On Calibration of Modern Neural Networks

ICML 2017

Summary

The authors of this paper address the issue of miscalibration in modern neural networks. Miscalibration in the realm of deep learning-based classification is defined as the gap between the output probability of the network and the likelihood of the prediction's correctness. In other words, does the output of the network indeed correspond to the probability of a correct prediction? The results presented indicate that popular modifications in modern deep learning architectures lead to increased miscalibration despite considerably higher classification scores. In order to counteract miscalibration, *Temperature Scaling*, a simple post-processing technique is proposed that improves the network calibration without influencing the classification score.

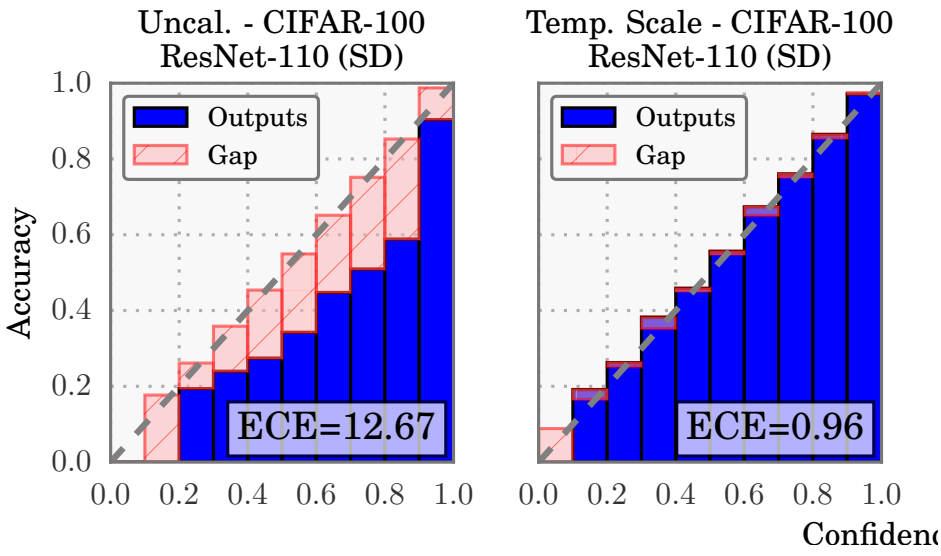

Figure 1: Reliability diagram of a ResNet-110 evaluated on CIFAR-100 with and without Temperature Scaling. The uncalibrated network outputs are misaligned with their likelihood of correctness which results in a large *Expected Calibration Error* (ECE). The proposed calibration method counteracts the misalignment and produces reliable confidence scores.

Main Contributions

- **Factors of Miscalibration** The authors present an investigation of the influence of several parameters such as model capacity, batch normalization and weight decay on the miscalibration of neural networks.
- **Method Evaluation** Different post-processing calibration methods are evaluated on a variety of image classification dataset.
- **Temperature Scaling** Temperature Scaling is proposed as a new approach to simply and efficiently calibrating neural networks.

Implementation Details

- **Reliability Diagrams** To illustrate the degree of miscalibration of the network, the expected prediction correctness is plotted against the network's confidence. To estimate the expected correctness from a finite number of samples, the confidence scores are assigned to M bins of size 1/M. A perfectly calibrated network corresponds to a diagonal reliability diagram. Deviations from the diagonal indicate over- or under-confident network predictions.
- **Expected Calibration Error** an approximation of ECE is applied to get a scalar measure indicating the degree of miscalibration. The ECE is defined as the expected difference between confidence and accuracy. This measure is approximated by taking the weighted average of each of the bins' confidence gap in the reliability diagram.
- **Temperature Scaling** The proposed calibration method is a simple extension of *Platt Scaling*. A single parameter T (referred to as the temperature) is used to scale the logits output of the network before feeding it to the softmax function. This parameter is optimized on the validation set using a cross entropy loss. The calibrated confidence score for a logits vector $\mathbf{z}_i$ is given by:

$$\hat{q}_i = \max_k \sigma_{\text{SM}}(\mathbf{z}_i/T)^{(k)}$$

Evaluation

- **Observing Miscalibration** The authors evaluate the influence of certain parameters on miscalibration by fixing network and training setup and observing the output for a single variable parameter.
  1. Model Capacity: Both increased width and depth induce higher ECEs. This might stem from the fact that a network can further lower a cross entropy loss by becoming more certain of its predictions.
  2. Batch Normalization: Despite positively influencing the accuracy, BN negatively affects the network's calibration. However, the authors provide no suggestions as to why that is the case.
  3. Weight Decay: While most modern neural networks are trained with little or no weight decay, a higher regularization parameter λ shows to positively affect the network's ECE.

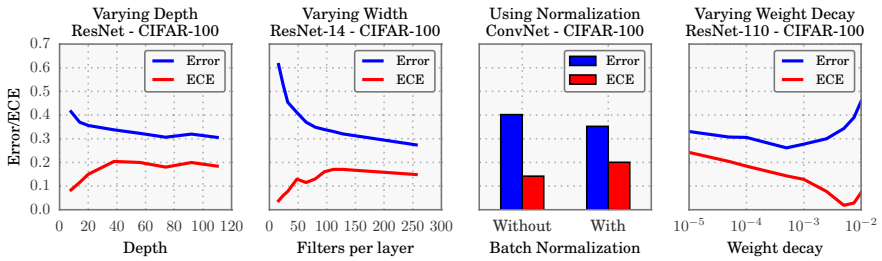

Figure 2: Classification Error and ECE of a ResNet-110 trained on CIFAR-100 as a function of different parameters affecting network and training process.

References

This summary is solely based on my understanding of the original paper. All images used here are taken from the original paper as well. The paper can be found under the following link:
`http://proceedings.mlr.press/v70/guo17a/guo17a.pdf`