Uncertainty Evaluation Metrics for Brain Tumour Segmentation

Raghav Mehta RAGHAV@CIM.MCGILL.CA
Centre for Intelligent Machines, McGill University, Canada

Angelos Filos ANGELOS.FILOS@CS.OX.AC.UK
Yarin Gal YARIN.GAL@CS.OX.AC.UK
Oxford Applied and Theoretical Machine Learning Group, University of Oxford, England

Tal Arbel ARBEL@CIM.MCGILL.CA
Centre for Intelligent Machines, McGill University, Canada

Abstract

In this paper, we describe and explore the metric that was designed to assess and rank uncertainty measures for the task of brain tumour sub-tissue segmentation in the BraTS 2019 sub-challenge on uncertainty quantification. The metric is designed to: (1) reward uncertainty measures where high confidence is assigned to correct assertions, and where incorrect assertions are assigned low confidence and (2) penalize measures that have higher percentages of under-confident correct assertions. Here, the workings of the metric is explored based on a number of popular uncertainty measures evaluated on the BraTS 2019 dataset.

Keywords: Brain Tumour Segmentation, Deep Neural Network, Uncertainty.

1. Introduction

Deep Neural Networks (DNN) have shown to outperform traditional machine learning methods on a variety of automatic medical image segmentation tasks, including tumour segmentation, as depicted by the highest ranking results on recent BraTS challenges (Bakas et al., 2018). However, errors in brain tumour segmentation deter the adoption of DNN frameworks in clinical contexts, particularly those that rely on high voxel-level accuracy, such as in image-guide neurosurgery. Although popular DNNs (Çiçek et al., 2016; Kamnitsas et al., 2017) for brain tumour segmentation provide tumour labels or "sigmoid"/"softmax" prediction probabilities, conveying an estimate of the resulting model uncertainties to clinicians would assist them in making a more informed decision. Although, several recent methods (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; Maddox et al., 2019) have been proposed to estimate uncertainties in deep neural networks, there is no established strategy in which their usefulness can be assessed and compared for particular clinical contexts. In this paper, we develop a metric to measure the quality of different uncertainty measures for the task of brain tumour segmentation, with the objectives: (1) when the network is correct it is confident in the predicted labels, and (2) when they are incorrect, it is not confident. This metric was used as the basis of ranking the uncertainties produced by participating teams in the BraTS 2019 sub-challenge on uncertainty quantification. It should be noted that this metric is designed assuming we do not have access to the "ground truth" uncertainty of inter-rater variability, which requires multiple annotations.

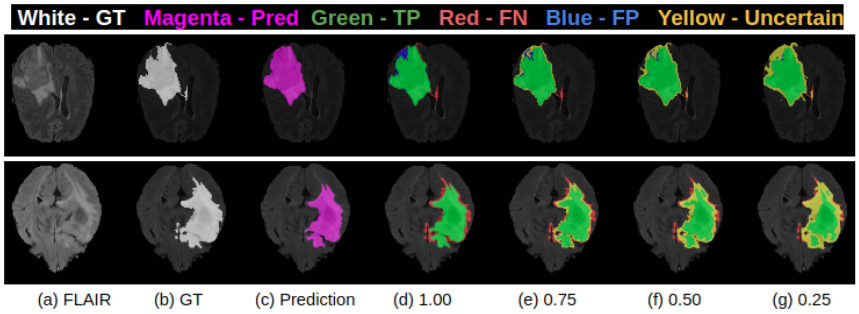

Figure 1: Effect of uncertainty thresholding on two examples for whole tumour segmentations (Top and bottom rows). (a) FLAIR MRI, (b) "Ground truth" labels, (c) Sample prediction, (d) Prediction with no filtering, and (e)-(g) Filtering with uncertainty thresholds (τ) of 0.75, 0.5 and 0.25.

	Dice			
	Dice at 1.00	**Dice at 0.75**	**Dice at 0.50**	**Dice at 0.25**
Example-1	0.94	0.96	0.965	0.97
Example-2	0.92	0.955	0.97	0.975
	Ratio of Filtered TPs ($(TP_{1.00} - TP\tau) / TP_{1.00}$)			
	FTP at 1.00	**FTP at 0.75**	**FTP at 0.50**	**FTP at 0.25**
Example-1	0.00	0.00	0.05	0.1
Example-2	0.00	0.00	0.15	0.25
	Ratio of Filtered TNs ($(TN_{1.00} - TN\tau) / TN_{1.00}$)			
	FTN at 1.00	**FTN at 0.75**	**FTN at 0.50**	**FTN at 0.25**
Example-1	0.00	0.0015	0.0016	0.0019
Example-2	0.00	0.0015	0.0026	0.0096

Table 1: Changes in Dice, Filtered True Positives (FTP), and Filtered True Negatives (FTN) with different uncertainty thresholds (τ) for two different examples.

