Maximum Entropy on Erroneous Predictions: Improving model calibration for medical image segmentation

Supplementary material

Agostina J. Larrazabal[1,4], César Martínez[1], Jose Dolz*[2,3], and Enzo Ferrante*[1]

[1]Research institute for signals, systems and computational intelligence, sinc(i), FICH-UNL / CONICET, Santa Fe, Argentina
[2]LIVIA, ETS Montreal, Canada
[3] International Laboratory on Learning Systems (ILLS)
[4] Tryolabs, Uruguay

Training loss		Segmentation performance				Calibration performance					
		Dice coefficient		HD		Brier (10^{-4})		Brier^{+}		ECE (10^{-3})	
		WMH	LA	WMH	LA	WMH	LA	WMH	LA	WMH	LA
$\mathcal{L}_{dice}$	–	0.768 (0.108)	**0.905 (0.024)**	20.499 (8.468)	24.311 (7.733)	6.742 (4.346)	24.940 (7.692)	0.258 (0.136)	0.092 (0.043)	0.667 (0.429)	24.698 (7.606)
	+ PS	0.764 (0.110)	0.904 (0.026)	20.553 (8.313)	23.412 (7.116)	6.120 (4.010)	21.248 (6.916)	0.253 (0.132)	0.099 (0.042)	1.663 (0.256)	10.933 (4.285)
	+ IR	0.770 (0.105)	0.904 (0.026)	20.518 (8.470)	**23.038 (6.535)**	5.616 (3.512)	21.034 (6.866)	0.194 (0.111)	0.097 (0.041)	1.605 (0.263)	**10.319 (5.483)**
	$+\mathcal{L}_H(\hat{Y})$	0.754 (0.122)	0.903 (0.029)	21.089 (7.475)	23.811 (8.902)	7.013 (4.643)	24.952 (8.964)	0.267 (0.157)	0.086 (0.051)	0.696 (0.461)	24.457 (8.765)
	$+\mathcal{L}_H(\hat{Y}_w)$	**0.786 (0.089)**	0.903 (0.025)	20.033 (7.566)	24.095 (8.357)	5.451 (3.492)	**19.565 (6.493)**	0.183 (0.095)	0.083 (0.036)	0.451 (0.287)	13.006 (5.160)
	$+\mathcal{L}_{KL}(\hat{Y}_w)$	**0.786 (0.093)**	0.900 (0.028)	**18.848 (6.513)**	23.600 (6.496)	**5.379 (3.502)**	20.106 (6.971)	**0.174 (0.094)**	**0.079 (0.042)**	**0.434 (0.285)**	12.561 (6.222)
$\mathcal{L}_{CE}$	–	0.770 (0.104)	0.890 (0.035)	18.928 (7.175)	26.596 (8.121)	6.256 (4.044)	22.458 (8.417)	0.259 (0.128)	0.113 (0.055)	0.602 (0.390)	16.700 (8.763)
	+ PS	0.775 (0.101)	0.893 (0.033)	19.385 (7.667)	25.508 (8.387)	5.296 (3.324)	21.092 (6.955)	0.194 (0.101)	0.088 (0.044)	1.637 (0.168)	12.140 (5.482)
	+ IR	0.775 (0.100)	0.892 (0.034)	19.649 (7.834)	26.041 (8.314)	5.236 (3.229)	21.023 (7.160)	**0.184 (0.096)**	**0.091 (0.045)**	1.590 (0.183)	**11.765 (5.758)**
	$+\mathcal{L}_H(\hat{Y})$	0.778 (0.092)	**0.896 (0.030)**	**18.554 (7.214)**	**25.137 (8.291)**	6.208 (3.842)	20.933 (7.552)	0.227 (0.103)	0.098 (0.044)	5.251 (0.504)	11.784 (6.481)
	$+\mathcal{L}_H(\hat{Y}_w)$	**0.779 (0.096)**	0.890 (0.036)	18.789 (7.205)	25.349 (6.265)	**5.169 (3.347)**	21.721 (7.639)	0.191 (0.093)	0.106 (0.053)	0.396 (0.257)	14.607 (7.872)
	$+\mathcal{L}_{KL}(\hat{Y}_w)$	0.775 (0.095)	0.895 (0.032)	20.949 (9.769)	25.576 (7.426)	5.269 (3.383)	**20.609 (7.260)**	0.187 (0.091)	0.097 (0.047)	**0.390 (0.248)**	12.811 (7.250)
$\mathcal{L}_{FL}$		0.780 (0.090)	0.891 (0.031)	19.759 (7.372)	26.447 (7.442)	5.621 (3.703)	21.269 (6.681)	0.216 (0.103)	0.102 (0.041)	0.472 (0.327)	14.249 (6.140)

Table 1: **Is our method backbone-agnostic?** This table includes the results for an additional experiment for both WMH and LA segmentation, using the ResUNet architecture as backbone (instead of UNet as used in the main manuscript), to show that our method is independent of the architecture. Our models are gray-shadowed and best results are highlighted in bold. Results show that the proposed regularizers typically lead to improvement on both model calibration and segmentation also for ResUNet architecture. This improvement is further stressed for the calibration metrics, where the four variants of our method significantly outperform the rest for WMH.

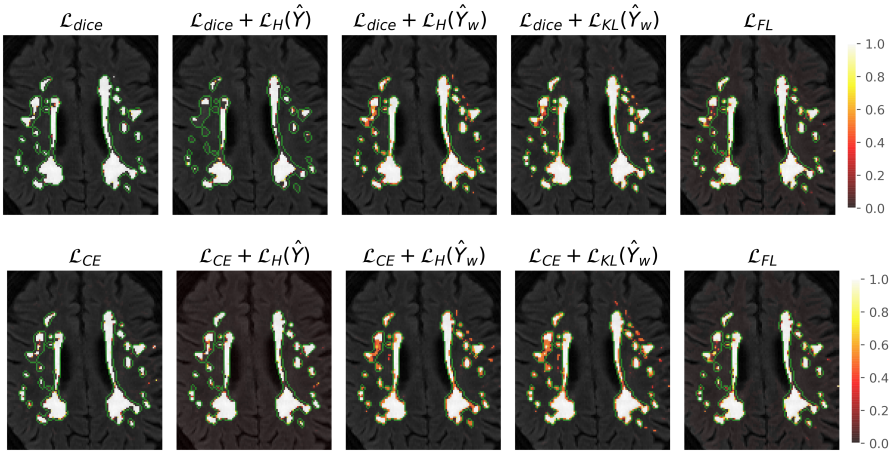

Fig. 1: **Qualitative results for WMH segmentation** We show exemplar cases of the probability maps obtained for each loss function for the WMH segmentation task. As it can be observed, the predictions made by the vanilla network trained with DSC loss tend to be highly overconfident, either assigning probabilities equal to 0 or 1. However, when employing the proposed regularizers, the models tend to use the full range of possible values, assigning scores around 0.5 (marked in red) to the more challenging pixels.