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

Adrian Galdran[1,2]*, Gustavo Carneiro[2], 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 screening programs. In a screening scenario, the amount of data to be reviewed manually by ophthalmic experts is massive, and efficient 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 unexpected performance deterioration. Such out-of-distribution/open-set data needs to be flagged in test time, but it is usually not be available during training. This short manuscript describes our solution to this task 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 to perform 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/.

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.

*Corresponding author: adrian.galdran@upf.edu

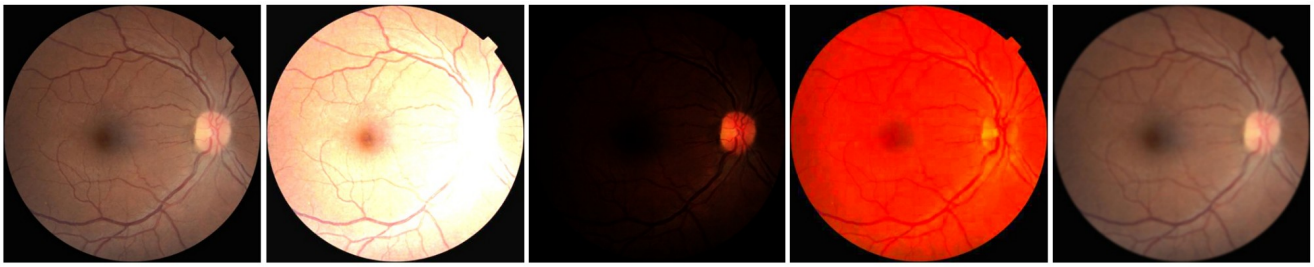

Figure 1. Original training image and synthetically degraded versions of it, degradations shown left to right: {`Brightness`, `Gamma`, `Saturation`, `Blur`}. Leveraging domain knowledge on the visual appearance of typical retinal image degradations, we construct a synthetic training set for learning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

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 aspects, like a small number of categories or class imbalance, impact the performance of this method. These factors are typical of medical image analysis problems, but are seldom considered in 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 technique 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. The 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. In test time, we simply generate from our model a probability of being OoD, and post-process it as outlined in the next section. In the remaining, we refer to this approach as *SynDK*.

2.3. Fixing a Decision Threshold for OoD

For both the *MSP* and *SynDK* techniques, the OoD probability is hard to interpret. In the former case, the OoD score lies in the $[0.5, 1]$ interval, and it is not clear what would be a reasonable threshold to make a binary decision. As for *SynDK*, in principle all images in the training set would be considered as gradable/undegraded by the model, but we notice that the AIROGS dataset is quite noisy and images of extremely poor quality are included as graded. This is visually verified in Fig. 2, where some examples of training images that are deemed gradable by our model using a threshold of $t_{\text{ungrad}} = 0.5$ are shown.

For this reason, in both cases, after training we use the

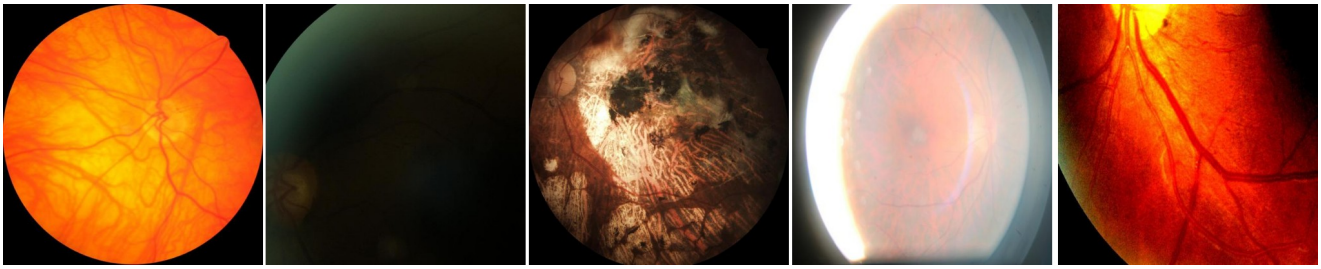

Figure 2. Images selected among the most ungradable by our classifier, which was trained on synthetic degradations. Note that these are samples from the training set and are therefore supposed to be gradable.

validation set to define a suitable threshold for declaring an image as ungradable. We apply our trained OoD models on the validation set and select the probabilistic threshold that classifies 0.1% of the validation images as ungradable, assuming that some mistakes have been made by annotators. We find this threshold to be $t_{\text{ungrad}} = 0.966$ for *MSP* and $t_{\text{ungrad}} = 0.017$ for *SynDK*. All images shown in Fig. 2 are considered ungradable by *SynDK* after fixing the decision boundary at $t_{\text{ungrad}} = 0.017$.

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).

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.

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	**Sns at Spc** β	**Kappa** γ	**AUC** δ
MSP	88.62	84.37	82.76	-1.19
SynthDK	88.62	84.37	83.54	32.22

4. Discussion

We ended up submitting a model similar to *SynthDK* but replaced batch-norm with instance-norm at test time. 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.

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

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

[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