Revisiting Recalibration

Jonathan Pearce
Department of Computer Science
McGill University
Montreal, QC H3A 0G4
jonathan.pearce@mail.mcgill.ca

Abstract

Modern deep neural networks have been shown to be overconfident. There are many ways to prevent this overconfidence, one popular technique is recalibration which is performed after model training is complete. Popular recalibration methods include temperature scaling, vector scaling and matrix scaling. However, since the initial research into these methods for recalibration was conducted there has been little follow-up analysis on the performance of these methods and potential ways to improve them. We provide three new recalibration experiments in our paper to evaluate temperature scaling, vector scaling and matrix scaling. First, we show that matrix scaling can be improved with regularization such as weight decay. We also compare NLL and focal loss as two functions to optimize these recalibration methods, we find that NLL leads to better results. Finally, we provide a more complete calibration evaluation of these three methods, using recently developed binning calibration metrics, we find that vector scaling is the best of these three recalibration methods.

1 Introduction

The work of Guo et al. [2017] was of one of the primary papers that led to an increase in calibration by binning research. This paper provided two main insights: modern deep neural networks are overconfident when trained using a standard procedure, and there exist recalibration methods that can help correct this overconfidence issue, such as temperature scaling and vector scaling. Since this work has been released there has been a great amount of attention and research drawn towards calibration by binning. Recently, Minderer et al. [2021] demonstrated that the observation, that deep neural networks are overconfident is less pronounced when more modern architectures (that have been introduced since 2017) are assessed for calibration. Our work seeks to revisit the secondary observation of Guo et al. [2017], that recalibration methods can help correct this overconfidence issue. Specifically we attempt to reproduce their findings and extend their methods with recent advancements in calibration research. Our work focuses on three of the methods that were originally evaluated: temperature scaling, vector scaling and matrix scaling. We provide three new experiments in our work. We attempt to fix the overfitting issue that matrix scaling experiences, particularly on the CIFAR-100 dataset. We compare NLL and focal loss as loss functions to optimize these recalibration methods. Finally, we evaluate calibration using expected calibration error (ECE) as well as classwise-ECE to obtain a more complete assessment of each methods calibration ability.

Preprint. Under review.

2 Background

2.1 Temperature scaling

A simple extension of Platt scaling [Platt, 1999] for multiclass models. Temperature scaling uses a single parameter $T > 0$ for all K classes. Let $\mathbf{z_i}$ be the logits vector produced before the softmax (SM) layer for input $\mathbf{x_i}$, the new confidence prediction is

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

Where T is denoted the temperature. With $T > 1$ the output entropy increases. As $T \to \infty$, the probability $\hat{q}_i$ approaches $\frac{1}{K}$, which represents maximum uncertainty. With $T = 1$, the original probability $\hat{p}_i$ is recovered, where $\hat{p}_i = \sigma_{SM}(z_i)$. As $T \to 0$, the probability collapses to a point mass (i.e. $\hat{q}_i = 1$). In our work T is optimized with respect to NLL or focal loss on the validation set.

2.2 Matrix and vector scaling

Two multi-class extensions of Platt scaling. Given the logit vector $\mathbf{z_i}$, matrix scaling applies a linear transformation $\mathbf{Wz_i} + \mathbf{b}$ to the logits:

$$\hat{q}_i = \max_k \sigma_{SM}(\mathbf{Wz_i} + \mathbf{b})^{(k)},$$

$$\hat{y}_i^{'} = \operatorname*{argmax}_k \sigma_{SM}(\mathbf{Wz_i} + \mathbf{b})^{(k)}.$$

Where $\hat{q}_i$ is the confidence prediction and $\hat{y}_i^{'}$ is the class prediction. In our work, the parameters $\mathbf{W}$ and $\mathbf{b}$ are optimized with respect to NLL or focal loss on the validation set. Vector scaling is a variant where $\mathbf{W}$ is restricted to be a diagonal matrix.

