On Calibration of Modern Neural Networks

Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger

Summary by Sri Datta (budaraju@kth.se)

The authors of the paper address the problem of calibration in modern deep neural networks. Calibration or Confidence calibration is defined as the prediction of probability that is a close representation of the true correctness likelihood. Deep neural networks are deployed in self-driving cars and the medical industry where being accurate is enough but being able to predict failure or convey the uncertainty (Jiang et al., 2012). That is a network should predict calibrated confidence along with the probability. Proper calibration of the networks is very important. Humans naturally have a cognitive intuition of probability (Cosmides & Tooby, 1996) and calibrated confidence from a neural network would make it more trustworthy and also help in interpretability besides just giving an extra bit of information.

Despite modern neural networks achieving very high accuracies, they are highly miscalibrated as visualized in the Reliability Diagram (DeGroot & Fienberg, 1983; Niculescu-Mizil & Caruana, 2005) (Figure 1), comparing a 5-layer LeNet Network (LeCun et al., 1998) with a 110-layer ResNet Network (He et al., 2016) on the CIFAR-100 dataset. Four major factors namely, width, depth of the network, weight decay, and Batch Normalization contribute to the miscalibration of the networks. If a network gives 100 predictions, each with the confidence of say, 0.7, then 70 of the total predictions should be classified correctly, only then a network is said to be perfectly calibrated.

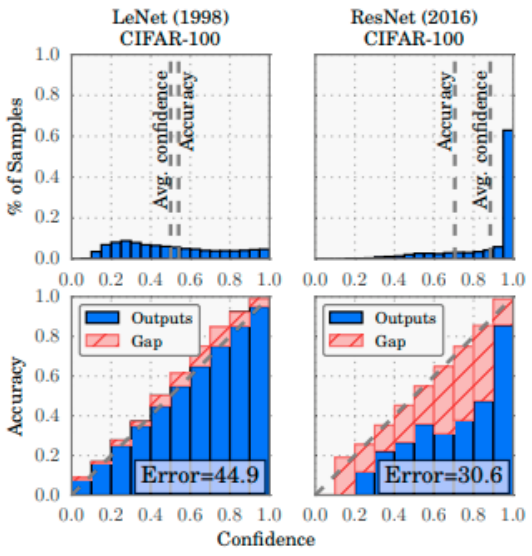

Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

Figure 1. Reliability Diagram | Source: Paper

Expected Calibration Error (ECE) (Naeini et al., 2015) can be used to approximate the miscalibration in a network. The predictions are partitioned into M equally-spaced bins (each bin is a range of prediction probability). Then the average confidence and the

accuracy of each bin are computed. The Expected Calibration Error is the weighted average of the Calibration Gap obtained from the difference between the accuracy and the average confidence of each bin. The procedure is illustrated in Figure 2.

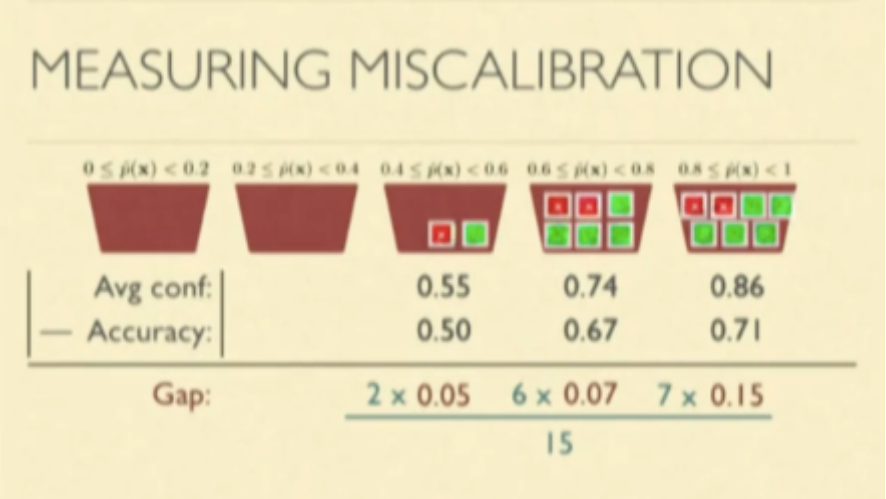

Figure 2. Calculation of ECE | Source: Conference Video

Maximum Calibration Error (MCE) (Naeini et al., 2015) can be used to minimize the worst-case deviation between confidence and accuracy measures. This is very useful for the applications of deep neural networks in high-risk scenarios where reliable confidence measures are crucial. The procedure is the same as the ECE but MCE is the maximum of the absolute difference between the average confidence and the accuracy of each bin. That is the MCE is the maximum calibration gap observed across all the bins where ECE is the weighted average of all calibration gaps. Though perfect calibration is practically impossible, a perfectly calibrated network will make MCE and ECE equal to zero.

