DATA-CENTRIC LABEL SMOOTHING FOR EXPLAINABLE GLAUCOMA SCREENING FROM EYE FUNDUS IMAGES

Adrian Galdran[1,2] *Miguel A. González Ballester*[2,3]

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[2] BCN Medtech, Dept. of Information and Communication Technologies,
Universitat Pompeu Fabra, Barcelona, Spain
[3] ICREA, Barcelona, Spain

ABSTRACT

As current computing capabilities increase, modern machine learning and computer vision system tend to increase in complexity, mostly by means of larger models and advanced optimization strategies. Although often neglected, in many problems there is also much to be gained by considering potential improvements in understanding and better leveraging already-available training data, including annotations. This so-called data-centric approach can lead to substantial performance increases, sometimes beyond what can be achieved by larger models. In this paper we adopt such an approach for the task of justifiable glaucoma screening from retinal images. In particular, we focus on how to combine information from multiple annotators of different skills into a tailored label smoothing scheme that allows us to better employ a large collection of fundus images, instead of discarding samples suffering from inter-rater variability. Internal validation results indicate that our bespoke label smoothing approach surpasses the performance of a standard resnet50 model and also the same model trained with conventional label smoothing techniques, in particular for the multi-label scenario of predicting clinical reasons of glaucoma likelihood in a highly imbalanced screening context. Our code is made available at `github.com/agaldran/justraigs`.

Index Terms— Data-Centric Computer Vision, Glaucoma Screening, Explainability, Label Smoothing

1. INTRODUCTION AND RELATED WORK

Since the introduction of the transformer architecture [1], many recent advances in the area of machine learning have resulted from the optimization of scaling up models and improving learning algorithms, but the field has also realized that it is equally important to handle training data effectively. Enhancing the quality and quantity of training data by engineering it before moving it into machine learning systems is an area of research known as data-centric artificial intelligence. As opposed to standard model-centric approaches, where we focus on designing better learning mechanism and models to improve performance, data-centrism investigates possible deficiencies in data, *e.g.* missing values [**?**], cleaning wrong labels [2], or finding and removing out-of-distribution samples prior to training [3]. In computer vision, applications range from semantic segmentation [4] or object detection [5]. A recent review on the topic can be found in [6].

In the medical image analysis field, the maximum exponent of data-centric strategies is arguably the widely popular nnU-net segmentation framework [7]. In this case, given a 3d medical image dataset, the system automatically generates a footprint indicating not only the type of CNN architecture to be used, but also a highly performing set of hyperparameters, e.g. volumetric patch size, batch size, pre-processing operations, and so on. Designed to compete in the Medical Decathlon challenge [8], the nnU-Net has since become the default baseline over which to build improvements on 3d medical segmentation tasks [9, 10, 11, 12].

In this article, we adopt a data-centric approach for the task of explainable glaucoma screening from retinal fundus images. Glaucoma is a sight-threatening disease that represents the second leading global cause of blindness, impacting over 91 million people worldwide [13, 14]. Due to the ease of acquisition and wide availability of public retinal image fundus data collections, a large set of classification models for the detection of glaucoma from fundus images have been proposed in recent years[15], including in the context of public competitions [16]. However, the black-box nature of deep neural networks has prevented these high-performing models from reaching common clinical practice[17]. Some authors have attempted to predict and regress common biomarkers such as vertical cup-to-disc ratio [18], or more recently the Rim Thickness Curve [19]. Another route towards greater explainability of glaucoma diagnosis is by means of directly predicting a set of relevant clinical features that result in a clinician declaring disease presence. Even if this might be an optimal solution, it requires richly annotated data that has only very recently been available to the community[20].

In this work, we focus on the engineering of a tailored label smoothing that reflects inter-rater disagreement. Our system is designed to take part in the JustRaigs competition, where the provided dataset [20] featured a large scale set of retinal images labeled by a pool of annotators of varying expertise. We quantitatively show that incorporating information on the amount of expertise into soft labels can enhance the predictive ability of a standard ResNet50 model for the tasks of glaucoma screening and justification.

2. METHODOLOGY

In this section, we first give an overview of our baseline model, composed of a standard computer vision architecture trained with conventional techniques. We then describe the main features of the dataset we used in this paper, focusing on the nature and quality of its annotations. The last subsection explains how we incorporated this knowledge into a specific label smoothing scheme suitable for our learning task.

