Estimating Uncertainty in Neural Networks for Segmentation Quality Control

Matthew Ng
Sunnybrook Research Institute
Toronto, ON M4N 3M5
matthewng.ng@mail.utoronto.ca

Fumin Guo
Sunnybrook Research Institute
Toronto, ON M4N 3M5
fumin.guo@sri.utoronto.ca

Labonny Biswas
Sunnybrook Research Institute
Toronto, ON M4N 3M5
lbiswas@sri.utoronto.ca

Graham A. Wright
Sunnybrook Research Institute
Toronto, ON M4N 3M5
gawright@sri.utoronto.ca

Abstract

Modelling uncertainty in neural networks is an important task in an automated image segmentation pipeline. In this work, we compared uncertainty estimates obtained using Monte Carlo (MC) Dropout and Bayes by Backprop (BBB) on a U-Net for cardiac MRI segmentation. We also showed a practical application of uncertainty measures in detecting inaccurate segmentation.

1 Introduction

Neural networks have been shown to perform well for automatic cardiac MR image segmentation [Bai et al., 2018, Bernard et al., 2018]. However, when using these methods in an automated image analysis pipeline, it is important to know which segmentation results are problematic and require further manual inspection. This may reduce segmentation errors for downstream analysis.

A few methods have been proposed to directly predict cardiac MR image segmentation quality using machine learning techniques. For example, Robinson et al. [2018] used a 3D residual network to directly predict the Dice score of a predicted segmentation. However, these methods add another black-box on top of the automated segmentation. Another approach is to look at model uncertainty. While uncertainty is not the same as accuracy, a model with well calibrated uncertainties would mean that segmentation outputs with low uncertainty are *likely* correct while outputs with high uncertainty are *likely* problematic. In terms of quality control, identifying segmentations with high uncertainty and correcting these cases with manual segmentation may lead to lower segmentation errors.

Several papers have explored segmentation uncertainty in medical images using MC Dropout to approximate Bayesian neural networks [Roy et al., 2018, Leibig et al., 2017]. However, there are some limitations with this method. For example, when using a constant dropout rate, the model uncertainty does not decrease as more data is observed [Osband, 2016] and the dropout rate needs to be tuned depending on model size and number of data points [Gal, 2016]. Other approaches to approximate Bayesian neural networks include Concrete Dropout and Bayes By Backprop; however, these methods have not yet been explored in medical imaging. In this paper, we compared two methods for estimating uncertainty - MC Dropout and Bayes by Backprop - in the context of cardiac MR image segmentation. In addition, we explored the use of uncertainty measures derived from these methods for detecting inaccurate segmentation.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

2 Methods

Bayesian Neural Networks Bayesian neural networks (BNNs) provide a theoretical framework for capturing model uncertainty. In BNNs, we would like to calculate a posterior distribution of weights, $p(\mathbf{w}|\mathbf{X}, \mathbf{Y})$ instead of a maximum likelihood or maximum-a-posteriori estimate of $\mathbf{w}$. Variational inference is a scalable technique that aims to learn an approximate posterior distribution of the weights, $q(\mathbf{w})$, by minimizing the KL divergence between the approximate and true posterior. This is equivalent to maximizing the evidence lower bound (ELBO): $\mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{Y}|\mathbf{X}, \mathbf{w})] - \mathrm{KL}[q(\mathbf{w})||p(\mathbf{w})]$ where $p(\mathbf{w})$ is the prior distribution of the weights. The first term is the data-driven term while the second term can be viewed as a regularizer. For classification problems, the log likelihood or $\log p(\mathbf{Y}|\mathbf{X}, \mathbf{w})$ is equivalent to negative cross-entropy.

