Deep Dirichlet uncertainty for unsupervised out-of-distribution detection of eye fundus photographs in glaucoma screening

Teresa Araújo[1,*], **Guilherme Aresta**[1,*], **and Hrvoje Bogunović**[1]

[1]Christian Doppler Laboratory for Artificial Intelligence in Retina, Department of Ophthalmology and Optometry, Medical University of Vienna, Austria
[*]Equal contribution

The development of automatic tools for early glaucoma diagnosis with color fundus photographs can significantly reduce the impact of this disease. However, current state-of-the-art solutions are not robust to real-world scenarios, providing overconfident predictions for out-of-distribution cases. With this in mind, we propose a model based on the Dirichlet distribution that allows to obtain class-wise probabilities together with an uncertainty estimation without exposure to out-of-distribution cases. We demonstrate our approach on the AIROGS challenge. At the start of the final test phase (8 Feb. 2022), our method had the highest average score among all submissions.

Correspondence: ***teresa.safinisterraaraujo@meduniwien.ac.at***

Introduction

The impact of glaucoma, one of the leading causes of blindness, can be significantly reduced if diagnosed early. Automatic systems can improve the success of screening programs by reducing the workload of specialists. However, current state-of-the-art-systems are usually not robust in real-world scenarios, providing over-confident predictions with out-of-distribution (OOD) cases. With this in mind, we propose an uncertainty-aware deep network that predicts a Dirichlet distribution on the class probabilities. During inference, this type of approach allows to obtain class-wise probabilities together with a sample-wise uncertainty $\in [0 1]$ of that same classification, and has already proven successful for uncertainty estimation in other tasks [1]. Additionally, to fully automate OOD detection, we exploit the assumption that referable glaucoma detection is only possible if the region of the optic disc (OD) has sufficient quality for diagnosis, since the primary manifestations of glaucoma occur there. This introduces an additional challenge, as the network has to additionally provide, without supervision, the location of the OD.

AIROGS challenge This paper describes our submission to the Artificial Intelligence for RObust Glaucoma Screening Challenge (AIROGS challenge) [2]. The main task was to develop an automatic method for *referable glaucoma* detection in eye fundus image. Additionally, the system should provide a soft and binary decision on whether each image isn't diagnosable (*ungradable*), i.e. automatically identify OOD samples and bad quality images. No definition or example of what an ungradable image was provided. Furthermore, usage of external datasets and annotations was forbidden.

Glaucoma classification with uncertainty

Dataset. The AIROGS development data [3] contains 101 442 images, from which 3 270 have referable glaucoma. For our experiments, we randomly split the data into training, validation and test sets with 80%, 10% and 10% of the data, respectively. Thus, both the validation and the test set contained 10 145 images, from which 327 were graded as referable glaucoma. All images were resized to the input size of the network. The test dataset has approximately 11 000 images. These images and their labels were hidden from the participants, and instead performance evaluation was performed by submitting the algorithm to the AIROGS web platform. However, prior to the final test phase on these images, the challenger organizers allowed to assess the performance of the algorithm on around 10% of the test data. A total of 3 attempts were possible for this preliminary test phase.

Base architecture The algorithm was developed using exclusively the AIROGS dataset [3]. The classification model is composed of the first two inception blocks from the Inception-V3 [4] network pre-trained on ImageNet [5]. Using only these blocks reduces the size of the receptive field which, as it will be addressed later, allows to identify in detail the relevant diagnosis regions and subsequently propose an OOD binary decision.

Deep Dirichlet uncertainty estimation. Our method is based on the direct modeling of the uncertainty following the evidential deep learning approach [6]. In particular, we deal with the K class probabilities as resulting from a Dirichlet distribution, i.e., a belief mass b_k is attributed to each singleton (i.e, class label) k, $k \in \{1,...,K\}$, from a set of mutually exclusive singletons, and an overall uncertainty mass u is provided, with $u \geq 0$, $b_k \geq 0$ and $u + \sum_{k=1}^{K} b_k = 1$. Each b_k is computed based on the evidence for that singleton e_k via $b_k = e_k/S$, where S is the total evidence. The prediction uncertainty u is:

$$u = \frac{K}{S} = \frac{K}{\sum_{k=1}^{K}(e_k + 1)}. \quad \textbf{(1)}$$

The uncertainty is thus inversely proportional to the total evidence, and in the extreme case of no evidence we have $b_k = 0, \forall k \implies u = 1$. This evidence can be modeled by a Dirichlet distribution characterized by K α_k parameters, with $\alpha_k = e_k + 1$. The probability $\hat{p_k}$ of the class k is given by the mean of the Dirichlet distribution parameters:

$$\hat{p_k} = \frac{\alpha_k}{S} \tag{2}$$

We utilize the uncertainty value u to detect OOD cases.

Network Training. The network receives as input 224×224 pixels RGB images and outputs per sample the glaucoma probability and the confidence of the prediction. For that, we first obtain $K = 2$ logits , which are clipped to $[-200, 200]$ and then converted to evidences (e) using a softplus activation. We train the model with two loss terms based on Kullback-Leibler (KL) divergence. The first term aims at increasing e for the correct class by assessing the divergence between the predicted α and the theoretically maximum $\alpha_{\text{max}} = 201$:

$$L_{\text{KL}_{\text{evid}}} = \text{KL}\left(D(p_i|\alpha_i) \;||\; D(p_i|y_{\text{gt}} \odot \langle\alpha_{\text{max}}, ..., \alpha_{\text{max}}\rangle)\right) \tag{3}$$

where y_{gt} is the reference categorical label. A second KL divergence term regularizes the distribution by penalizing the divergence from the uniform distribution in the uncertain cases:

$$L_{\text{KL}_{\text{unif}}} = \text{KL}\left(D(p_i|\hat{\alpha_i}) \;||\; D(p_i|\langle 1, ..., 1\rangle)\right) \tag{4}$$

where $D(p_i|\langle 1, ..., 1\rangle)$ is the uniform Dirichlet distribution and $\hat{\alpha_i}$ is the Dirichlet parameters after removing the non-misleading evidence from the α_i parameters for sample i: $\hat{\alpha_i} = y_i + (1 - y_i) \odot \alpha_i$. The final loss is then defined as:

$$L = L_{\text{KL}_{\text{evid}}} + a_t L_{\text{KL}_{\text{unif}}} \tag{5}$$

with a_t being the annealing coefficient that increases as the training progresses. In particular, $a_t = \min(1, t/s)$, where t is the current training epoch and s is the annealing step, gradually increasing the effect of the second term in the final loss, avoiding the premature convergence to the uniform distribution for misclassified images in the beginning of the training [1]. Our model was trained with balanced batches and the data was randomly augmented with flips, translations, rotations, scales, Gaussian blur and brightness modifications.

Out-of-distribution binary decision. The challenge required participants to indicate, both with a continuous score and a binary label, if an image is ungradable. Since no examples of ungradable images were provided, we made the assumption that diagnosis is only possible if the OD has enough image quality for diagnosis, as glaucoma main structural manifestation occurs in that region. Thus, we artificially created OOD images by zeroing the regions of the images where their Grad-CAM [7] is greater than 0.5. This allowed us to produce in-distribution (ID) and OOD samples in our validation set, with which we computed the threshold for the binary ungradability decision. In particular, we contructed a receiver operating characteristic (ROC) curve using u_{ID} and u_{OOD}. The ROC curve was used for selecting two decision thresholds, one at 0.5 sensitivity ($u = 0.35$) and the other at the optimal operating point ($u = 0.13$). We tested both thresholds on the preliminary test phase, and we kept $u = 0.35$ as it performed better on that data.

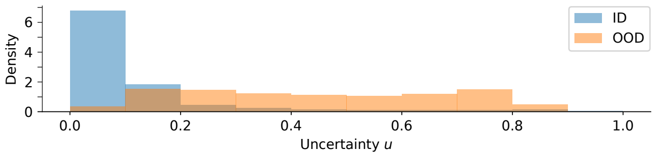

Fig. 1. Uncertainty histogram of the in-distribution (ID) and the artificial out-of-distribution (OOD) cases.

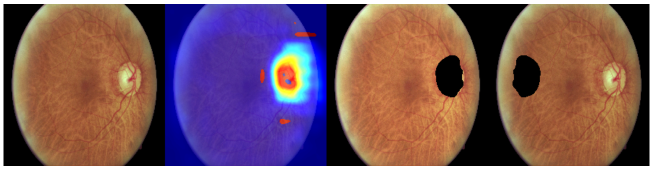

Fig. 2. Representative example of (left-to-right) original image ($u = 0.08$), Grad-CAM overlay, out-of-distribution by optic disc obscuring ($u = 0.54$), and out-of-distribution by flipping the binarized Grad-CAM ($u = 0.08$).

Evaluation and Results

The uncertainty histogram (Fig 1) for ID and OOD shows that the predicted uncertainty u is a viable metric to identify images where the OD is not visible. To ensure that this behaviour was due to the OD being obscured, we compared the AUC values for detecting our OOD cases with the values for detecting the cases where the corresponding Grad-CAM mask was flipped vertically (see Fig. 2). The achieved AUC values were 0.905 and 0.506, respectively, thus validating our hypothesis that the OD image quality is pivotal for this task.

The challenge participants were evaluated using four metrics: 1) the partial area under the ROC curve (90-100% specificity) for referable glaucoma (pAUC), 2) sensitivity at 95% specificity (TPR@95), 3) Cohen's kappa score between the reference and the decisions provided by the challenge participants on image ungradability (κ) and 4) the ungradability AUC (gAUC). Table 1 shows our results on our test set and on the preliminary test phase. As shown, besides a 10% performance drop at TPR@95 and an over-optimistic estimation of κ, which were expected given the reduced number of glaucoma cases and complexity of the ungradability task, our approach shows a similar behaviour on both datasets. Importantly, the model shows high scores across all the metrics. In fact, at the time of the opening of the final test phase (Feb. 8th, 2022), our method had the highest average score among all submissions.

Table 1. Performance of the proposed method using the challenge metrics on the preliminary test phase.

Test data	pAUC	TPR@95	κ	gAUC
Our test set	0.9187	0.8990	0.6915	0.9049
Pr. test phase	0.8464	0.7813	0.4452	0.8691

Conclusion

In this paper, we presented our method for the AIROGS challenge which, being based on the Dirichlet distribution, allows to obtain a probability of referable glaucoma and the corresponding prediction uncertainty. Even without explicit supervision, the model is capable of detecting OOD cases while maintain a high performance on the classification task.

Bibliography

1. Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, 2018.
2. Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, Bram van Ginneken, Hans G. Lemij, and Clara I.Sánchez. AIROGS: Artificial Intelligence for RObust Glaucoma Screening Challenge, 2021.
3. Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, Bram van Ginneken, Hans G. Lemij, and Clara I.Sánchez. Rotterdam EyePACS AIROGS train set, 2021.
4. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. ISSN 08866236. doi: 10.1002/2014GB005021.
5. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. ISSN 15731405. doi: 10.1007/s11263-015-0816-y.
6. Arthur P. Dempster. A Generalization of Bayesian Inference. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 73–104. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 9783540343561. doi: 10.1007/978-3-540-44792-4{_}4.
7. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Venice, 10 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.74.