Deeper and wider neural networks are observed to generalize the dataset better. Though increasing the neural capacity of a network decreases the classification error, it is observed to increase the miscalibration of the network. Similarly, Batch Normalization which is known to fasten the training process and reduces the need for regularization is also found to contribute to the miscalibration of deep networks. Weight decay, which is another way to regularize and prevent neural networks with capacity from overfitting is also found to contribute to miscalibration. The above-mentioned methods used to improve the performance, capacity, and generalization, have a high effect on miscalibration as illustrated in Figure 3.

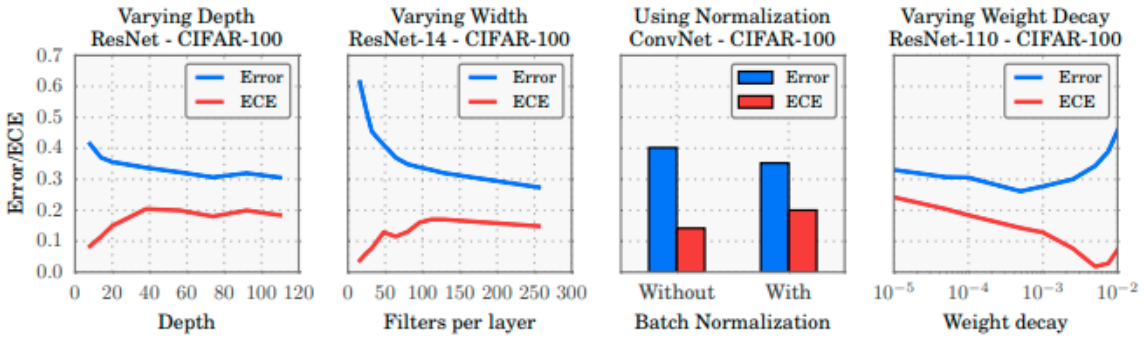

Figure 3: Factor contributing to miscalibration | Source: Paper

Negative log-likelihood or NLL is a standard measure of a probabilistic model's quality (Friedman et al., 2001). In deep learning, it is most frequently referred to as cross-entropy loss (Bengio et al., 2015). It has been noticed that modern deep networks tend to overfit on the negative log-likelihood while still improving on the error as visualized in Figure 4. This phenomenon is not yet completely understood. But the same trend has been particularly noticed in miscalibrated networks. This could explain the problem of miscalibration.

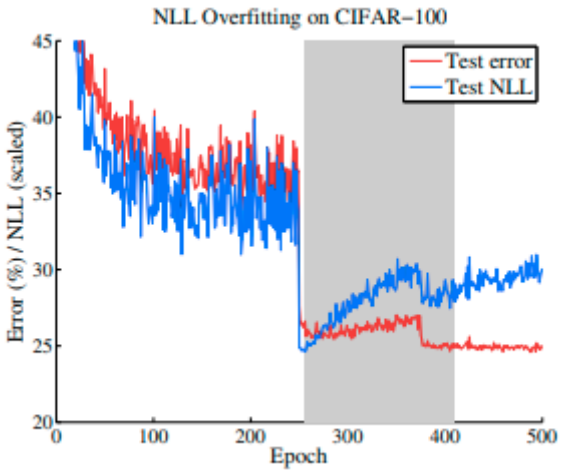

Figure 4: NLL overfitting | Source: Paper

There have been many post-processing methods like Histogram binning, Isotonic regression, Bayesian Binning into Quantiles, and Platt scaling to calibrate neural networks. Temperature scaling is an extension of Platt scaling (Platt et al., 1999). It reduces the miscalibration of the network just by adding another parameter T also referred to as Temperature to soften the Softmax function as illustrated in Figure 5. Temperature is optimized with respect to NLL on the validation set. As T does not change the max of softmax, the class prediction remains the same. That is it does not affect the accuracy. This technique is also used in knowledge distillation.

