Neural Network Calibration for Medical Imaging Classification Using DCA Regularization

Gongbo Liang[1] Yu Zhang[1] Nathan Jacobs[1]

Abstract

Empirically, neural networks are often miscalibrated and dramatically overconfident in their predictions, which could be problematic in any automatic decision-making system. In this work, we focus on the medical field because neural network miscalibration has the potential to lead to significant treatment errors. We propose a novel approach to neural network calibration that maintains the overall classification accuracy while significantly improving model calibration. Our approach can be easily integrated into any classification task as an auxiliary loss term. We show that our approach reduces calibration error significantly across various architectures and datasets.

1. Introduction

Recently, many high-performance deep learning models for various medical imaging analysis tasks have been developed (Esteva et al., 2017; Yang et al., 2018; Mihail et al., 2019). Researchers are actively working on convolutional neural network (CNN) architecture development that is pursuing higher accuracy (Ronneberger et al., 2015; Ribli et al., 2018; Zhang et al., 2019). However, uncertainty quantification is often ignored when evaluating these models. Uncertainty quantification of neural networks is as important as achieving higher accuracy, if not more, especially in automatic decision-making settings in the medical field. An automated method that achieves higher accuracy, but captures uncertainty inaccurately (i.e., providing an inaccurate confidence or probability of a specific prediction) could be dangerous (Jiang et al., 2011).

Unfortunately, modern deep learning neural networks are poorly calibrated (Pereyra et al., 2017; Widmann et al., 2019), which tend to be overconfident in their predictions (Guo et al., 2017; Kumar et al., 2018). Though the real cause of miscalibration is unclear, one reasonable explanation is overfitting the cross-entropy loss for classification models. With larger capacities, classification models can overfit the cross-entropy loss easily without overfitting the $0/1$ loss (e.g., accuracy) (Zhang et al., 2016).

Temperature scaling (Hinton et al., 2015; Guo et al., 2017) is a widely-used, state-of-the-art approach for deep learning calibration. It fixes the miscalibration issue with dividing the predicted probability by a single parameter T ($T > 0$). The method is easy to use and performs well, in general. However, it treats model calibration as a post-processing task and does not affect the model's learning ability.

We propose to add the difference between the predicted confidence and accuracy (DCA) as an auxiliary loss term to cross-entropy loss for classification model calibration. Unlike temperature scaling, the DCA regularization term integrates model calibration into the training stage. As an auxiliary loss, the DCA term may help a deep learning model to learn a better feature representation, which improves the ability of a model to recover the true probability better (Figure 1).

We evaluate the proposed method across four large, publicly available, medical datasets and four widely used CNN architectures. The results show that our approach reduces calibration error significantly by an average of 65.72% compared to uncalibrated methods (from 0.1006 ECE to 0.0345 ECE), while maintaining the overall accuracy across all the experiments—83.08% and 83.58% for the uncalibrated method and our method, respectively. The proposed method is also approximately 20% better on calibration than temperature scaling, the state-of-the-art calibration approach, on average.

2. Background

The problem we address in this paper is the miscalibration issue of supervised classification tasks using modern deep learning networks.

[1] Department of Computer Science, University of Kentucky, Kentucky, USA. Correspondence to: Gongbo Liang <gb.liang@uky.edu>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. Copyright 2020 by the author(s).

