Measuring Calibration in Deep Learning

Jeremy Nixon
Google Brain
jeremynixon@google.com

Michael W. Dusenberry
Google Brain
dusenberrymw@google.com

Linchuan Zhang
Google
linchzhang@google.com

Ghassen Jerfel
Google Brain, Duke University
ghassen@google.com

Dustin Tran
Google Brain
trandustin@google.com

Abstract

The reliability of a machine learning model's confidence in its predictions is critical for high-risk applications. Calibration—the idea that a model's predicted probabilities of outcomes reflect true probabilities of those outcomes—formalizes this notion. Current calibration metrics fail to consider all of the predictions made by machine learning models, and are inefficient in their estimation of the calibration error. We design the Adaptive Calibration Error (ACE) metric to resolve these pathologies and show that it outperforms other metrics, especially in settings where predictions beyond the maximum prediction that is chosen as the output class matter.

1. Introduction

The reliability of a machine learning model's confidence in its predictions is critical for high risk applications, such as deciding whether to trust a medical diagnosis prediction (1; 6; 10). One mathematical formulation of the reliability of confidence is calibration (8; 2). Intuitively, for class predictions, calibration means that if a model assigns a class with 90% probability, that class should appear 90% of the time.

Recent work proposed Expected Calibration Error (ECE; 9), a measure of calibration error which has lead to a surge of works developing methods for calibrated deep neural networks (e.g., 5; 7). In this paper, we show that ECE has numerous pathologies, and that recent calibration methods, which have been shown to successfully recalibrate models according to ECE, cannot be properly evaluated via ECE.

Issues with calibration metrics include: not computing calibration across all predictions, issues coming out of fixed calibration ranges, and an inefficient bias-variance tradeoff. We solve these issues through including all predictions, adaptive calibration ranges, and thresholding.

We identify and examine challenges in measuring calibration and propose several new calibration metrics that are designed to resolve them: Static Calibration Error (SCE), Adaptive Calibration Error (ACE), and Thresholded Adaptive Calibration Error (TACE). We perform experiments across MNIST, Fashion MNIST, CIFAR-10/CIFAR-100, and ImageNet 2012. They indicate, for example, that ECE does not work well when class predictions beyond the maximum prediction matter more, and so our ability to evaluate recalibration methods such as temperature scaling suffers. TACE, ACE and SCE and other more flexible calibration metrics are more robust.

In general, we recommend the use of Adaptive Calibration Error in calibrating a model. Our recommendation is grounded in experiments demonstrating that the evaluation of calibration is more effective with ACE and TACE (Figure 1, Table 2) than the others. TACE should be used in settings where the class count is high (ex., 100+ in our experiments) to use bins more efficiently and compute a score that focuses on likely predictions.

2. Background & Related Work

2.1. Measurement of Calibration

Assume the dataset of features and outcomes $\{(x, y)\}$ are i.i.d. realizations of the random variables $X, Y \sim \mathbb{P}$. We focus on class predictions. Suppose a model predicts a class y with probability $\hat{p}$. The model is *calibrated* if $\hat{p}$ is always the true probability. Formally,

$$\mathbb{P}(Y = y \mid \hat{p} = p) = p$$

for all probability values $p \in [0, 1]$ and class labels $y \in \{0, \ldots, K-1\}$. The left-hand-side denotes the true data distribution's probability of a label given that the model predicts $\hat{p} = p$; the right-hand-side denotes that

value. Any difference between the left and right sides for a given p is known as *calibration error*.

Expected Calibration Error (ECE). To approximately measure the calibration error in expectation, ECE discretizes the probability interval into a fixed number of bins, and assigns each predicted probability to the bin that encompasses it. The calibration error is the difference between the fraction of predictions in the bin that are correct (accuracy) and the mean of the probabilities in the bin (confidence). Intuitively, the accuracy estimates $\mathbb{P}(Y = y \mid \hat{p} = p)$, and the average confidence is a setting of p. ECE computes a weighted average of this error across bins:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \left|\mathrm{acc}(b) - \mathrm{conf}(b)\right|,$$

where n_b is the number of predictions in bin b, N is the total number of data points, and $\mathrm{acc}(b)$ and $\mathrm{conf}(b)$ are the accuracy and confidence of bin b, respectively. ECE as framed in (9) leaves ambiguity in both its binning implementation and how to compute calibration for multiple classes. In (5), they bin the probability interval $[0, 1]$ into equally spaced subintervals, and they take the maximum probability output for each datapoint (i.e., the predicted class's probability). We use this for our ECE implementation.

