Intelligence for RObust Glaucoma Screening Challenge. We compare two approaches, namely: 1) directly measuring the confidence of a glaucoma classifier in terms of the maximum probability it produces, a popular and generalistic Open-Set recognition technique, and 2) synthetic generation of Open-Set data based on Domain Knowledge in order to train an auxiliary model for performing Open Set Recognition.

1. Introduction

Early glaucoma detection can prevent visual impairment, and screening for this disease can have a great impact in the general population. For this reason, this task has attracted much attention in the computerized medical image analysis community in recent years, see [4], or a recent review in [2]. However, in a real scenario, atypical data that comes from a distribution not matching the data used for training a model can break a model and result in serious misdiagnosis. Therefore, techniques that can deal with this situation, like Out of Distribution (OoD) detection or Open Set Recognition (OSR) algorithms, hold great promise in this context.

Note that we do make a small difference between OoD detection and OSR in this paper. We consider the OoD task as rejecting in test time samples that do not belong to the training data distribution,, but without addressing any kind of classification problem. An example of OoD would be training an autoencoder as a one-class classification algorithm, in which we would expect OoD samples to incur in larger reconstruction errors than in-distribution data. On the other hand, OSR would be the task of jointly performing multi-class classification on in-distribution data and OoD detection. In this case, we refer to the classes used for training as the Closed Set, and the categories to which the OoD data belongs conform the Open Set.

In this paper, we present the details of our participation in the Artificial Intelligence for RObust Glaucoma Screening Challenge (AIROGS challenge) [1]. The proposed task was to train a model to perform referable glaucoma detection and simultaneously discard OoD data that would be presented to the algorithm in test time. The organization specified that OoD data would amount in this context to ungradable images, *i.e.* images for which an expert ophthalmologist decided there was not enough information to formulate a diagnosis. No further information on the visual aspect of ungradability, nor access to ungradable examples, were provided to the participants. In addition, employing extra fundus images, or models pretrained on external fundus images, was prohibited by the organization. More information on the dataset construction, challenge evaluation, and public leaderboards can be found at https://airogs.grand-challenge.org/.

*Corresponding author: adrian.galdran@upf.edu

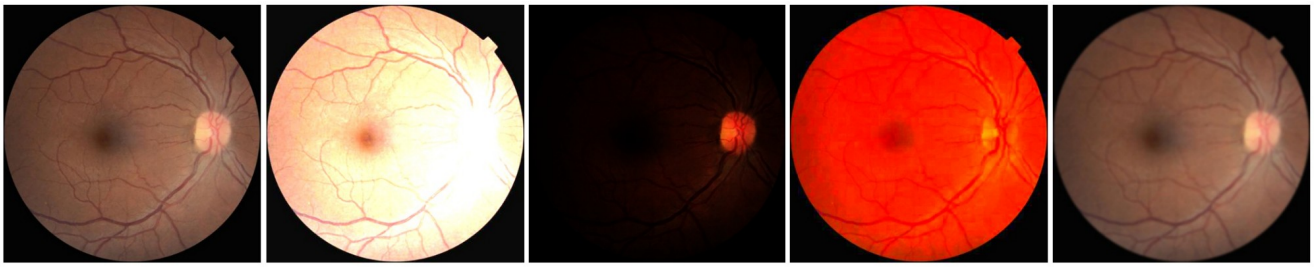

Figure 1. Original training image and synthetically degraded versions of it, degradations shown left to right: {`Brightness`, `Gamma`, `Saturation`, `Blur`}. Leveraging domain knowledge on the visual appearance of typical retinal image degradations, we construct a synthetic training set for learning to detect low-quality images in test-time without the need of labeled data.

