arXiv:2010.14290v1 [eess.IV] 27 Oct 2020

POST TRAINING UNCERTAINTY CALIBRATION OF DEEP NETWORKS FOR MEDICAL IMAGE SEGMENTATION

Axel-Jan Rousseau[⋆], *Thijs Becker*[⋆], *Jeroen Bertels*[†], *Matthew B. Blaschko*[†,*], *Dirk Valkenborg*[⋆,*]

[⋆] I-Biostat, Data Science Institute, Hasselt University, Belgium
[†] Processing Speech and Images, Department of Electrical Engineering, KU Leuven, Belgium
[*] shared last authorship

ABSTRACT

Neural networks for automated image segmentation are typically trained to achieve maximum accuracy, while less attention has been given to the calibration of their confidence scores. However, well-calibrated confidence scores provide valuable information towards the user. We investigate several post hoc calibration methods that are straightforward to implement, some of which are novel. They are compared to Monte Carlo (MC) dropout. They are applied to neural networks trained with cross-entropy (CE) and soft Dice (SD) losses on BraTS 2018 and ISLES 2018. Surprisingly, models trained on SD loss are not necessarily less calibrated than those trained on CE loss. In all cases, at least one post hoc method improves the calibration. There is limited consistency across the results, so we can't conclude on one method being superior. In all cases, post hoc calibration is competitive with MC dropout. Although average calibration improves compared to the base model, subject-level variance of the calibration remains similar.

Index Terms— Medical image segmentation, uncertainty estimation, confidence calibration, deep learning

1. INTRODUCTION

Medical image segmentation is an invaluable aspect for many medical image analysis tasks, such as radiotherapy treatment planning, tumor volume measurement, analysis of tissue morphology, and others. Convolutional neural networks (NN), such as U-Net [1], are being investigated to automate the segmentation process. Although the accuracy of segmentation models has improved, their ability to quantify their predictive uncertainty is often overlooked. However, the ability of models to capture predictive confidence is a desirable property. This means that the model's (voxel-wise) predictions would directly reflect their classification accuracy. Calibrated confidences provide valuable extra information towards the end user and increase the trustworthiness of the model. For example, in a medical setting a low confidence could be a signal to hand over control to human doctors.

Dice and Jaccard metrics are among the most commonly used performance metrics for medical image segmentation. These metrics can be optimized directly by using their relaxations as loss functions, e.g. soft Dice (SD) and soft Jaccard, leading to a better segmentation performance compared to the traditional use of a cross-entropy (CE) loss which optimizes voxel-wise accuracy [2]. However, recent results indicate that models trained on SD do not produce well-calibrated models compared to CE loss [3, 4, 5]. In fact, SD optimization leads to hard predictions, even when there is true uncertainty present in the task, leaving no room for a direct interpretation as confidence scores [3]. This irreducible uncertainty is called aleatoric uncertainty and is caused by noise, artifacts, or incomplete class labels.

Bayesian probability theory is used to reason about uncertainty, and Bayesian neural nets (BNN) [6] attempt to learn distributions over the model weights and allow for uncertainty estimation, albeit the training process is generally considered computationally expensive. Gal and Ghahramani showed that Monte Carlo (MC) dropout, where dropout is performed during inference, is an approximation of BNN [7]. This method is one of the most popular for image segmentation, as it is simple to implement, at least if the original NN has dropout layers. Another approach is the use of deep ensembles [8]. Deep ensembles train multiple networks from random initialization and use model averaging to obtain the uncertainty estimate. A clear downside of deep ensembles is the need to train multiple models from scratch. Networks can also be modified to learn to estimate the aleatoric uncertainty [9]. These networks return a probability distribution and have two outputs: a mean logit output for the prediction together with its variance. Auxiliary networks trained on top of baseline models have also been proposed. DeVries and Taylor use such an auxiliary network to produce uncertainty maps starting from the baseline model's output to have an additional quality measure of the segmentation [10]. Inspired by this, Jungo et al. [11] used a simple auxiliary network consisting of three

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

convolutional layers to predict voxel-wise segmentation errors from feature maps of the segmentation network. Guo et al. found that for classification tasks, modern deep NNs are not well calibrated [12]. They evaluated the performance of various post hoc calibration methods on NN classifiers and found an extension of Platt scaling to be surprisingly effective. For segmentation tasks such post hoc calibration methods are almost never included, with some exceptions [13].