2.3 Focal Loss

We utlize focal loss [Lin et al., 2017] as an alternative to NLL for optimizing these recalibration methods. Focal loss is given as,

$$\mathcal{L}(q_i) = -(1 - q_i)^{\gamma}\text{log}(q_i).$$

The intuition behind using focal loss is to direct the network's attention during training towards samples for which it is currently predicting a low probability for the correct class, since trying to reduce the NLL on samples for which it is already predicting a high probability for the correct class is liable to lead to NLL overfitting, and thereby miscalibration. Mukhoti et al. [2020] have shown that using focal loss can help improve model calibration.

2.4 Calibration

2.4.1 Expected Calibration Error

A model is said to be perfectly calibrated when for each sample $(\mathbf{x}, y) \in \mathcal{D}$, the confidence of the model $\hat{p}$ in the class prediction $\hat{y}$ to be a true probability. For example, of all the data samples that a perfectly calibrated model assigns a prediction confidence of 0.8, 80% of those samples will be predicted correctly. This can be expressed more formally,

$$\mathbb{P}(\hat{y} = y|\hat{p} = p) = p. \tag{1}$$

In practical settings, perfect calibration is impossible. Additionally, the probability in (1) cannot be computed using finitely many samples since $\hat{p}$ is a continuous. Therefore empirical approximations are required to capture the essence of (1).

A popular metric used to measure model confidence calibration is the expected calibration error (ECE) [Naeini et al., 2015]. ECE approximates the difference in expectation between model confidence and model accuracy, more formally written as,

$$\mathbb{E}_{\hat{p}}[|\mathbb{P}(\hat{y} = y|\hat{p} = p) - p|]. \tag{2}$$

Due to finite data, ECE cannot in practice be computed using (2). Instead, we group predictions into M interval bins (each of size $1/M$) and calculate the accuracy and confidence of each bin. Let B_m be the set of indices of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$. The accuracy and confidence of B_m are defined as

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}_i = y_i),$$

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i.$$

Where y_i is the true class label for sample i and $\hat{p}_i$ is the model confidence for the class prediction $\hat{y}_i$. ECE is a weighted average of the absolute difference between the accuracy and confidence of each bin,

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|.$$

2.4.2 Classwise-ECE

ECE only considers the probability of the predicted class, which means it does not consider how well calibrated a model is with respect to the $K - 1$ other probabilities that a model outputs. A stronger definition of calibration requires the probabilities of all the classes for every data instance to be calibrated [Kull et al., 2019, Nixon et al., 2019, Widmann et al., 2019, Kumar et al., 2019, Vaicenavicius et al., 2019].

Classwise-ECE is a simple extension of ECE that accounts for all predictions [Kull et al., 2019, Nixon et al., 2019]. For Classwise-ECE, we group predictions by the K classes and then into M interval bins (each of size $1/M$) and calculate the accuracy and confidence of each bin. Let $B_{k,m}$ be the set of indices of samples from class k whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$. The accuracy and confidence of $B_{k,m}$ are defined as

$$\text{acc}(B_{k,m}) = \frac{1}{|B_{k,m}|} \sum_{i \in B_{k,m}} \mathbb{1}(k = y_i),$$

$$\text{conf}(B_{k,m}) = \frac{1}{|B_{k,m}|} \sum_{i \in B_{k,m}} \hat{p}_{i,k}.$$

Where $\hat{p}_{i,k}$ is the model confidence that sample i belongs to class k. Classwise-ECE is a weighted average across all K classes and their M bins, of the absolute difference between the bin accuracy and confidence,

$$\text{Classwise-ECE} = \frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{|B_{k,m}|}{N} |\text{acc}(B_{k,m}) - \text{conf}(B_{k,m})|.$$

3 Results

3.1 Experimental Details

We conduct image classification experiments using the CIFAR-10/100 datasets [Krizhevsky, 2009]. We use a train/validation/test split of 50,000/5,000/5,000 images for both CIFAR-10 and CIFAR-100. We train and evaluate our methods with the ResNet56, VGG11, MobileNet-v2 and ShuffleNet-v2 architectures. We provide a basic benchmark of the trained models with no calibration (NC). We optimize temperature scaling (TS), vector scaling (VS) and matrix scaling (MS) with both NLL and focal loss using the LBFGS optimizer, with a learning rate of 0.01 and 200 iterations. For focal loss we set $\gamma = 1.0$. We optimize matrix scaling with weight decay (MS-WD) with only NLL as this method with focal loss had convergence issues, we utilize stochastic gradient descent with a learning rate of 0.01, 200 iterations and a weight decay of 0.25. ECE and classwise-ECE are calculated with 15 and 50 bins respectively.

