Paper: *On Calibration of Modern Neural Networks*

Nirajan Koirala

CSC 9010_001

Summary

Guo et al. carried out experiments to find the most important factors influencing calibration for the classification models. They evaluate the performance of various post-processing calibration methods on state-of-the-art architectures with image and document classification datasets. Authors assert that the neural networks today despite being more accurate, are less well-calibrated than older networks. In figure 1 they show how the 1998 LeNet model's calibration level is superior to the 2016 ResNet model even though ResNet has significantly higher accuracy levels. In section 3 they identify some recent changes that are responsible for the miscalibration phenomenon and even though they don't claim causality they find that increased model capacity and lack of regularization to be closely related to model miscalibration.

They point that normalization techniques have enabled the developmental of very deep architectures and techniques like batch normalization improves training time, reduces the need for additional regularization and in some cases even improves the accuracy of the networks. It is observed that batch normalization also negatively affects the calibration regardless of the hyper-parameters used on the batch normalization model. Weight decay, which is also one the predominant regularization techniques is found to negatively affect the calibration if implemented with low decay values. Also, negative log likelihood is found to be a useful metric to indirectly measure model calibration. In section 4, they review existing calibration methods as well as introduce some new variants of their own.

For calibrating the binary models, they use histogram binning, isotonic regression and platt scaling and provide their description in section 4.1. For models dealing with more than 2 classes, one of the methods they use is the extension of binary histogram binning. Other methods they use for multi-class models are matrix and vector scaling and temperature scaling. They apply the discussed calibration methods to image classification and document classification neural networks. 6 different kinds of datasets are used and trained using various state-of-the-art CNNs. When analyzing the calibration results it is found that most datasets and models experience some degree of miscalibration with the ECE usually between 4-10%. However, this miscalibration is not architecture specific and SVHN and Reuters experience very low ECE values. Surprisingly they found that temperature scaling outperforms all other models on vision tasks and performs comparably to other methods on NLP datasets. They found that network miscalibration is intrinsically low dimensional.

Hence, their studies show that recent advances in neural network architecture and training like normalization and regularization have strong effects on network calibration. They also present simple techniques to effectively remedy the miscalibration and temperature scaling is found to be the simplest, fastest and most straightforward of all the techniques they use.

Strengths

- A very important study for evaluating the credibility of model's prediction is using confidence calibration is done in the paper.
- Authors have used six different type of datasets are used for evaluation.

Weaknesses

- I think a more thorough elaboration of results in table 1 could have been done.

Points of Confusion

- I am not sure why there is a slight uptick of calibration at the end of the graph of weight decay in figure 2.
- How does BBQ differs from histogram binning?

Discussion Questions

- Was there any specific reason for them for using CNNs instead of any other type of neural networks for their evaluations?
- What are some of the fields in which machine learning is heavily deployed, calibration metrics could be the most useful?
- Discussion on section 4.2.