Automatic Classification and Triage of Diabetic Retinopathy from Retinal Images based on a Convolutional Neural Networks (CNN) Method

Adrian Galdran[1]*, Hadi Chakor[2], Ryad Kobbi[2], Argyrios Christodoulidis[2], Jihed Chelbi[2], Marc-André Racine[2] , and Ismail ben Ayed[1]
[1]Ecole de Technologie Superieure - ETS, Montréal (Canada), [2]Diagnos INC, Montréal (Canada)

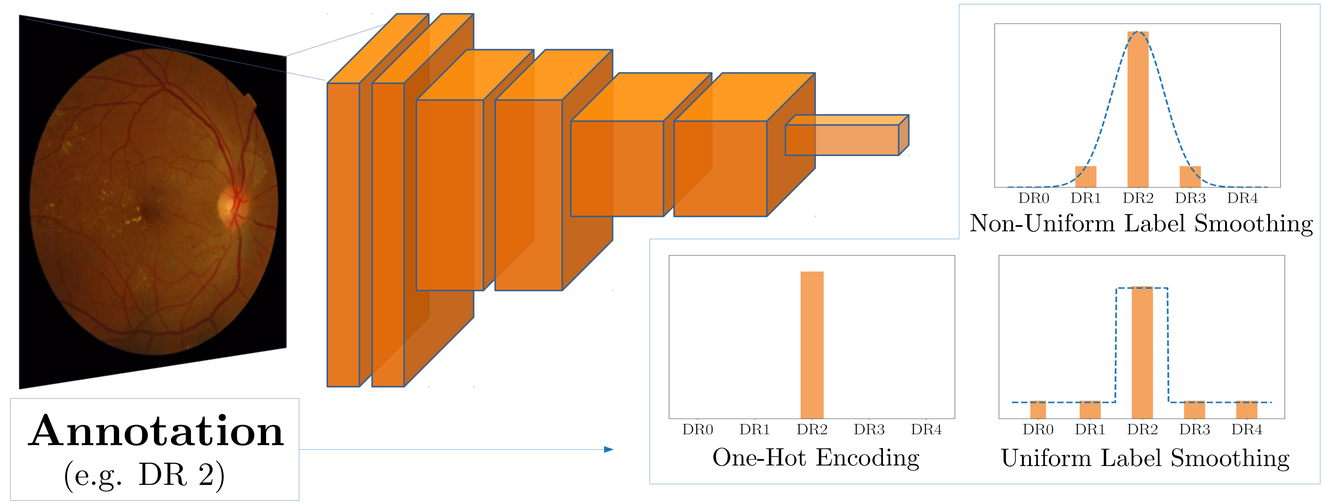
Summary

PURPOSE: We introduce a new technique to improve Deep Learning (DL) models designed for automatic grading of Diabetic Retinopathy (DR) from retinal fundus images by enhancing predictions consistency.

METHODS: A Convolutional Neural Network (CNN) was optimized in three different manners to predict DR grade from eye fundus images. The optimization criteria were 1) the standard Cross-Entropy (CE) loss, 2) CE supplemented with Label Smoothing (LS), a regularization approach widely employed in computer vision tasks, and 3) our proposed Non-Uniform Label Smoothing (N-ULS), a modification of LS that models the underlying structure of expert annotations.

RESULTS: Performance was measured in terms of per-grade Area Under the ROC curve, and also with suitable metrics for analyzing diagnostic consistency, like Quadratic-weighted Kappa score (Quad-Kappa). Whilst LS generally harmed the performance of the CNN, N-ULS statistically significantly (s.s.) improved performance with respect to CE in predicting DR grade 0 (0.947+/-0.004 vs. 0.940+/-0.004) and grade 2 (0.950+/-0.004 vs. 0.923+/-0.002), without any s.s. performance decrease in the remaining grades. N-ULS achieved this while simultaneously s.s. increasing the Quad-Kappa score (0.732+/- 0.013 vs. 0.777+/-0.012) and all other analyzed metrics.

CONCLUSIONS: For extending standard modeling approaches from DR detection to the more complex task of DR grading, it is essential to consider the underlying structure of expert annotations. The approach introduced in this paper can be easily implemented in conjunction with DNNs to increase their consistency without sacrificing per-class performance.



Methods

- **Standard Approach**: Cross-Entropy (**CE**) error and annotations represented in one-hot encoding [1,2].
- We are interested in a simple but much successful approach to regularization in CNNs, a technique called **Label Smoothing** [3]. Replace "hard labels" by a smoothed version of them, in which part of their "truth value" is redistributed in an uniform manner (**LS**) among the rest of the labels. **ULS** can contribute to increase learning speed, benefits overall accuracy, and improves model calibration [3,4].
- We propose to modify the standard **LS** regularization technique by replacing the uniform label noise distribution by a Gaussian distribution centered around the true label k with a decay factor σ selected in such a way that 95% of the probability mass still falls within its neighboring grades. Mathematically:

$$\bullet\, y_k = y_k(1\text{-}a) + a \bullet G_{k,\sigma},$$

  for each class k in {0, 1, 2, 3, N=4}, where * represents standard convolution.
- By implementing a Non-Uniform Label Smoothing scheme (**N-ULS**), we expect to bias the learning of a DR grading CNN towards a model that, when mistaken, produces more consistent errors. This is because **N-ULS** reflects in a more suitable manner inter-observer disagreements: two human graders differing in their opinion will most likely do so by neighboring grades than by far way ones [5]. **N-ULS** introduces in this way new information into the optimization process, since when the CNN observes a new data-point with associated annotation, it must also learn a notion of the underlying DR grading structure.