2.1. Baseline Model

In this paper, we considered a ResNet50 architecture, which has been shown as an extremely competitive computer vision model with a good compromise between complexity and accuracy [21].

Since we intend to take a data-centric approach, we do not spend much time optimizing hyper-parameters. This is, we optimize the network with an Adam algorithm and a default learning rate of $l = 1e - 4$, batch-size of 8, minimizing a regular Cross-Entropy loss for as long as we do not detect overfitting on a separate validation set comprising 20% of the data (we run a 5-fold training ensemble). We crop the images to their field of view, and apply common data augmentation strategies, *e.g.* random flipping and rotations, or small image intensity perturbations.

2.2. Data and Annotations

The dataset provided by the competition organizers contains 113,893 retinal images labeled for glaucoma analysis. Specifically, there is a main annotation reporting the status of glaucoma in the patient, but also a subset of the images has a rich collection of supplementary annotations, as explained next.

In order to understand our approach, it is important to explain the details of the annotation process. Initially, a pool of graders, both ophthalmologists and optometrists, graded the images. For each image, a randomly selected pair of raters reported the image to be with referable glaucoma (RG), no referable glaucoma (NRG), or ungradable (U). Whenever there was disagreement between grader 1 (G1) and grader 2 (G2), a glaucoma expert (G3) resolved the grading. On the other hand, performance of graders was monitored and part of them abandoned the study if their accuracy was not deemed sufficient. Their images were reannotated, unless G3 had already gave a diagnosis, in which case the low-quality diagnosis was simply removed. As a consequence, a small subset of the labels for G1 or G2 would become unavailable (NaN). Eventually, in order to reach a binary label, an annotation considered as final whenever there is agreement between G1 and G2, or in case of disagreement, the final label is the one provided by by the specialist G3.

For the explainability task, images identified as showing signs of RG were given annotations for 10 clinically relevant glaucomatous features $f_1, ..., f_{10}$ [20]. In this case, if both G1 and G2 initially reported RG, then the dataset contains two sets $f_1^1, ..., f_{10}^1$ and $f_1^2, ..., f_{10}^2$ that may not be equal, but there was no adjudication in cases of disagreement at the glaucoma feature level. In addition, if the image was deemed glaucomatous by one of both graders, and G3 agreed, then we also have two sets of feature values that may not coincide. For purposes of evaluation, the competition organizers discarded any feature values that showed disagreement.

A straightforward approach would consist of training a model using only final referrable glaucoma annotations and discarding all information regarding disagreement. In contrast, we attempt to incorporate this into the labels used for training, as explained in the next subsection.

2.3. Multi-Rater Data-Centric Label Smoothing

Conventional Label Smoothing strategies for binary classification attempt to regularize the training process by substituting "hard labels" $y \in \{0, 1\}$ for hard-coded soft values $\tilde{y} \in \{0.1, 0.9\}$, thereby preventing a network trained with cross-entropy loss from becoming overly confident [22, 23].

We propose label smoothing to encode inter-rater disagreement by following a set of rules described below:

- Whenever the final label $y = 0$ or $y = 1$ but a rater considered the image as ungradable (U), we use soft labels $\tilde{y} = 0.1$ or $\tilde{y} = 0.9$ instead.
- Whenever there was disagreement between G_1 and G_2, instead of directly adopting the decision of G_3, we soften it so that if $G3 = G_i = 0, G_j = 1$, we use $\tilde{y} = 0.15$; conversely, if $G3 = G_i = 1, G_j = 0$, we use $\tilde{y} = 0.85$.
- If an annotation is missing, *i.e.* G_1=NaN, and $G_2! = G_3$, instead of considering G_3 as the final decision, if $G_3 = 0$ then we use $\tilde{y} = 0.2$ and if $G_3 = 1$, we consider $\tilde{y} = 0.8$.

For the feature-level labels, we have sets of ten binary annotations, that we pre-process as follows:

- If $G_i = 1 \neq G_j$, with $G_3 = 1$, we have two sets of feature values. When there is disagreement at the feature level, we assign $f \in \{0.1, 0.9\}$ favoring the opinion of G3.
- If $G_i = 1 \neq G_j$, with $G_3 = 0$, we have one set of feature values. For each feature that has a value of $f = 1$, we consider a label of $f = 0.25$
- This still leaves cases where $G_i = G_j = 1$, but not at the feature level. In case of disagreement at the feature level in this context, we use $f = 0.5$.