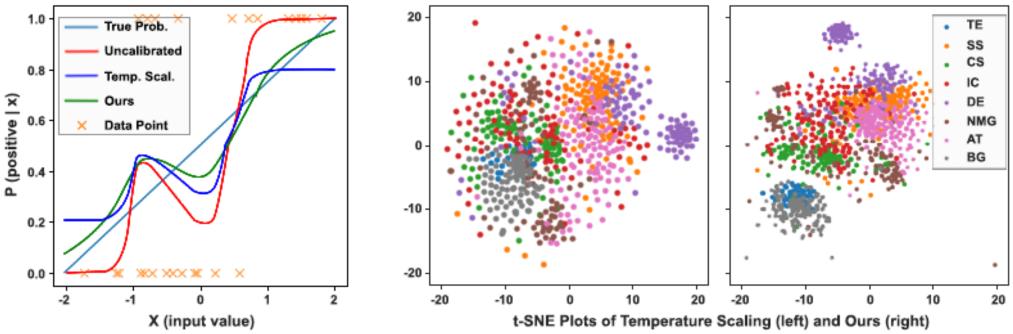

Figure 1. **Left**: The figure shows the ability to recover the true probability distribution of a random dataset. The diagonal line is the true distribution. The red line is the prediction of the uncalibrated model. The predicted probability distribution is far away from the ground truth with many overconfident predictions. Temperature scaling (blue) reduces the prediction confidence of the uncalibrated model, but the calibrated result is still quite far to the ground truth. Our method (green) can better recover the true probability. **Right**: The t-SNE plots of the representations learned by the uncalibrated model and temperature scaling (left) and the proposed method (right). The samples of many classes in the plot of the uncalibrated model and temperature scaling are spreading evenly over the feature space (left). However, the same class samples are densely close in the plot of ours; the TE and BG classes are nicely separated from the rest of the classes (right).

2.1. Problem Definition

Mathematically, the problem can be defined in the following way. The input $X \in x$ and label $Y \in y = \{1, ..., k\}$ are random variables that follow a joint distribution $\pi(X, Y) = \pi(Y|X)\pi(X)$. Let h be a modern neural network with $h(X) = (\hat{Y}, \hat{P})$, where $\hat{Y}$ is the predicted class label and $\hat{P}$ is the associated confidence. We would like the confidence estimate $\hat{P}$ to be calibrated, which intuitively means that $\hat{P}$ represents a true probability. For instance, given 100 predictions with the average confidence of 0.95, we expect that 95 predictions should be correct. In reality, the average confidence of a modern neural network is often higher than its accuracy (Guo et al., 2017; Pereyra et al., 2017; Kumar et al., 2018). The perfect calibration can be defined as:

$$\mathbb{P}\left(\hat{Y} = Y|\hat{P} = p\right) = p, \forall p \in [0, 1]. \tag{1}$$

Difference in expectation between confidence and accuracy (i.e., the calibration error) can be defined as:

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

We want to reduce the calibration error as much as possible.

2.2. Expected Calibration Error

Expected Calibration Error (ECE) is the main criteria that is used to measure neural network calibration error. ECE (Naeini et al., 2015) approximates Equation (2) by partitioning predictions into M bins and taking a weighted average of the accuracy/confidence difference for each bin. All the samples need to be grouped into M interval bins according to the prediction. Let B_m be the set of indices of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$, $m \in M$. The accuracy of B_m is

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i), \tag{3}$$

where $\hat{y}_i$ and y_i are the predicted and ground truth label for sample i. The average prediction confidence of bin B_m can be defined as

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \tag{4}$$

where $\hat{p}_i$ is the confidence of sample i. ECE can be defined with $\text{acc}(B_m)$ and $\text{conf}(B_m)$

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{5}$$

where n is the number of samples.

2.3. Temperature Scaling

Temperature scaling is a two-phase or post-processing method for neural network calibration. The first step is to train a deep learning model. Once the model is trained, a single temperature parameter, T ($T > 0$), is added to the model. The temperature parameter is trained on the validation set using cross-entropy loss while all the other parameters are frozen (Guo et al., 2017). The temperature parameter will be used for calibration at the testing time. The calibrated confidence, $\hat{q}_i$, using temperature scaling is

$$\hat{q}_i = \max_k \theta_{SM}(\frac{z_i}{T})^{(k)}, \tag{6}$$

where k is the class label ($k = 1, ..., K$), $\theta_{SM}(z_i)$ is the predicted probability. As $T \to \infty$, the probability $\hat{q}_i$ approaches $1/K$, which represents maximum uncertainty. With $T \to 0$, the probability collapses to a point mass.