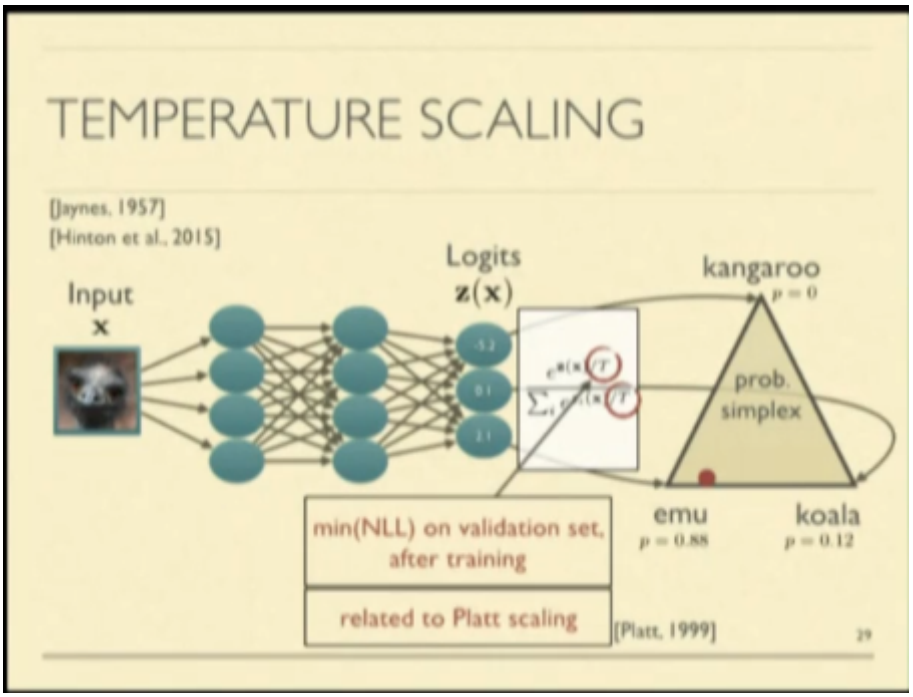

Figure 5. Temperature Scaling | Source: Conference Video

TakeAway

Though the error rates of modern deep networks are decreasing every day, they are being more and more miscalibrated. Important techniques like batch normalization, which increases the accuracy have some mysterious side effects that make networks express high confidence even for the wrong classifications. There are a few techniques like temperature scaling to calibrate the networks to some extent efficiently. However, more effort and study is needed to understand the phenomena.

Strong Points

The authors did a great job in explaining the phenomena of miscalibration and the ways to measure miscalibration. A decent amount of background and a wide range of techniques are discussed in a very intuitive manner. In addition to this, the authors concisely present numerous experiments and results comparing the various methods to tackle miscalibration. The authors also provide an example code for temperature scaling for the readers to try out.

Weak Points

The main weakness of the paper is that the authors only describe the miscalibration in classification tasks. Though they presented numerous experiments on different datasets, It is necessary to evaluate if the same techniques like temperature scaling work on other tasks (like time series or generative models) or if new techniques have to be introduced.

The comparison of the methods shows that temperature scaling performs better only on vision datasets and not NLP datasets. Hence as only temperature scaling does not affect the accuracy and the others do, the comparison of the overall accuracy after calibration is missing.

References

Jiang, Xiaoqian, Osl, Melanie, Kim, Jihoon, and Ohno Machado, Lucila. Calibrating predictive model estimates to support personalized medicine. Journal of the American Medical Informatics Association, 19(2):263–274, 2012.

Cosmides, Leda and Tooby, John. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. cognition, 58(1):1–73, 1996.

LeCun, Yann, Bottou, Leon, Bengio, Yoshua, and Haffner, ´ Patrick. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278– 2324, 1998.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In CVPR, pp. 770–778, 2016.

DeGroot, Morris H and Fienberg, Stephen E. The comparison and evaluation of forecasters. The statistician, pp. 12–22, 1983.

Niculescu-Mizil, Alexandru and Caruana, Rich. Predicting good probabilities with supervised learning. In ICML, pp. 625–632, 2005.

Naeini, Mahdi Pakdaman, Cooper, Gregory F, and Hauskrecht, Milos. Obtaining well calibrated probabilities using bayesian binning. In AAAI, pp. 2901, 2015.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. The elements of statistical learning, volume 1. Springer series in statistics Springer, Berlin, 2001.

Bengio, Yoshua, Goodfellow, Ian J, and Courville, Aaron. Deep learning. Nature, 521:436–444, 2015.
Platt, John et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, 10(3): 61–74, 1999.