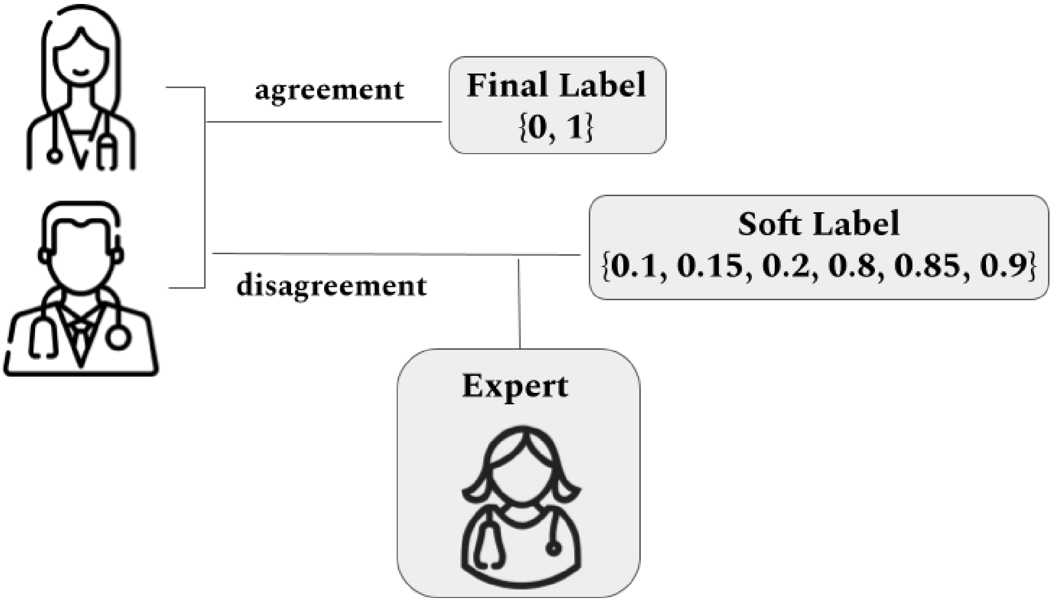

Fig. 1: Data-Centric Label Smoothing: depending on the sort of disagreement present on annotations, and the skill of the involved annotators, the degree of smoothing appliced to labels will vary.

The resulting set of different smooth labels we use is illustrated in Fig. 1.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Final	93.12	92.05	92.97	**94.33**	92.51
LS	92.97	91.59	92.35	93.57	90.67
DC-LS	**93.27**	**93.43**	**93.27**	93.87	**92.66**

Table 1: Five fold sens@95spec for Glaucoma screening using different label configurations. Best results **boldfaced**.

3. EXPERIMENTAL RESULTS

We trained the same model five times by separating the training and validation sets in 80/20 proportions[1]. In Table 1 we show, for the referrable glaucoma problem (binary classification) the results of these experiments when using only data with final decisions (Final), when using all data but a uniform label smoothing (LS), and when using the data-centric label smoothing (DC-LS) scheme described in the previous section. The metric of choice is sensitivity at 95% specificity, as indicated in the competition guidelines.

We can see how the data-centric label smoothing results in improvements in terms of sensitivity in all folds but one. In addition, even when standard label smoothing leads to degraded performance, our approach still improves the model's results. This insight is confirmed in Table 2, where we report results for the explainable glaucoma feature prediction task, terms of Hamming error between the true feature vector and the predicted one. We again compare against training on data with full agreement or expert opinion, plus when using label smoothing. We again see that data-centric label smoothing results on lower Hamming losses, indicating that the extra training data and the adapted soft labels can improve performance of a standard classifier also on this multi-label problem.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Final	0.2251	0.2286	0.2127	0.2129	0.2352
LS	0.1823	0.1764	0.1748	0.1743	0.1799
DC-LS	**0.1468**	**0.1440**	**0.1455**	**0.1488**	**0.1476**

Table 2: Five fold Hamming Loss for Glaucoma features prediction with different label schemes. Best results **boldfaced**.

4. CONCLUSION

In this paper we describe Data-Centric Label Smoothing, an adaptation of conventional label smoothing that takes into account multi-rater disagreement s and different expert skills in order to define a set of soft labels, both for binary and a multi-label classification tasks. Our experiments show that introducing otherwise discared data with soft labels into the training of a standard Resnet50 model leads to substantial performance increases.