2.4. MMCE

MMCE is a trainable calibration method that uses kernel embeddings (Kumar et al., 2018). The method proposes an auxiliary loss term (MMCE) to the cross-entropy loss. The MMCE auxiliary loss term is computed in a reproducing kernel Hilbert space (RKHS) (Gretton, 2013). The completely loss function can be written as:

$$\text{Loss} = \text{CrossEntropy} + \lambda(\text{MMCE}_m^2(D))^{\frac{1}{2}}, \tag{7}$$

where D denotes a dataset, MMCE is the auxiliary loss term in RKHS, and λ is the weight of the auxiliary loss term. During the training process, MMCE term needs to be reweighted due to the imbalance prediction (e.g., the number of correct predictions is usually larger than the number of incorrect predictions). The re-weighted MMCE term can be written as:

$$\begin{aligned}\text{MMCE}_w^2 = &\sum_{c_i=c_j=0} \frac{p_i, p_j, k(p_i, p_j)}{(m-n)^2} \\ &+ \sum_{c_i=c_j=1} \frac{(1-p_i)(1-p_j)k(p_i, p_j)}{n^2} \\ &- 2 \sum_{c_i=1, c_j=0} \frac{(1-p_i)p_j k(p_i, p_j)}{(m-n)n},\end{aligned} \tag{8}$$

where c is the predicted label, m is the number of corrected predictions, n is the batch size, and k is a universal kernel.

3. Proposed Method

We propose to add the difference between confidence and accuracy (DCA) as an auxiliary loss term to the cross-entropy loss function for classification tasks. DCA is based on expected calibration error, which is a standard metric for quantifying model calibration error. The proposed DCA term is easy to implement and suitable for any classification tasks.

The DCA term provides a more principled fix to the miscalibration issue by penalizing the overfitting of cross-entropy loss. The DCA term penalizes deep learning models when the cross-entropy loss can be reduced, but the 0/1 loss does not change (i.e., when the model is overfitting the cross-entropy loss). In general, the classification loss function can be written as the following:

$$\text{Loss} = \text{CE} + \beta DCA, \tag{9}$$

where CE indicates cross-entropy, β is a weight scalar. The DCA term can be computed for each mini-batch using the following equation:

$$\text{DCA} = \left| \frac{1}{N} \sum_{i=1}^{N} c_i - \frac{1}{N} \sum_{i=1}^{N} p(\hat{y}_i) \right|, \tag{10}$$

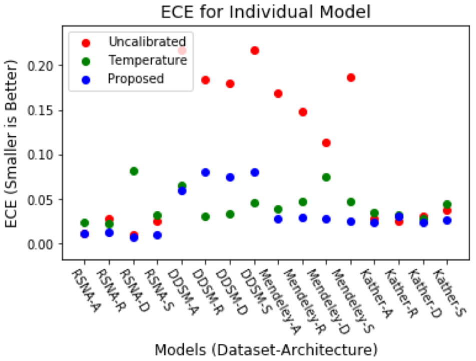

Figure 2. The calibration error (ECE) of each individual model. The model names are formed as the "Dataset-Architecture". (-A: AlexNet, -R: ResNet, -D: DenseNet, -S: SqueezeNet.)

where $\hat{y}_i$ is the predicted label; $p(\hat{y}_i)$ is the predicted probability; $c_i = 1$, if $\hat{y}_i = y_i$, y_i is the true label; otherwise, $c_i = 0$. The final loss function of a classification task can be written as:

$$\text{Loss} = \text{CE} + \beta \left| \frac{1}{N} \sum_{i=1}^{N} c_i - \frac{1}{N} \sum_{i=1}^{N} p(\hat{y}_i) \right|. \tag{11}$$

DCA is differentiable in the prediction confidence term but not strictly in the prediction accuracy term due to the argmax step for computing the predicted label. During the training phase, backpropagated gradients can be done through the confidence terms but not through the accuracy.