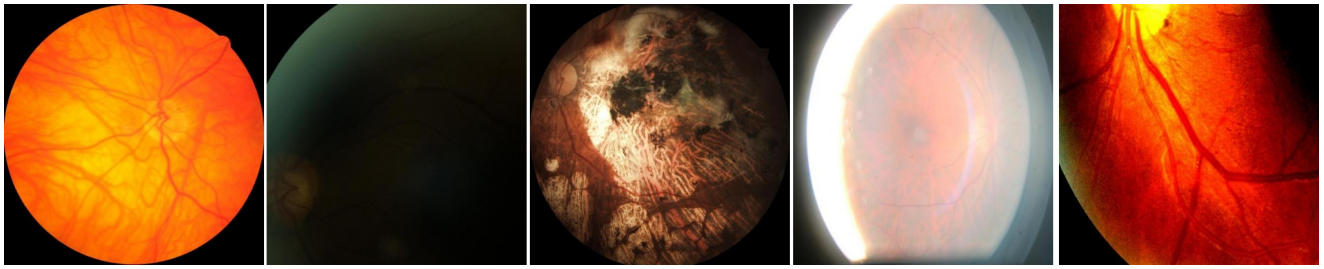

Figure 2. Images selected as ungradable by our classifier, which trained on synthetic degradations. Note that these are samples from the training set and are therefore supposed to be gradable.

2. Generic Open Set Recognition versus Domain Knowledge

Below we give the details of the two Open Set Recognition approaches we submitted to the AIROGS challenge.

2.1. Open Set Recognition - Max over Softmax as a strong baseline

In an OSR problem, we begin with a labeled training set $\mathcal{C}_{\text{train}}$ that contains examples from N known classes $\mathcal{K} = \{k_1, ..., k_N\}$, which compose the *Closed Set*. In test time samples, from an *Open Set* $\mathcal{O}_{\text{test}}$ can appear. This set is composed of examples from M unknown categories $\mathcal{U} = \{u_1, ..., u_M\}$ not seen by the model during training, *i.e.* $\mathcal{D}_{\text{test}} = \mathcal{C}_{\text{test}} \cup \mathcal{O}_{\text{test}}$. The purpose of an Open Set algorithm is to accurately classify test samples from $\mathcal{C}_{\text{test}}$ and at the same time abstaining from formulating a prediction on samples from $\mathcal{O}_{\text{test}}$.

It has been recently shown in [5] that a widely popular baseline OSR method can achieve state-of-the-art results, if it is adequately tuned. One can train a CNN U_θ by minimizing the cross-entropy loss between one-hot labels y and softmax probabilities $p_\theta(y|x)$ for $x \in \mathcal{C}_{\text{train}}$, and then define an OSR score as the Maximum Softmax Probability (*MSP*) $S(y \in \mathcal{C}_{\text{test}}|x) = \max_{y \in \mathcal{C}} p_\theta(y|x)$. If we assumes that U_θ distributes probabilities with high entropy for unknown classes, resulting in a low $S(y \in \mathcal{O}_{\text{test}}|x)$ value, this provides a robust OSR technique.

2.2. Domain-Knowledge OSR

The above approach was introduced for general OoD and can be implemented for any classifier that produces a vector of probabilities, on visual and non-visual domains. However, the price to pay for generality is several shortcomings. For example, *MSP* relies heavily on the correct calibration of the underlying predictive model, and modern over-parametrized neural networks are known to suffer when calibration is measured under domain shift. Also, it is not clear how other factors, like a small number of categories or the presence of class imbalance, impact the performance of this method. Such factors are typical of medical image analysis problems, but are seldom considered in general computer vision benchmarks. Finally, when using *MSP* there is no leverage of domain knowledge that can help it reach better OoD. For this challenge, we intend to compare *MSP* with a different Domain-Knowledge based approach, that we describe next.

In principle *MSP* is a techinque suitable for detecting any kind of OoD data. However, in the AIROGS challenge, we are given the information that Open Set data consists indeed of glaucoma-ungradable fundus images, and this reduces substantially the extent of data we can encounter in test time, since low-quality retinal images have been studied thoroughly in the literature [3]. We therefore proceed to train a new classifier that can tell apart the original AIROGS training set from a randomly degraded version of it. For

that, we follow the same training protocol as above, but in this case each time an image is sampled we apply a degradation with a probability of $p = 0.5$, and also sample the parameters that define each degrading operator from a uniform distribution defined over a given interval.