In this paper, we investigate several post hoc calibration methods for segmentation, including Platt scaling, and compare them with MC dropout. We introduce simple auxiliary networks that can be applied to most segmentation methods and that improve calibration. We furthermore show that if one takes a network trained with SD, its calibration can be significantly improved by fine-tuning the last layer with CE, with only a small drop in Dice score. Finally, we investigate the subject-level calibrations for both the base and calibrated networks.

2. METHODS

2.1. Performance metrics for calibration

Consider a model with class predictions $\hat{y}$ and associated confidences $\hat{p}$. The confidence $\hat{p}$ is found by taking the logit output of the network z through a sigmoid function: $\hat{p} = \sigma(z)$. A model is perfectly calibrated when its class prediction $\hat{y}$ and associated confidence $\hat{p}$ are correct with a probability p:

$$P\left(\hat{y} = y|\hat{p} = p\right) = p, \tag{1}$$

where y is the true class. To measure how well-calibrated a model is, we can therefore define

$$E\left[|P(\hat{y} = y|\hat{p} = p) - p|\right], \tag{2}$$

where the average is taken over all predictions [12]. This equation cannot be calculated, but can be approximated with the expected calibration error (ECE) [14]. The predictions are binned into equally spaced bins B_k, $k \in [1, K]$ (we use $K = 20$). The accuracy of bin B_k is defined as $\text{acc}(B_k) = \sum_{i \in B_k} \mathbb{1}(\hat{y}_i = y_i)/|B_k|$ and its confidence as $\text{conf}(B_k) = \sum_{i \in B_k} \hat{p}_i/|B_k|$, where $|B_k|$ is the number of elements in the bin. The ECE is the weighted average of the difference in accuracy and confidence of the bins:

$$\text{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} |\text{acc}(B_k) - \text{conf}(B_k)|, \tag{3}$$

with N is the total number of elements in all bins. This can be represented visually with reliability diagrams. These plot the accuracy of the bins against their confidence. A reliability diagram for a perfectly calibrated model plots the identity function.

2.2. Calibration methods

We investigate the following four calibration methods. Source code is available at `github.com/AxelJanRousseau/PostTrainCalibration`

Platt scaling: A simple parametric method that is often used for classification tasks [15]. It involves fitting a logistic regression model on top of the logit outputs z of the network, whilst keeping the model's weights fixed. For a segmentation task, this is equivalent to training a convolutional layer with a single kernel of size 1 on top of the logit outputs of the network, resulting in only 2 extra weights that need to be optimized using CE.

Auxiliary network: Keeping in line with this convolution analogy, we also tested single convolutional layers using larger kernel sizes. These can be seen as a simple, shallow auxiliary network, or as a generalized version of Platt scaling, where the scaling is influenced by neighbouring voxels. Kernel sizes of k = 5, 7, 9 and 11 were tested, but sizes larger than 5 provided little benefit. Therefore all reported values are for $k = 5$.

Fine-tuning: Instead of adding extra layers, we consider the effect of fine-tuning the last learnable layer of the network. Starting from the base network, the weights of all layers except the last one are frozen, and training is resumed using CE loss. As SD loss is expected to lead to poor calibration, we expect that fine-tuning with CE can improve the calibration of models trained using SD.

MC Dropout: Dropout is enabled during test time and multiple predictions are sampled (in this case 20). We either retrain the base model with dropout layers included, or we add dropout layers at test time without retraining.

2.3. Datasets

We tested the calibration performance on the publicly available training datasets of BraTS 2018 [16, 17, 18] (BR18) and ISLES 2018 [19] (IS18). The BR18 dataset contains multimodal MRI scans of 285 patients with brain tumors. For each patient T1-weighted (T1), post-contrast T1-weighted (T1c), T2-weighted (T2) and T2 Fluid Attenuated Inversion Recovery (FLAIR) scans were provided along with manual consensus delineations of whole tumor (WT), active tumor (AT) and tumor core (TC) tissue. Here we only consider the WT segmentation task. The IS18 dataset contains CT perfusion scans of 94 cases with an acute ischemic stroke. The available perfusion parameter maps are cerebral blood volume (CBV), cerebral blood flow (CBF), mean transit time (MTT), time to peak of the residue function (Tmax) and the raw CT perfusion data. The input volumes were resampled to a 2 mm isotropic voxel size and center patches of size 136×136×82 were used. All available modalities were used and concatenated along the feature dimension as input to the network, except for raw perfusion data for IS18, which was neglected. Because of the bin-size weighting in the ECE metric, the often

highly-confident and accurate background pixels have a large effect. We therefore only consider those voxels belonging to the brain for the ECE calculation, similar to [11].