Bayes by Backprop (BBB) A simple way to parameterize the posterior distribution of the weights is to use a fully factorized Gaussian and perform gradient updates using the "reparameterization trick". Each weight in the neural network is drawn independently from a Gaussian distribution with mean μ and standard deviation σ which is parameterized by $\mathrm{softplus}(\rho)$. The training procedure, known as Bayes by Backprop (BBB) [Blundell et al., 2015], is as follows:

1. Sample $\epsilon \sim \mathcal{N}(0, I)$. Then, set $w = \mu + \mathrm{softplus}(\rho) \circ \epsilon$
2. Calculate the loss function (-ELBO): $\mathcal{L} = \text{cross-entropy} + \alpha \mathrm{KL}[(q(\mathbf{w})||\mathcal{N}(\mu_{\mathbf{prior}}, \sigma_{\mathbf{prior}}\mathbf{I})]$
3. Update all parameters, μ and ρ, with a gradient descent optimizer (e.g., Adam)

MC Dropout MC Dropout [Gal and Ghahramani, 2016] is a commonly used method because it is easy to implement and does not require additional parameters or weights. This can be interpreted as choosing the posterior distribution $q(\mathbf{w})$ to be a mixture of two Gaussians with very small variances, one at 0 and the other at the weight. Dropout is applied during training and testing in order to obtain segmentation samples.

Dataset We used short-axis b-SSFP cine MR images from the UK Biobank dataset and trained models for the segmentation of the left ventricle blood pool (LV), left ventricle myocardium (Myo) and right ventricle (RV). 156, 103, and 569 subjects were used for training, validation, and testing, respectively. Each subject has, on average, 20 images slices.

Bayesian Segmentation Network We used a basic 2D U-Net [Ronneberger et al., 2015] with either MC Dropout or BBB. The basic U-Net consists of 10 layers with 3x3 filters and 2 layers with 1x1 convolutions followed by a softmax layer. The number of filters ranges from 32 to 512. In both methods, the final prediction was obtained by averaging the softmax probabilities of 50 samples.

For MC Dropout, we experimented with adding dropout on all layers or only on the central layers with different dropout rates: 0.5, 0.3, 0.1. These settings effectively tune the amount of uncertainty in the model. For BBB, we experimented with different standard deviations of the prior distribution: $\sigma_{prior} =$: 0.1, 1.0, 10, 30 and different coefficients for the prior term: $\alpha = 0.1$ or 1.0. We used the Dice coefficient and average symmetric surface distance (ASSD) to compare the quality of the segmentation and evaluate the average per-pixel negative log likelihood and calibration plots to compare the uncertainty estimates.

Structural Uncertainty Measures Similar to Roy et al. [2018], we defined two structural uncertainty measures as follows:

1. $\mathrm{Dice}_{\text{MeanToSamples}} = \text{Mean}\left(\left\{\mathrm{Dice}(\bar{S}, S_i)\right\}_{i=1\ldots T}\right)$
2. $\mathrm{ASSD}_{\text{MeanToSamples}} = \text{Mean}\left(\left\{\mathrm{ASSD}(\bar{S}, S_i)\right\}_{i=1\ldots T}\right)$

where $\bar{S}$ is the mean predicted segmentation and $S_i, i \in \{1 \ldots T\}$, are predicted segmentation samples from the neural network. We use the standard definitions of the Dice coefficient and ASSD [Bai et al., 2018] except for cases where one of the segmentations is blank (the structure is not present in the slice). The Dice coefficient was set to 1 when both segmentations are blank and 0 when one of the segmentations is blank. ASSD was set to 0 when both segmentations are blank and to the average diameter of the non-blank segmentation when exactly one segmentation is blank.

3 Results

Training Time UNet-BBB has twice as many parameters as UNet-MCDropout and requires 1.5 - 2x the amount of time for training. Both methods require similar time for inference.

Segmentation Performance and Uncertainty Estimates For each method, we report test performance of the model which gave the best validation log likelihood. Among the MC Dropout models, adding dropout on the central layers with a dropout rate of 0.5 performed the best. For BBB, $\alpha = 30$ with $\sigma_{prior} = 1.0$ performed the best. Table 1 shows that UNet-MCDropout and UNet-BBB performed equally well in terms of the Dice coefficient, ASSD, and test log likelihood. Both methods also have excellent calibration based on plots of confidence *vs* accuracy (not shown here). Variance of the segmentation probability maps was observed to be higher around the edges of the ventricles and near the base and apex of the heart where segmentation is poor.