2. Metric for Assessing and Comparing Uncertainty Measures

In the context of the BraTS 2019 challenge, each team provided their multi-class brain tumour segmentation output labels and the voxel-wise uncertainties for each of the associated tasks: whole tumour (WT), tumour core (TC) and enhanced tumour (ET) segmentations. For each task, the uncertain voxels were filtered out at several predetermined (N) number of uncertainty thresholds, τ, and the model performance was assessed based on the metric of interest (here, Dice score) on the remaining voxels at each of the thresholds. For example, at $\tau = 0.75$, all voxels with uncertainty values ≥ 0.75 are marked as uncertain. The associated predictions are filtered out, and Dice values are calculated for the remaining predictions based on the unfiltered voxels. This evaluation rewards models where the confidence in the incorrect assertions (False Positives - FPs, and False Negatives - FNs) is low and high for correct assertions (True Positives - TPs and True Negatives - TNs). For these models, it is expected that as more uncertain voxels are filtered out, the Dice score should increase on the remaining predictions.

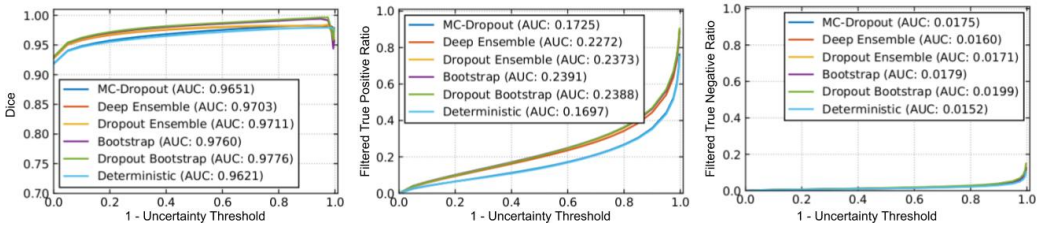

Figure 2: Effect of changing uncertainty threshold (τ) on whole tumour for entropy measure.

The proposed strategy does not keep track of the number of correctly labeled voxels that are filtered at each threshold level along with the uncertain incorrect labels. In order to penalize filtering out many correctly predicted voxels (TPs, TNs) when attaining high Dice values, an additional assessment component is added to keep track of the filtered TP and TNs voxels. Given that tumour segmentation is expected to have a high-class imbalance between tumour and healthy tissues, the system keeps track of the filtered TPs and TNs separately. The ratio of filtered TPs (FTP) at different thresholds (τ) is measured relative to the unfiltered values ($\tau = 1.00$) such that FTP = ($TP_{1.00}$ - TP_{τ}) / $TP_{1.00}$. The ratio of filtered TNs is calculated in a similar manner. This evaluation essentially penalizes approaches that filter out a large percentage of TP or TN relative to $\tau = 1.00$ voxels in order to attain the reported Dice value.

Figure 1 and Table 1 shows the workings of the assessment metric for example cases based on images from BraTS 2019. Decreasing the threshold (τ) leads to filtering out voxels with incorrect assertions, leading to an increase in the Dice value for the remaining voxels. Case 2 shows a marginally higher Dice value than Case 1 at uncertainty thresholds $\tau = 0.50$ and 0.25. However, the Ratio of Filtered TPs and TNs indicates that this is at the expense of marking more TPs and TNs as uncertain.

3. Experiments and Results

A modified 3D U-Net architecture (Çiçek et al., 2016) generates the segmentation outputs and corresponding uncertainties. We train (228), validate (57), and test (50) this network based on the publicly available Brain Tumour Segmentation (BraTS) challenge 2019 training dataset (335) (Bakas et al., 2018). The performances of whole tumour segmentation with the Entropy uncertainty measure (Gal et al., 2017) using MC-Dropout (Gal and Ghahramani, 2016), Deep Ensemble (Lakshminarayanan et al., 2017), Dropout Ensemble (Smith and Gal, 2018), Bootstrap, Dropout Bootstrap, and Deterministic, are shown in Figure 2. [1] Dropout bootstrap shows the best Dice performance (highest AUC), but also has the worst performance for Filtered True Positive and Filtered True Negative curves (highest AUC). This result shows that, here, the higher performance in Dice is at the expense of a higher number of filtered correct voxels. Overall, the metric is working in line with the objectives. However, there is no clear winner amongst these uncertainty methods in terms of rankings.

1. Due to space constraints, we showed results for whole tumour segmentation and the Entropy uncertainty measure. If the paper is accepted, results for other tumour sub-types and other uncertainties measures will be presented.

4. Conclusion

This paper provides a rationale behind the design of a metric presented at the MICCAI BraTS 2019 sub-challenge to evaluate and rank uncertainties produced by different methods for brain tumour segmentation. Using two different examples it was demonstrated that the designed metric is able to reward methods which convey higher uncertainty for incorrect assertions and penalize methods which have higher uncertainties for correct assertions.

References

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, pages 13132–13143, 2019.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.