2.4. U-Net architecture and training

The network architecture used is the modified 3D U-Net-like network from [2]. To speed up convergence the model was pre-trained using CE loss, after which training continued using SD or CE loss. Weights for these models are indicated by CE-SD and CE, respectively. In addition, models were also trained directly on SD loss. We used the same 5-fold cross-validation and training scheme as in [2]. No data-augmentation was used for training of the calibration methods.

Training of Platt scaling and the auxiliary networks was done on batches of 64 z-slices taken from the 3D volumes. The Adam optimizer was used with an initial learning rate of $5 \cdot 10^{-3}$ to optimize the CE. The learning rate decreased with a factor 10 when the loss reached a plateau. Training continued for a maximum of 50 epochs, or until the loss on the validation fold stopped improving.

For fine-tuning, training was done on batches of 2 3D volumes. The Adam optimizer, loss function, learning rate schedule, and early stopping are used as before, but the initial learning rate was 10^{-4} for models trained with CE loss and 10^{-3} for models trained on SD loss.

Because initially no dropout layers were used to train the base models, they were inserted to the base models later. We then retrain all layers after the inserted dropout layers to ensure the dropout does not wreck the input distributions of the subsequent layers. Two dropout settings used in previous literature are considered [20]: In the first setting, referred to as MC-Decoder, dropout is added before each decoder block of the network and before the last convolutional output. The second setting, referred to as MC-Center, applies dropout after the lowest two encoder blocks and before the lowest two decoder blocks. Retraining of the layers after dropout was done with SD for models using SD-based model weights, and CE for models using the CE-based model weights. There was no single learning rate setting that produced good results across dataset, weight or dropout settings. We therefore trained each model using three different initial learning rate settings of 10^{-3}, 10^{-4} and 10^{-5} and report on the best performing model. A learning rate schedule together with early stopping were used. Not training after adding dropout typically lowered calibration compared to the base model. Training always resulted in better performance.

3. RESULTS AND DISCUSSION

We calculated the mean of the per-volume Dice and ECE of every subject in the validation folds. Tables 1 and 2 show the results for the base model and the four calibration methods on the BR18 and IS18 datasets. As expected, the results show a superior segmentation performance in terms of Dice score for SD-based methods. Surprisingly, on BR18 the calibration of CE-based models is not better compared to SD-based models.

Table 1. Results on BR18. Values in bold indicate the best result, or not significantly different from the best result (Wilcoxon signed-rank test $p < 0.05$).

weights→	CE	CE-SD	SD	CE	CE-SD	SD
method	Dice			ECE%		
base model	0.848	**0.870**	**0.878**	1.765	1.429	1.367
Platt	0.851	**0.869**	0.877	1.246	1.186	1.087
auxiliary	0.851	**0.869**	0.877	1.224	1.158	1.067
fine-tune	0.851	0.869	0.876	1.269	**1.141**	**1.028**
MC-Decoders	0.854	0.863	0.873	1.231	1.323	1.247
MC-Center	**0.855**	**0.867**	0.875	**1.180**	1.408	1.308

Table 2. Results on IS18. Values in bold indicate the best result, or not significantly different from the best result (Wilcoxon signed-rank test $p < 0.05$).

weights→	CE	CE-SD	SD	CE	CE-SD	SD
method	Dice			ECE%		
base model	0.462	**0.538**	**0.527**	1.902	2.765	2.981
Platt	0.200	0.499	0.407	4.173	1.917	2.248
auxiliary	0.451	0.509	0.466	1.800	2.076	**2.116**
fine-tune	**0.467**	0.523	0.491	1.808	**1.822**	2.191
MC-Decoder	0.463	0.515	0.498	1.711	2.352	2.312
MC-Center	0.455	0.519	0.497	**1.694**	1.985	2.305

On BR18, all methods improve the ECE over the baseline without appreciably lowering the Dice score. In fact, there is a slight increase in the Dice score for the CE model. Fine-tuning turns out to be superior for SD-based models, although the results are close to both Platt scaling and auxiliary networks. The MC-Center method turns out to be superior for CE-based models for both Dice and ECE. However, MC-Center has the highest ECE for all calibration methods for the CE-SD and SD model. The MC-Decoders method consistently scores amongst the worst calibration methods on BR18.