Table 1: Segmentation performance of the U-Net with MC Dropout or BBB. ↑ indicates higher is better. ↓ indicates lower is better. Format: Mean (Standard Deviation)

	Dice ↑			ASSD (mm) ↓			Test Log Likelihood ↑
	LV	Myo	RV	LV	Myo	RV	($\times 10^{-3}$)
UNet-MCDropout	0.938 (0.038)	0.875 (0.032)	0.899 (0.045)	1.05 (0.38)	1.08 (0.34)	1.76 (0.71)	-4.80 (1.70)
UNet-BBB	0.937 (0.040)	0.872 (0.031)	0.898 (0.044)	1.07 (0.42)	1.08 (0.31)	1.77 (0.70)	-4.88 (1.56)

Segmentation Quality Control For each method, we considered the predicted segmentation to be poor when True Dice < 0.85 or True ASSD > 1.5 mm. These numbers are loosely based on the inter-observer variability reported in Bai et al. [2018]. We then calculated the uncertainty measures, $\text{Dice}_{\text{MeanToSamples}}$ and $\text{ASSD}_{\text{MeanToSamples}}$, using the network prediction samples alone and evaluated how well these could identify poor segmentation.

Figure 1 shows the relationship between the number of images with poor segmentation remaining in the dataset and the number of images flagged for manual correction as we change the uncertainty threshold, i.e., (positives - true positives) *vs* (true positives + false positives) where positive represents poor segmentation. As we restrict the predictions to be the ones in which we are more certain, we flag more images for manual correction and the number of images with poor segmentations decreases. The ideal curve for these plots would be towards the bottom left. Figure 1 shows that the two methods are comparable. MC Dropout is better than BBB in terms of average precision for detection of poor segmentation.

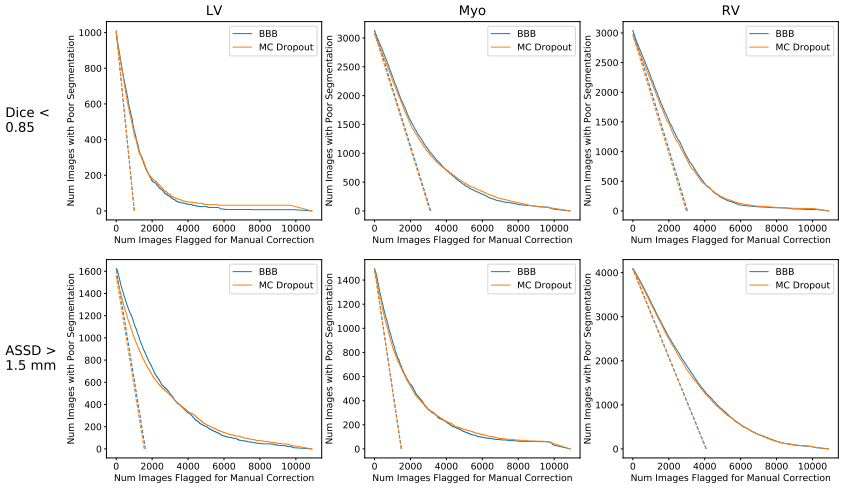

Figure 1: Number of images with poor segmentation remaining after flagging images for manual correction. Rows: Different criteria for poor segmentation. Columns: Different structures. Dashed line represents ideal curve.

4 Conclusions

In this work, we showed that MC Dropout and BBB demonstrated similar performance in a U-Net for cardiac MRI segmentation. Uncertainty measures derived from either method may be used in detecting inaccurate segmentation. Having the ability to know when a segmentation is inaccurate is useful to reduce downstream errors.

Acknowledgments

The authors would like to to thank Guodong Zhang, David Duvenaud, Anne Martel, and Nilesh Ghugre for helpful discussions. This project was funded by the Federal Economic Development Agency for Southern Ontario (FedDev Ontario) and the Canadian Institutes of Health Research (CIHR). This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada. This research has been conducted using the UK Biobank Resource, Application 2964.

References

Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65, 2018.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 2018.

Robert Robinson, Ozan Oktay, Wenjia Bai, Vanya V. Valindria, Mihir M. Sanghvi, Nay Aung, José M. Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Bernhard Kainz, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Chris Page, Daniel Rueckert, and Ben Glocker. Real-time prediction of segmentation quality. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–585. Springer, 2018.

Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Inherent brain segmentation quality control from fully convnet monte carlo sampling. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer, 2018.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):17816, 2017.

Ian Osband. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *Proceedings of the NIPS* 2016 Workshop on Bayesian Deep Learning*, 2016.

Yarin Gal. *Uncertainty in deep learning*. PhD thesis, University of Cambridge, 2016.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.