Multi-Scale Evaluation of Uncertainty Quantification Techniques for Deep Learning based MRI Segmentation

Benjamin Lambert[1,3], Florence Forbes[2], Senan Doyle[3], Alan Tucholka[3], Michel Dojat[1]

[1]Univ. Grenoble Alpes, Inserm U1216, Grenoble Institut Neurosciences, GIN, 38000, Grenoble, France

[2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000, Grenoble, France

[3]Pixyl, Research and Development Laboratory, 39000, Grenoble, France

Abstract

Deep learning systems are powerful models that tend to be considered as black-boxes. To facilitate their acceptance in clinical routine, models that *know when they don't know* are desired. A multitude of methods have been proposed to quantify uncertainty of Deep Learning predictions. Here, we propose to compare three of them : Monte Carlo Dropout, Deep Ensemble, and Heteroscedastic models. We illustrate this analysis on a task of segmentation of White-Matter Hyperintensities in FLAIR MRI sequences of Multiple-Sclerosis patients. Evaluation is carried-out at 3 different scales : the voxel, the lesion, and the whole image. Results show the superiority of the Heteroscedastic approach, outperforming competing methods at the voxel and image-level, and ranking second in the lesion task. This technique is also the fastest and the most memory-efficient. Alternative methods achieved lower uncertainty estimates quality, while also requesting longer inference time or larger computational requirements.

Introduction

Deep Learning (DL) techniques have become the gold standard for biomedical image segmentation, but they tend to be considered as black-boxes. This is partly due to the inability of neural networks to express the uncertainty in their predictions. In recent years, massive efforts have been carried out to develop models that can express their confidence[1].

In this work, we propose an in-depth evaluation of 3 state-of-the-art approaches to quantify uncertainty attached to DL predictions : Monte Carlo Dropout[2] (M1), Deep Ensemble[3] (M2), and Heteroscedastic network[4] (M3). Uncertainty estimates are evaluated at the voxel, lesion and image levels. We illustrate this comparison on an automatic segmentation task to detect White-Matter Hyperintensities (WMH) from T2-weighted FLAIR MRI sequences of Multiple-Sclerosis (MS) patients.

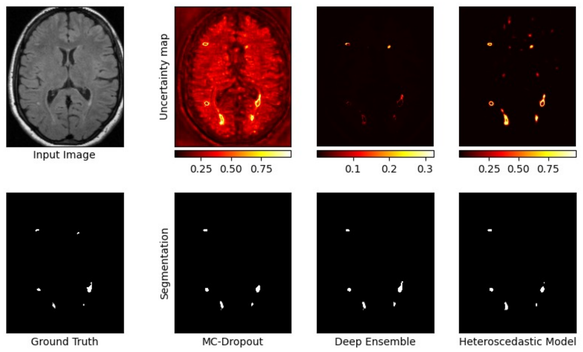

Figure 1 : Illustration of segmentation masks and uncertainty maps obtained with each technique.

Methodology

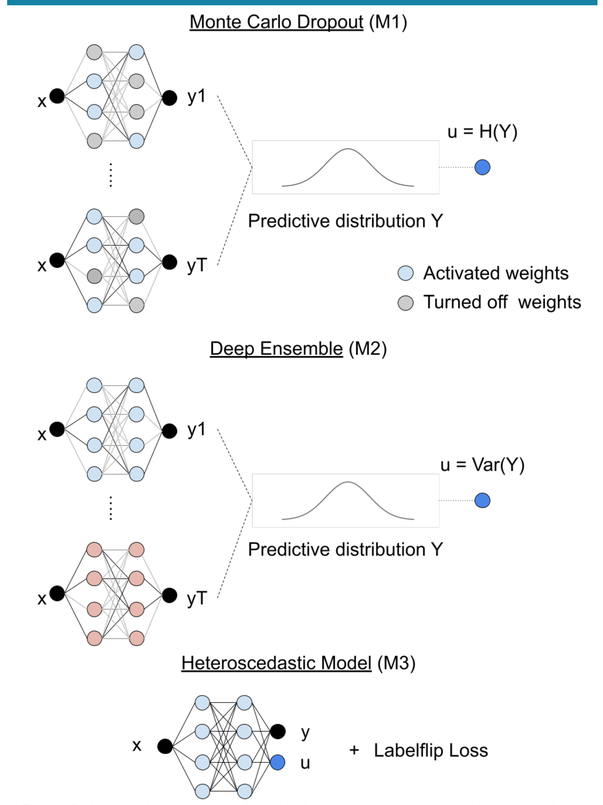

Figure 2 : Uncertainty computation pipeline for each technique. y is the predicted lesion probability, u is the uncertainty attached to the prediction.

Experiments

Dataset

We used a brain dataset composed of 238 T2-weighted FLAIR MRI sequences of MS patients, with ground truth segmentations of WMH. The dataset was split into 187 scans for training and 51 for testing. Example of segmentation masks and uncertainty maps are illustrated in Figure 1.