3. Issues With Calibration Metrics

3.1. Not Computing Calibration Across All Predictions

Expected Calibration Error was crafted to mirror reliability diagrams, which are structured around binary classification such as rain vs not rain (3). A consequence is that the error metric is reductive in a multi-class setting. In particular, ECE is computed using only the predicted class's probability, which implies the metric does not assess how accurate a model is with respect to the $K - 1$ other class probabilities. We examine increasing the prevalence of those predictions via label noise, and find that ECE becomes a worse approximation of the calibration error (Figure 1).

3.2. Fixed Calibration Ranges

One major weakness of evenly spaced binning metrics is caused by the dispersion of data across ranges. In computing ECE, there is often a large leftward skew in the output probabilities, with the left end of the region being sparsely populated and the rightward end being densely populated. (That is, network predictions are typically very confident.) This causes only a few bins to contribute the most to ECE—typically one or two as bin sizes are 10-20 in practice (5).

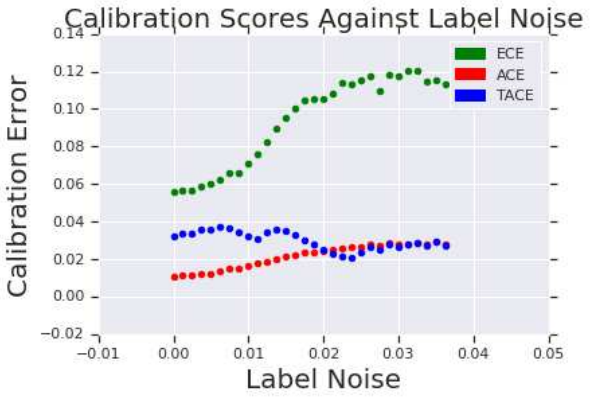

Figure 1. As label noise increases, ECE is outperformed by ACE and TACE. This shows that ECE becomes a worse approximation of true calibration error as class predictions beyond the predicted one matter more. (The $x = 0$ extreme has a true data distribution with deterministic y; $x \to \infty$ extreme has a true data distribution with uniform y.)

More broadly, sharpness, which is the desire for models to always predict with high confidence, i.e., predicted probabilities concentrate to 0 or 1, is a fundamental property (4). Because of the above behavior, ECE conflates calibration and sharpness when a model is highly accurate.

3.3. Bias-Variance Tradeoff

Selecting the number of bins has a bias-variance tradeoff as it determines how many data points fall into each bin and therefore the quality of the estimate of calibration from that bin's range. In particular, a larger number of bins causes more granular measures of calibration error (low bias) but also a high variance of each bin's measurement as bins become sparsely populated. This tradeoff compounds particularly with the problem of fixed calibration ranges, as certain bins have many more data points than others.

3.4. Pathologies in Static Binning Schemes

Metrics that depend on static binning schemes like ECE suffer from issues where you can get near 0 calibration error due to positive and negative predictions overlapping in the same bin. For example, assuming the dataset is 45% positive we could simply output a prediction in the range of (0.41, 0.43) for the negative examples and (0.47, 0.49) for the positive examples to create a set of predictions that has 1.0 AUC, 0 ECE and yet be uncalibrated.

4. New Calibration Metrics

4.1. Multiclass & Static Calibration Error

We first introduce Static Calibration Error (SCE), which is a simple extension of Expected Calibration Error to every probability in the multiclass setting. SCE bins predictions separately for each class probability, computes the calibration error within the bin, and averages across bins:

$$\text{SCE} = \frac{1}{K}\sum_{k=1}^{K}\sum_{b=1}^{B}\frac{n_{bk}}{N}\left|\text{acc}(b,k) - \text{conf}(b,k)\right|.$$

Here, $\text{acc}(b,k)$ and $\text{conf}(b,k)$ are the accuracy and confidence of bin b for class label k, respectively; n_{bk} is the number of predictions in bin b for class label k; and N is the total number of data points.

4.2. Adaptivity & Adaptive Calibration Error

Adaptive calibration ranges are motivated by the bias-variance tradeoff in the choice of ranges, suggesting that in order to get the best estimate of the overall calibration error the metric should focus on the regions where the predictions are made (and focus less on regions with few predictions). This leads us to introduce Adaptive Calibration Error (ACE). uses an adaptive scheme which spaces the bin intervals so that each contains an equal number of predictions.