Such degradation is randomly picked from a set of four image processing operation that we expect to render the retinal fundus ungradable. To that end, we define parameter ranges of our degradations so that they produce extreme visual results that we expect to destroy visual cues that enable diagnosis. An example of these four operations is shown in Fig. 1. During training, the model learns to predict whether the image has undergone a degradation. When training is over, we use the validation set to define a suitable threshold for declaring an image as ungradable. Ideally, our model in this case would classify each image in the original training set as gradable/undegraded, but we notice that the AIROGS dataset is quite noisy and images of extremely poor quality are included as graded. For that reason, we apply our trained degradation model on the training set and select a probabilistic threshold that classifies 0.1% of the images as ungradable. Some examples of training images that are deemed ungradable by our model are shown in Fig. 2.

A great advantage of *MSP* is that it can in principle deal with any kind of OoD data. However, in the AIROGS challenge, we are given the information that Open Set data consists indeed of glaucoma-ungradable fundus images, and this reduces substantially the extent of data we can encounter in test time, since low-quality retinal images have been studied thoroughly in the literature [3]. We therefore proceed to train a new classifier that can tell apart the original AIROGS training set from a randomly degraded version of it. For that, we follow the same training protocol as above, but in this case each time an image is sampled we apply a degradation with a probability of $p = 0.5$, and also sample the parameters that define each degrading operator from a uniform distribution defined over a given interval.

3. Experimental Preliminary Results

Our models[1] were trained on the provided AIROGS dataset [1], with around 102,000 gradable images. The test set, hidden to the participants, contained about 11,000 images, which could be both gradable and ungradable.

The evaluation was be based on both screening performance and OoD detection. Screening accuracy was assessed by means of the partial Area Under the receiver operator characteristic Curve (pAUC), which covers a 90-100% specificity range, for referable glaucoma (α) and sensitivity at 95% specificity (β). OoD detection was evaluated in terms of Cohen's kappa score (γ) between the binary decisions generated by the system and expert labels, as well as the AUC computed from ungradability labels and the submitted ungradability soft probabilities (δ).

[1]Training details, together with code, data and pretrained weights to reproduce our results are available at github.com/agaldran/airogs

Table 1. Performance on Closed Set and Open Set tasks on the preliminary test phase of the AIROGS challenge

	Closed Set		**Open Set**	
	pAUC α	**Spec** β	**Kappa** γ	**AUC** δ
MSP	—	—	—	—
SynthDK	—	—	—	—

There was a preliminary test phase, in which three submissions were allowed, and the challenge platform computed submission performance on a reduced test set. A final test phase in which performance would be derived from the entire test set was held afterwards, but at the time of writing only results of the preliminary phase were available. We compare the performance of Maximum over Softmax Probabilities method (**MSP**) and Synthetic Degradations based on Domain Knowledge (**SynthDK**) in Table 1.

4. Discussion

At the time of writing we do not know yet the performance of our approach on the larger test set of the last challenge phase. This section will be updated once we learn about the final results.

References

[1] Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, Bram van Ginneken, Hans G. Lemij, and Clara I. Snchez. Rotterdam EyePACS AIROGS train set, Dec. 2021. Type: dataset. 1, 3

[2] Yuki Hagiwara, Joel En Wei Koh, Jen Hong Tan, Sulatha V. Bhandary, Augustinus Laude, Edward J. Ciaccio, Louis Tong, and U. Rajendra Acharya. Computer-aided diagnosis of glaucoma using fundus images: A review. *Computer Methods and Programs in Biomedicine*, 165:1–12, Oct. 2018. 1

[3] Ziyi Shen, Huazhu Fu, Jianbing Shen, and Ling Shao. Modeling and Enhancing Low-Quality Retinal Fundus Images. *IEEE Transactions on Medical Imaging*, 40(3):996–1006, Mar. 2021. Conference Name: IEEE Transactions on Medical Imaging. 2, 3

[4] Naoto Shibata, Masaki Tanito, Keita Mitsuhashi, Yuri Fujino, Masato Matsuura, Hiroshi Murata, and Ryo Asaoka. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Scientific Reports*, 8(1):14665, Oct. 2018. 1

[5] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-Set Recognition: A Good Closed-Set Classifier is All You Need. In *International Conference on Learning Representations*, Apr. 2022. 2