Voxel-wise Evaluation

At the voxel-level, we assessed the quality of uncertainty estimates with Area Under Confidence-Classification Characteristic curves (AUCCC)[5] (Figure 3). This metric is agnostic of the model segmentation performance, which is desired for a fair comparison. An AUCCC of 0.5 indicates that the uncertainties of correct and incorrect voxels are confounded, hence meaningless. Alternatively, an AUCCC of 1 indicates that correct voxels are systematically assigned with a lower uncertainty than incorrect voxels.

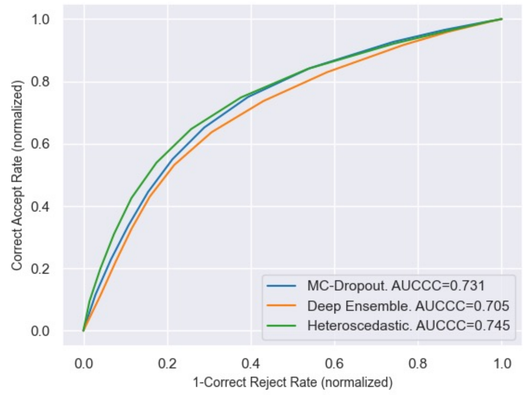

Figure 3 : Confidence-Classification Characteristic curves and their associated AUCCC scores.

Lesion-wise Evaluation

Lesion uncertainty is obtained by computing the mean of the connected components uncertainties . We progressively filtered out lesions according to their uncertainty and monitored the evolution of the predicted segmentations (number of Lesion True Positives (LTP) and number of Lesion False Positives (LFP)). This results in a Lesion Stratification curve (Figure 4), where each point corresponds to a couple (LTP, LFP) for a given uncertainty threshold. A quantitative score was finally obtained by computing the Area Under the Lesion Stratification Curve (AULSC).

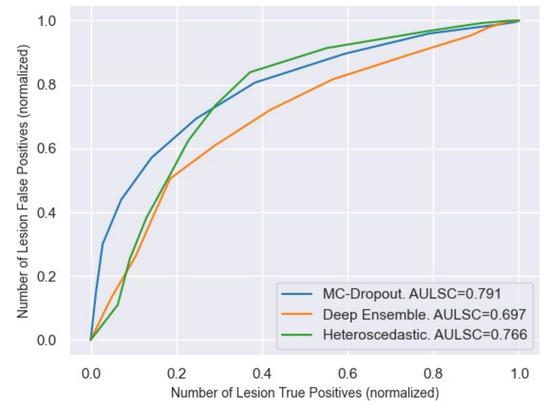

Figure 4 : Lesion Stratification curves and their associated AULSC scores. .

Image-wise Evaluation

Image uncertainty is calculated as the mean of uncertainties of all voxels predicted as lesions. We plotted correlation curves between images uncertainties and their Dice scores (segmentation quality) (Figure 5). We used the Pearson correlation coefficient (PCC) between both quantities to assess the quality of image uncertainties, and the Dice scores to estimate segmentation performance.

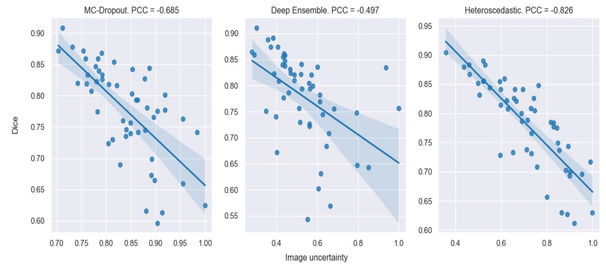

Figure 5 : Correlation curves between segmentation performance and uncertainty estimates.

Results and Conclusion

Results are presented in Figure 6, with top-scoring techniques highlighted in green. The Heteroscedastic approach (M3) outperformed competing approaches on 2 of the 3 evaluation scales regarding uncertainty estimates as well as for segmentation performance. This technique is also the fastest, as uncertainty and segmentation are simultaneously obtained in a single-step, contrary to MC-Dropout (M1) and Deep Ensemble (M2) that require the aggregation of multiple predictions.

Scale & Metrics \ Method	MC-Dropout (M1)	Deep Ensemble (M2)	Heteroscedastic (M3)
Voxel - AUCCC (E1)	0.731	0.705	**0.745**
Lesion - AULSC (E2)	**0.791**	0.697	0.766
Image - PCC (E3)	-0.685	-0.497	**-0.826**
Image - Dice (E3)	0.777	0.780	**0.784**

Figure 6 : Evaluation results of the presented methods. Top-performing method is highlighted in green.

References

1. Abdar, Moloud, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion (2021):243-297.
2. Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16). 1050–1059.
3. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17): 6402-6413.
4. Richard McKinley, Michael Rebsamen, et al. Uncertainty-Driven Refinement of Tumor-Core Segmentation Using 3D-to-2D Networks with Label Uncertainty. BrainLes at MICCAI (1) 2020: 401-411.
5. Huang X, Yang J, Li L, Deng H, Ni B, Xu Y. Evaluating and boosting uncertainty quantification in classification. arXiv preprint arXiv:1909.06030. 2019.
6. Roy AG, Conjeti S, Navab N, Wachinger C. Inherent brain segmentation quality control from fully convnet monte carlo sampling. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2018 Sep 16 (pp. 664-672). Springer, Cham.