Paper Review - Self

By

Debaditya Roy

For the course - **FID3018**

Debaditya Roy

May 2021

Introduction

Neural networks are compelling machine learning methods that are heavily used in application and research. Nevertheless, deterministic neural networks involved in classification tasks fail to provide well-modeled probabilities for each class. To address this issue neural network calibration was reintroduced a couple of years back. Since then, considerable research has spanned that tries to calibrate the probabilities produced from the neural network. In this review, we summarize three such methods.

On calibration of modern neural networks [2]

Motivation

This paper was a pioneer not because it introduced the concept of neural network calibration but it reintroduced the notion of calibration in modern neural networks. The authors highlight the importance of confidence calibration to ensure safety and reliability. They found out that despite producing excellent results, modern neural networks suffer in providing reliable predictions. Empirically they found out certain reasons behind the miscalibration and proposed a simple yet effective recipe called *temperature scaling* that mitigates the issue.

Contributions

There are two main contributions in this paper:

1. Empirical findings of the reasons affecting calibration.
2. Proposing a simple and effective scaling method of calibration (that is a simple extension of an earlier method called *Platt scaling*.
3. Comparisons with some standard calibration measures.

The main factors that influence calibration are:

Model Capacity

With the increased capacity of neural networks, predictions tend to get miscalibrated.

Regularization methods

While the authors have discussed two different factors namely *batch normalization* and *weight decay* separately. I think it will be better to group it as regularization methods. Older regularization methods such as weight decay, l2-regularization perform better in terms of calibration. Modern methods such as *batch normalization* that improve the classification result and results in faster training perform inversely in terms of calibration.

Negative log-likelihood (NLL)

This was perhaps the most exciting observation that identified one of the significant reasons for miscalibration. The paper showed that neural networks trained with log-likelihood overfits NLL, and that improves the classification accuracy. However, overfitting the NLL hampers confidence calibration of the probabilities. Thus modern neural networks improve accuracy at the expense of well-modeled probabilities.

Temperature scaling

To calibrate the probabilities of a trained neural network, the authors adopt an older parametric method known as Platt-scaling [4] and simplify some assumptions of the method. Platt scaling is about optimizing a linear neural network that takes into input the logits outputted by the trained neural network on a validation set. The formulation of Platt scaling is $q_i = \sigma(za + b)$, where $z = logits$, a and b are parameters of the neural network. The paper suggests using a single scaler T instead of a and having $b = 0$. This formulation enables the optimization algorithm to learn a temperature parameter. The learned temperature parameter scales the logits and hence softens the softmax output.

Weaknesses

Although the paper does a great job with merging the important confidence calibration task in mainstream neural networks, there are specific points it fails to address.

- Any theoretical or intuitive explanation does not back the empirical observation of regularization and calibration.
- While temperature scaling is offered as a solution, the evaluation of temperature scaling was done on a binning method. Therefore, the impact of changing the number of bins is important w.r.t calibration error. This was not shown experimentally.
- The paper does not explore the role of softmax in overemphasizing the positive class.

Verified Uncertainty Calibration [3]

Motivation

This paper is selected in the review because it addressed some shortcomings of the previous paper and improved upon them. Parametric methods such as Platt scaling (or variants of them) are used for calibration, but they are not reliable because they depend on histogram-based methods for estimating the calibration error. However, they are sample efficient as they might require fewer points to fit a curve. Histogram-binning, a non-parametric method, can be used because apart from re-calibrating the model, they also estimate calibration error. However, they are not sample-inefficient. The authors combine both the methods to produce *scaling binning calibrator*, that is sample efficient in calibrating the model, and estimate calibration error natively.

Contributions

The primary contribution of this paper is identifying the sample inefficiency of the histogram-based methods and the inability of scaling methods to quantify calibration error. Both these constructs are shown empirically as well as with mathematical proof.

The paper followed a solid theoretical lead to establish that the scaling method that relies on binning techniques to capture the calibration error underestimates the error. Increasing the number of bins unanimously increases the calibration error. The method proposed in this paper offers a certain theoretical bound within which the calibration error is limited. The strength of this paper is that it has shown the mathematical proofs and empirical studies of all the claims it made.

Weaknesses

Like the previous method, this method is also heavily dependent on the validation dataset selection. Although they have provided the size of the validation sets in their experiment, they did not comment on the expected distribution of the validation set required for the optimization. Furthermore, it may suffer from the same convergence problems an optimization algorithm working on a small data set.

Learning of Single-Shot Confidence Calibration in Deep Neural Networks through Stochastic Inferences [5]

Motivation

This paper is discussed in this essay as an alternative to validation set-based post-processing methods. Existing calibration methods optimize a different function and the training loss function on a different validation set to calibrate the neural network. This paper proposes optimizing a loss through stochastic inferences without having a separate validation set for it.

Contribution

The theoretical foundation related to stochastic inference presented in this paper has already been established by Gal et al. [1]. However, the relation between predictive variance, confidence, and accuracy as shown by empirical experiment and backed by the theoretical foundation is novel and central to their primary goal. The authors proposed a custom loss function that adds a term to the ground-truth based cross-entropy loss term (i.e. the standard cross-entropy loss). This additional term is a cross-entropy loss of the prediction with a uniform distribution. The predictive variance of the inference weights both

these terms. The significance of each of these terms is: If the confidence of the prediction is low, then the variance is high and more weight is put on the second term, which makes the prediction closer to the uniform distribution. In the other case, the prediction is more confident and will enhance the ground truth cross-entropy term. Thus, it creates a gap between confident and under-confident examples. The paper shows, validates, and verifies the effectivity of the new loss term experimentally and compares it with temperature scaling.

Weakness

Although the paper compares it with temperature scaling, the accuracy comparison did not account for the validation dataset that temperature scaling-based methods set aside. The paper bypasses the need of validation set by incorporating stochastic inferences and its variance in loss function. However, it shows experimentally that at least four stochastic inferences are needed for most of the cases to calibrate the confidence. It would have been interesting to get a performance comparison of doing these stochastic inference vs. optimization via a validation set. Also, it compares with only one calibration mechanism with its method.

References

[1] Yarin Gal and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[2] Chuan Guo et al. "On calibration of modern neural networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1321–1330.

[3] Ananya Kumar, Percy Liang, and Tengyu Ma. "Verified uncertainty calibration". In: *arXiv preprint arXiv:1909.10155* (2019).

[4] John Platt et al. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74.

[5] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. "Learning for single-shot confidence calibration in deep neural networks through stochastic inferences". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9030–9038.