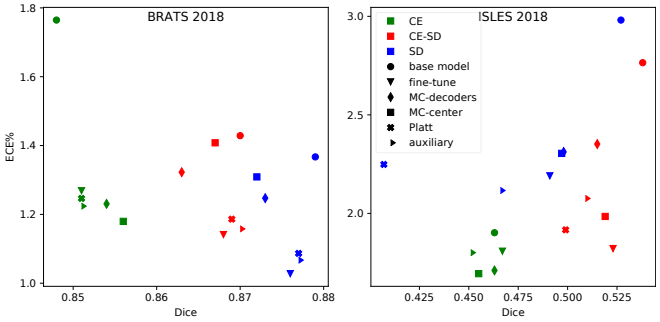


Fig. 1. Scatter plots of ECE% versus Dice score.

Figure 1 shows scatter plots of ECE against Dice. For the SD trained weights we observe that the MC methods are Pareto-dominated (ECE and Dice can be improved by choosing another method), and fine-tuning and auxiliary networks are both Pareto efficient. Only SD trained models are on the Pareto frontier.

Results on IS18 show a different picture. The segmentation of ischemic stroke lesions from CT data is more challenging as MRI is considered to be the gold standard for lesion annotation. This means there is significant aleatoric uncertainty present in the IS18 dataset. All but one calibration methods lower the Dice score significantly, with Platt scaling having the largest negative impact. If each network is considered separately, fine-tuning is always Pareto efficient as can be seen on the scatterplot in Figure 1. In contrast to the results on BR18, no networks trained with SD are on the Pareto frontier. For the CE trained weights, the MC methods performs well: both are Pareto optimal on IS18.

Overall, we can conclude that simple post hoc methods have similar performance compared to MC dropout. Models trained on CE seem more suited for MC dropout. In contrast to previous results [3, 4, 5], we observe that networks trained with SD can be surprisingly well-calibrated, as observed on BR18. On IS18, the calibration of networks trained with SD is significantly worse. This highlights the strong dependency on the dataset for calibration results.

It was noted by Jungo et al. [11] that good average calibration doesn't necessarily translate to good subject-level calibrations. Although their model was well-calibrated on average, a significant number of subjects were under- or overconfident. We examined the distribution of the accuracy in each bin and find that their standard deviations and interquartile ranges do not change much after calibration. The calibration plots in figure 2 show that the difference in mean accuracy and confidence of each bin decreases, and as such the plot moves closer to the ideal line after calibration for SD weights. With CE weights, this shift is not apparent on the plot. As the standard deviations of the bins stays the same, we can conclude that calibration has no adverse effects on the amount of under- or overconfident subjects. We further investigate the subject-level quality by looking at the subject-level distribution over several bins in the calibration curve, as shown in figure 3. Despite improving on average, the amount of outliers doesn't considerably change.

4. CONCLUSION

Neural networks for segmentation are often trained to achieve optimal performance on metrics such as Dice and Jaccard, while less attention is given to their uncertainty calibration. We investigated post hoc methods that calibrate models without needing to retrain them. There are large differences in calibration performance depending on the segmentation task. We find that simple post hoc methods are competitive with the popular MC dropout. MC dropout is most efficient for models trained with cross-entropy. Surprisingly, models trained with soft Dice are better calibrated than those trained with cross-entropy for BraTS 2018. Consistent with previous results [11], we find that despite an improvement in average calibration performance, several subjects remain under- or overconfident. Using segmentation maps as feedback for uncertainty should therefore be done with caution.

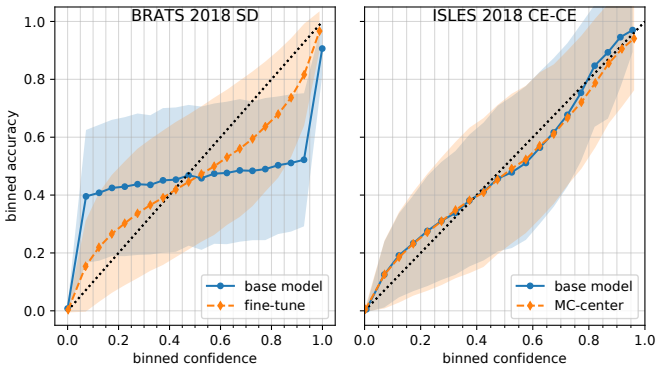


Fig. 2. Reliability diagrams. For each bin the accuracy (points) and standard deviation of the volumes (shaded area) is shown.

