Jenish Maharjan **Week 12**

Paper: *On Calibration of Modern Neural Networks*

Summary:

Guo et al study the issue of confidence calibration which is an important issue for classification models in many applications. Due to factors such as depth, width, weight decay and batch normalization, modern neural networks have been demonstrated to be poorly calibrated unlike the networks from a decade ago. The authors describe and evaluate the performance of various post-processing calibration methods on state-of-the-art architectures with image and document classification datasets. Their study presents an insight into neural network learning and also provides a simple method of calibration for practical settings.

The accuracy of neural networks has improved drastically with recent advances in deep learning. This has led to the usage such neural networks in many complex-decision-making applications. Such decision-making systems need to be accurate and also be able to indicate when they are likely to be incorrect. In instances like a self-driving car or medical diagnosis, it is important for the control to be passed on to human experts when the confidence of a classifier is low. And it is very important for the probability associated with the predicted class label should reflect its ground truth correctness likelihood. In other words, the classifiers should provide a calibrated confidence measure in addition to its prediction. While studies from a decade ago showed that neural networks typically produced well-calibrated probabilities on binary classification tasks, the modern neural networks are no longer well-calibrated. A well calibrated classifier basically means if there are 100 predictions each with confidence of x/100, it is expected that x predictions are correct.

A really good tool to check for the calibration of a classifier is a reliability diagram. A reliability diagram plots expected accuracy and the average confidence showing the gap between the output confidence and the expected probabilities. Expected Calibration Error (ECE) makes it convenient to have a scalar summary statistic of calibration. ECE is the weighted average of difference in expectation between confidence and accuracy partitioned into equally-spaced bins. Maximum Calibration Error (MCE) estimates the worst-case deviation between confidence and accuracy. A standard measure of a probabilistic model's quality is the Negative log likelihood (NLL), also known as cross entropy loss. The authors discuss that the miscalibration of models are closely related to model capacity and lack of regularization. The increase in depth and width while reducing classification error has been observed to have negative effects on model calibration. Batch Normalization has also been observed to improved the optimization of neural networks by minimizing distribution shifts in activation within the neural network's hidden layers but also have negative effect in calibration. Weight decay, which is decreasingly utilized when training modern neural networks, is a regularization mechanism. The use of a very small weight decay or none at all is another cause of miscalibration.

The authors have presented calibration methods for binary classification models and their extensions for multi-class models. The method of Histogram binning is a simple non-parametric calibration method in which all uncalibrated predictions are divided into mutually exclusive bins and assigned a calibrated score. Another non-parametric calibration method, Isotonic regression, learns a piecewise constant function to transform uncalibrated outputs. It produces the function to minimize the square loss. An extension to histogram binning is Bayesian Binning into Quantiles (BBQ) which applies Bayesian model averaging to histogram binning. A parametric approach to calibration is Platt scaling that trains on the validation set to

learn the best parameter to return the probabilities. Parameters for Platt scaling can be optimized using the NLL loss over the validation set. For problems with multiple classes, the authors present three methods. Extension of binning methods treats the problem as K one-versus all problems. Matrix and vector scaling are two multi-class extensions of Platt scaling where the logits vectors produced before a linear transformation to the logits. Temperature scaling is another extension of Platt scaling where a single scalar parameter is learned for all classes by minimizing the NLL on the validation set. The learned parameter known as temperature "softens" the softmax output.

In the studies performed by the authors, the various calibration methods are implemented with state-of-the-art classification models for image classification and NLP tasks using various datasets. The results from these studies shows that temperature scaling is the most effective calibration technique. This study successfully demonstrates the aforementioned factors on model calibration and also compare various model calibration techniques.

Strengths:

- They have experiments and results backing the claims they make about some factors having negative effects on model calibration.
- The study provides a good idea of how various calibration techniques compare with each other.
- The content of the paper is well structured.

Weaknesses:

- They could have expanded explanations on each of the calibration techniques with more details.

Confusions:

- How is a reliability diagram plotted?
- The concept of Bayesian Binning into Quantiles is a bit confusing to me.

Discussion Questions:

- What are other tools like Reliability Diagrams that can be used to demonstrate model calibration?
- How does Batch Normalization affect model calibration?
- Since temperature scaling uses NLL/cross entropy to learn the temperature parameter, can the temperature parameter be learned during training by modifying the softmax function?