[1] Code is available at `github.com/agaldran/justraigs`

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available by the JustRaigs challenge organizers in an open access Zenodo repository (link). Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

Adrian Galdran was partially funded by Marie Sklodowska-Curie Fellowship No. 892297.

7. REFERENCES

[1] A. Vaswani et al., “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, 2017.

[2] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang, “Confident Learning: Estimating Uncertainty in Dataset Labels,” *Journal of Artificial Intelligence Research (JAIR)*, vol. 70, pp. 1373–1411, 2021.

[3] Rashmiranjan Nayak et al., “A comprehensive review of datasets for detection and localization of video anomalies: a step towards data-centric artificial intelligence-based video anomaly detection,” *Multimedia Tools and Applications*, Dec. 2023.

[4] Adrian Galdran et al., “A No-Reference Quality Metric for Retinal Vessel Tree Segmentation,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 2018, pp. 82–90.

[5] Ulyana Tkachenko et al., “ObjectLab: Automated Diagnosis of Mislabeled Images in Object Detection Data,” in *ICML Workshop on Data-centric Machine Learning Research*, 2023.

[6] Daochen Zha et al., “Data-centric Artificial Intelligence: A Survey,” June 2023, arXiv:2303.10158 [cs].

[7] Fabian Isensee et al., “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[8] Michela Antonelli et al., “The Medical Segmentation Decathlon,” *Nature Communications*, vol. 13, no. 1, pp. 4128, July 2022.

[9] Lalith K. S. Sundar et al., “Fully Automated, Semantic Segmentation of Whole-Body 18F-FDG PET/CT Images Based on Data-Centric Artificial Intelligence,” *Journal of Nuclear Medicine*, vol. 63, no. 12, pp. 1941–1948, Dec. 2022.

[10] J. Wasserthal et al., “TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images,” *Radiology: Artificial Intelligence*, vol. 5, Sept. 2023.

[11] Niccolò McConnell et al., “Exploring advanced architectural variations of nnUNet,” *Neurocomputing*, vol. 560, pp. 126837, Dec. 2023.

[12] Fabian Isensee et al., “Extending nnU-Net Is All You Need,” in *Bildverarbeitung für die Medizin 2023*, 2023.

[13] Anshul Thakur et al., “Predicting Glaucoma before Onset Using Deep Learning,” *Ophthalmology. Glaucoma*, vol. 3, no. 4, pp. 262–268, 2020.

[14] Yeganeh Madadi et al., “Domain Adaptation-Based Deep Learning Model for Forecasting and Diagnosis of Glaucoma Disease,” *Biomedical Signal Processing and Control*, vol. 92, June 2024.

[15] Ruben Hemelings et al., “A generalizable deep learning regression model for automated glaucoma screening from fundus images,” *npj Digital Medicine*, vol. 6, no. 1, pp. 1–15, June 2023.

[16] José Ignacio Orlando et al., “REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs,” *Medical Image Analysis*, vol. 59, Jan. 2020.

[17] Jo-Hsuan Wu et al., “Performances of Machine Learning in Detecting Glaucoma Using Fundus and Retinal Optical Coherence Tomography Images: A Meta-Analysis,” *American Journal of Ophthalmology*, vol. 237, pp. 1–12, May 2022.

[18] Ruben Hemelings et al., “Deep learning on fundus images detects glaucoma beyond the optic disc,” *Scientific Reports*, vol. 11, no. 1, pp. 20313, Oct. 2021.

[19] Anna M. Wundram et al., “Leveraging Probabilistic Segmentation Models for Improved Glaucoma Diagnosis: A Clinical Pipeline Approach,” in *Medical Imaging with Deep Learning*, 2024.

[20] H. G. Lemij et al., “Characteristics of a Large, Labeled Data Set for the Training of Artificial Intelligence for Glaucoma Screening with Fundus Photographs,” *Ophthalmology Science*, vol. 3, no. 3, Sept. 2023.

[21] Ross Wightman et al., “ResNet strikes back: An improved training procedure in timm,” Oct. 2021.

[22] Adrian Galdran et al., “Multi-Head Multi-Loss Model Calibration,” in *Medical Image Computing and Computer Assisted Intervention 2023*, 2023, pp. 108–117.

[23] Balamurali Murugesan et al., “Calibrating segmentation networks with margin-based label smoothing,” *Medical Image Analysis*, vol. 87, pp. 102826, July 2023.