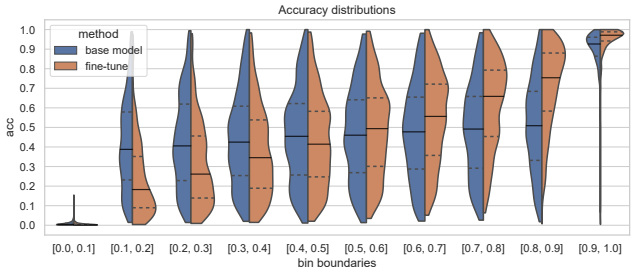


Fig. 3. Violin plots of the subject-level accuracy distributions of the bins, for the BR18 SD model. Median (full lines) and first and third quartiles (dashed lines) are shown.

Acknowledgments

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI. The authors have no relevant financial or non-financial interests to disclose.

Compliance with Ethical Standards

This research study was conducted retrospectively using human subject data made available in open access at `https://ipp.cbica.upenn.edu/` for BraTS 2018 and `https://www.smir.ch/ISLES/Start2018` for ISLES 2018. Ethical approval was not required as confirmed by the license attached with the open access data.

5. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI 2015*, Cham, 2015, pp. 234–241, Springer International Publishing.

[2] Tom Eelbode, Jeroen Bertels, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B. Blaschko, "Optimization for Medical Image Segmentation: Theory and Practice when evaluating with Dice Score or Jaccard Index," *IEEE Trans. Med. Imaging*, pp. 1–1, jun 2020.

[3] Jeroen Bertels, David Robben, Dirk Vandermeulen, and Paul Suetens, "Theoretical analysis and experimental validation of volume bias of soft dice optimized segmentation maps in the context of inherent uncertainty," *Medical Image Analysis*, vol. 67, pp. 101833, 2021.

[4] Jörg Sander, Bob D. de Vos, Jelmer M. Wolterink, and Ivana Išgum, "Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI," in *Medical Imaging 2019: Image Processing*. sep 2019, vol. 10949, p. 44, SPIE-Intl Soc Optical Eng.

[5] Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur, "Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation," *IEEE Trans. Med. Imaging*, pp. 1–1, jul 2020.

[6] Radford M Neal, *Bayesian learning for neural networks*, vol. 118, Springer Science & Business Media, 2012.

[7] Yarin Gal and Zoubin Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," *ICML 2016*, vol. 3, pp. 1651–1660, 2016.

[8] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," *NeurIPS*, vol. 2017-Decem, pp. 6403–6414, dec 2016.

[9] Alex Kendall and Yarin Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *NeurIPS*, 2017, vol. 2017-Decem, pp. 5575–5585.

[10] Terrance DeVries and Graham W Taylor, "Leveraging uncertainty estimates for predicting segmentation quality," *arXiv preprint arXiv:1807.00502*, 2018.

[11] Alain Jungo, Fabian Balsiger, and Mauricio Reyes, "Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation," *Front. Neurosci.*, vol. 14, apr 2020.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, "On calibration of modern neural networks," *ICML 2017*, vol. 3, pp. 2130–2143, jun 2017.

[13] Davood Karimi, Qi Zeng, Prateek Mathur, Apeksha Avinash, Sara Mahdavi, Ingrid Spadinger, Purang Abolmaesumi, and Septimiu E. Salcudean, "Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images," *Medical Image Analysis*, vol. 57, pp. 186–196, oct 2019.

[14] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht, "Obtaining well calibrated probabilities using Bayesian Binning," in *Proceedings of the National Conference on Artificial Intelligence*, 2015, vol. 4, pp. 2901–2907.

[15] John C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *Advances in Large Margin Classifiers*, pp. 61–74, 1999.

[16] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, 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*, 2019.

[17] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, 2017.

[18] Bjoern H. Menze, Andras Jakab, Stefan Bauer, et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[19] Oskar Maier, Bjoern H. Menze, Janina von der Gablentz, Levin Häni, et al., "ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI," *Medical Image Analysis*, vol. 35, pp. 250–269, jan 2017.

[20] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla, "Bayesian segnet: Model uncertainty in deep

convolutional encoder-decoder architectures for scene understanding," in *BMVC 2017*. nov 2017, BMVA Press.