In detail, ACE takes as input the predictions P (usually out of a softmax), correct labels, and a number of ranges R.

$$\text{ACE} = \frac{1}{KR}\sum_{k=1}^{K}\sum_{r=1}^{R}\left|\text{acc}(r,k) - \text{conf}(r,k)\right|.$$

Here, $\text{acc}(r,k)$ and $\text{conf}(r,k)$ are the accuracy and confidence of adaptive calibration range r for class label k, respectively; and N is the total number of data points. Calibration range r defined by the $\lfloor N/R \rfloor$th index of the sorted and thresholded predictions.

4.3. Thresholding & Thresholded Adaptive Calibration Error

One initial challenge is that the vast majority of softmax predictions become infinitesimal (Figure 2). These tiny predictions can wash out the calibration score, especially in the case where there are many classes, where a large proportion of them model's predictions correspond to an incorrect class. One response is to only evaluate on values above a threshold ϵ.

These predictions overlap with the predictions evaluated by ECE (all maximum values per datapoint), leading them to have similar reactions to recalibration methods (Table 1 & 2).

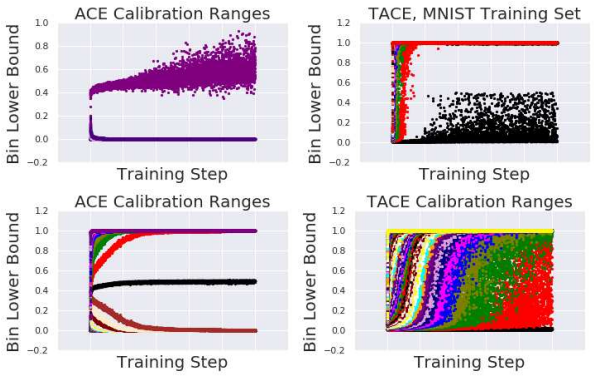

Figure 2. **Top Left:** Lower bounds of calibrations ranges over the course of training for adaptive calibration error on Fashion-MNIST, focusing almost entirely on small ranges and motivating thresholding. **Top Right:** On the MNIST training set with thresholding, so few values are small that the bottom of the lowest range often spikes to .99 and higher due to every datapoint being fit. **Bottom Left:** ACE on Fashion-MNIST validation with 100 calibration ranges. **Bottom Right:** Thresholded adaptive calibration with 50 calibration ranges over the course of training on Fashion-MNIST's validation set.

Method	**ECE**	**TACE**	**SCE**	**ACE**
Uncalibrated	19.64%	10.03%	0.41%	0.13%
Temp. Scaling	**2.16%**	**0.52%**	0.06%	0.06%
Vector Scaling	2.27%	0.61%	**0.04%**	**0.01%**
Matrix Scaling	12.11%	3.98%	0.26%	0.26%
Isotonic Regr.	17.85%	2.83%	0.35%	0.12%

Table 1. ECE, TACE, SCE, and ACE (with 15 bins) on a ResNet-110 applied to CIFAR-100 before calibration, and after the application of post-processing methods. The best recalibration method depends on the metric, which motivates its study.

Method	**ECE**	**TACE**	**SCE**	**ACE**
Uncalibrated	6.63%	2.51%	0.02%	0.015%
Temp. Scaling	5.42%	2.64%	0.01%	0.001%
Vector Scaling	**1.44%**	**1.25%**	**0.002%**	0.004%
Matrix Scaling	5.06%	1.98%	0.01%	0.001%
Isotonic Regr.	3.474%	1.862%	0.01%	**0.000%**

Table 2. ECE, TACE, SCE, and ACE (with 15 bins) on a ResNet-50 applied to ImageNet before calibration, and after the application of various extensions to Platt scaling and Isotonic regression. These percentages across metrics are not directly comparable.

References

[1] Cynthia S Crowson, Elizabeth J Atkinson, and Terry M Therneau. Assessing calibration of prognostic risk scores. *Statistical methods in medical research*, 25(4):1692–1706, 2016. 1

[2] A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. 1

[3] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. 2

[4] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007. 2

[5] 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-Volume 70*, pages 1321–1330. JMLR. org, 2017. 1, 2

[6] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2011. 1

[7] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*, 2018. 1

[8] Allan H Murphy and Edward S Epstein. Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology*, 6(5):748–755, 1967. 1

[9] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 1, 2

[10] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. 2018. 1