Table 1: Dataset Summary

Nr. of Images (Unique Individuals)			46,865 (27,361)	
Age (Average +/- StdDev)			59.6 +/- 14	
Female/Male			17658/9233	
	Total	Training	Validation	Test
No DR	31,447 (67,1%)	23,585	3,145	4,717
Mild DR	1264 (2,7%)	948	126	190
Moderate DR	6822 (14,6%)	5117	682	1,023
Severe DR	230 (0,5%)	172	23	35
Proliferative DR	683 (1.5%)	512	68	103
Ungradability	6419 (13,7%)	-	-	-

Results

In order to obtain confidence intervals, human annotations and model predictions where bootstrapped (n=1000) in a stratified manner with respect to the relative presence of each grade. We computed statistical significance (s.s.) of the proposed Non-Uniform Label Smoothing (**N-ULS**) scheme when compared with standard Cross-Entropy (CE) and ordinary Label Smoothing (**LS**). We observe that:

1) When comparing **N-ULS** with respect to ULS, the AUROC was s.s. better for diagnosing DR grade 0 (p=0.009), grade 2 (p=0.002), and grade 3 (0.014). For the remaining two grades, differences in performance were not s.s..

2) Regarding CE, when comparing it against **N-ULS** we found that the latter achieved s.s. greater AUROC for DR grade 0 (p=0.004) and grade 2 (0.000). For the other three grades, again there were no s.s. differences.

We see that while the introduction of standard LS in the training of a network seems to generally harm performance, this is not the case for the **N-ULS** approach.

Also, the average performance of CE and **N-ULS** was similar, but the quadratic kappa was significantly higher for the **N-ULS** approach, which confirms our hypothesis that the error consistency can be improved by means of a simple domain-specific label smoothing strategy.

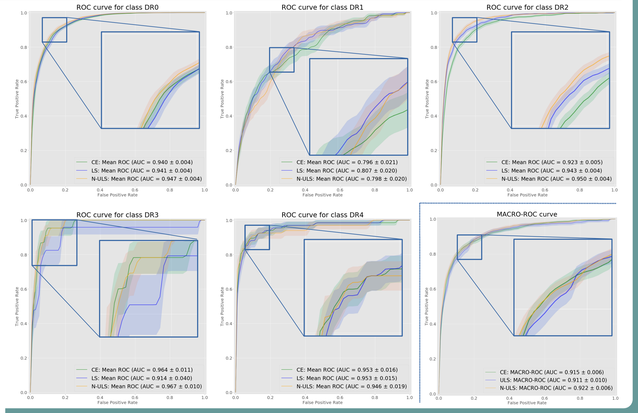


Table 2: Performance analysis in terms of per-class AUROCs and Quadratic-weighted Kappa for each of the considered approaches. Statistically significant results marked bold.

	AUROC -DR0	AUROC -DR1	AUROC -DR2	AUROC -DR3	AUROC -DR4	Quad-Kappa
CE	0.940 +/- 0.004	0.796 +/- 0.021	0.923 +/- 0.005	0.964 +/- 0.011	0.955 +/- 0.016	0.732 +/- 0.013
LS	0.941 +/- 0.004	0.807 +/- 0.020	0.943 +/- 0.004	0.914+/- 0.004	0.953 +/- 0.015	0.722 +/- 0.016
N-ULS	**0.947 +/- 0.004**	0.798 +/- 0.020	**0.950 +/- 0.004**	**0.967 +/- 0.010**	0.946 +/- 0.019	**0.777 +/- 0.012**

Table 3: Performance analysis including several other metrics of interest

	Macro-AUROC	Weighted-Precision	Weighted-Recall	Weighted-F1	MCC	Kendall-τ
CE	0.915 +/- 0.006	0.865 +/- 0.005	0.844 +/- 0.005	0.853 +/- 0.005	0.597 +/- 0.013	0.697 +/- 0.012
LS	0.911 +/- 0.010	0.852 +/- 0.006	0.870 +/- 0.004	0.851 +/- 0.004	0.607 +/- 0.012	0.678 +/- 0.012
N-ULS	0.922 +/- 0.006	0.873 +/- 0.005	0.891 +/- 0.004	0.879 +/- 0.004	0.680 +/ 0.012	0.745 +/- 0.011

References

[1] Krause, J., Gulshan, V., Rahimy, E. et al., Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy, AAO, 2018.

[2]Sahlsten, J., Jaskari, J.,Kivinen, J., et al., Deep Learning Fundus Image Analysis for DR and DME Grading, Nature Scientific Reports, 2019.

[3] Szegedy , C., Vanhoucke , V., Ioffe, S., et al., Rethinking the Inception Architecture for Computer Vision, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] Müller, R., Kornblith, S., and Hinton, G., When Does Label Smoothing Help?, arXiv:1906.02629, 2019.

[5] Ruamviboonsuk, P., Teerasuwanajak, K., Tiensuwan, M., et al., Interobserver Agreement in the Interpretation of Single-Field Digital Fundus Images for Diabetic Retinopathy Screening, Ophthalmology, 2006.