Table 1: ECE(%)

			NLL				**Focal Loss**		
Dataset	Method	NC	TS	VS	MS	MS-WD	TS	VS	MS
CIFAR10	ResNet56	3.63	1.90	1.25	1.65	**1.08**	1.64	3.44	2.53
	VGG11	4.90	2.10	**1.32**	2.06	1.33	1.36	3.19	2.75
	MobileNet-v2	3.75	1.55	**1.21**	1.54	1.70	2.54	4.06	3.47
	ShuffleNet-v2	4.67	1.72	1.83	2.07	**1.24**	1.35	3.16	1.61
CIFAR100	ResNet56	14.17	4.85	2.98	9.16	3.19	3.28	**2.77**	6.85
	VGG11	15.18	6.67	4.99	6.10	3.89	5.75	4.23	**3.54**
	MobileNet-v2	9.76	3.86	2.80	7.85	4.74	2.65	**2.20**	5.16
	ShuffleNet-v2	12.36	3.31	**1.45**	11.68	3.23	1.48	2.08	9.90

Table 2: Classwise-ECE(%)

			NLL				**Focal Loss**		
Dataset	Method	NC	TS	VS	MS	MS-WD	TS	VS	MS
CIFAR10	ResNet56	0.97	0.73	**0.67**	0.70	0.75	0.88	1.21	1.03
	VGG11	1.21	0.87	**0.81**	0.83	0.88	0.97	1.31	1.11
	MobileNet-v2	0.95	0.75	**0.68**	0.69	0.85	1.02	1.33	1.17
	ShuffleNet-v2	1.27	0.97	**0.95**	**0.95**	1.03	1.06	1.19	1.00
CIFAR100	ResNet56	0.41	**0.35**	0.37	0.42	0.38	0.38	0.38	0.39
	VGG11	0.44	0.34	**0.32**	0.37	0.35	0.36	0.34	0.38
	MobileNet-v2	0.37	0.35	**0.34**	0.37	0.39	0.35	**0.34**	0.36
	ShuffleNet-v2	0.45	0.40	**0.39**	0.48	0.43	**0.39**	0.40	0.48

3.2 Discussion

With respect to ECE (Table 1) generally our results align with the findings of Guo et al. [2017]. Temperature scaling (TS) and vector scaling (VS) are successful at recalibrating all model architectures across both CIFAR10 and 100 datasets. Matrix scaling (MS) works well for CIFAR10 but appears to overfit the validation set on CIFAR100, leading to poor calibration. This issue is reduced by utilizing weight decay (MS-WD). Additionally, we observe between optimizing with NLL and focal loss there is no clear superior method with respect to ECE, however one advantage with NLL was that MS-WD converged when optimized with NLL, where as with focal loss there were convergence issues. For classwise-ECE (Table 2) the results differ slightly. Vector scaling optimized with NLL is the clear best method, achieving the lowest classwise-ECE in 7 of the 8 model architecture-dataset combinations. More generally methods optimized with NLL appear to have lower classwise-ECE than those optimized with focal loss. The test set error was similar between all methods.

4 Conclusion

In our work we have three key results. We confirm that regular matrix scaling can be improved with regularization such as weight decay. We demonstrate that optimizing these recalibration methods with NLL is preferrred to focal loss. Using both ECE and classwise-ECE we show that vector scaling optimized with NLL appears to be the most consistent method for recalibration.

There are many avenues for future work in this area. Immediate investigations can be conducted for the lack of convergence with matrix scaling weight decay when optimized with focal loss. Sensitivity analyses can be pursued to study the effect of the weight decay parameter on the performance of matrix scaling and similarly how the value of gamma in focal loss effects calibration.

References

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. In *Advances in Neural Information Processing Systems*, 2019.

Ananya Kumar, Percy S Liang, and Tengyu Ma. In *Advances in Neural Information Processing Systems*, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc., 2021.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 1999.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

David Widmann, Fredrik Lindsten, and Dave Zachariah. In *Advances in Neural Information Processing Systems*, 2019.