4. Results

We compared the proposed method with temperature scaling and uncalibrated models (trained with cross-entropy loss without the application of any calibration methods) on four medical imaging datasets across four popular CNN networks. The MMCE method is not compared because the pure MMCE method performs worse than temperature scaling (Kumar et al., 2018).

The evaluation results show that the proposed method significantly improves model calibration while maintaining the overall classification accuracy. The proposed model reduces calibration error by an average of 65.72% compared to uncalibrated methods (from 0.1006 ECE to 0.0345 ECE) and performs approximately 20% better than the temperature scaling method on all the tested cases. Figure 2 shows the ECE of each individual model.

4.1. Experiment Setup

4.1.1. Datasets and CNN Models

Four large, publicly available, medical imaging datasets (RSNA (RSNA, 2019), DDSM (Heath et al., 2000), Mende-

Table 1. Datasets used in this study.

Name	Modality	# of Images	# of Classes
RSNA	Head CT	674257	2
DDSM	Mammography	10480	2
Mendeley	Chest X-ray	5856	2
Kather	Histological	5000	8

ley (Kermany & Goldbaum, 2018), and Kather (Kather et al., 2016)) were used in this study for both binary and multi-class classification tasks. See Table 1 for more details.

Four transfer learning CNN models were evaluated. More specifically, AlexNet (Krizhevsky et al., 2012), ResNet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2017), and SqueezeNet 1-1 (Iandola et al., 2016) were used. All the models were pre-trained on ImageNet. The parameters of the convolutional (Conv) layers in the original networks were frozen and used as the feature extractors. A Conv layer with 1×1 kernels were added after each feature extractor. Only the parameters of the new Conv layer and the fully connected layers are optimized during the training.

4.2. Binary Classification Calibration Results

Table 2 shows the expected calibration error (ECE) and the accuracy of the uncalibrated models (Unca.), the temperature scaling (Temp.), and the proposed method (DCA). Each model was trained for at least two times. The values in the tables are the averaged result of all the trials.

The table shows that our method is constantly better than the uncalibrated method on model calibration, which reduces the ECE by 70.05% (from 0.1242 to 0.0372). The temperature scaling method has the second smallest average ECE, 0.0454, which is 22% worse than the proposed method. On average, the uncalibrated method and temperature scaling method have 80.16% accuracy, while the proposed method is 80.74%. The proposed method increases the accuracy of 9 out of 12 tests.

4.3. Multi-Class Classification Calibration Results

Table 3 shows the expected calibration error (ECE) and the accuracy of the uncalibrated models (Unca.), the temperature scaling (Temp.), and the predicted method (DCA) on multi-class classification tasks. On average, the uncalibrated model has 0.03 ECE. The proposed method reduces the number by 12.33% to 0.0263 ECE. However, temperature scaling increases the error by over 15% to 0.0347 ECE.

The Kather dataset is a relatively simple and large dataset for its task. The dataset is considered as the MNIST of histology images. It is speculated have a sufficient amount of training data to train a model end-to-end, with a smaller overfitting effect (i.e., miscalibration). In such a case, since the temperature parameter (T) of temperature scaling is learned on only the validation set, it may actually hurt the calibration. However, the proposed method jointly optimizes the accuracy and modal calibration simultaneously, it can still reduce the calibration error.

Table 2. Expected Calibration Error and Accuracy for Binary Classification Tasks

Dataset	**Model**	**ECE** (smaller is better)			**Accuracy**[1] (larger is better)	
		Unca.	**Temp.**	**DCA**	**Unca.**	**DCA**
RSNA	AlexNet	**0.0113**	0.0239	0.0120	0.8376	**0.8488**
	ResNet	0.0276	0.0231	**0.0122**	0.8569	**0.8762**
	DenseNet	0.0102	0.0814	**0.0077**	0.8502	**0.8543**
	SqueezeNet	0.0253	0.0317	**0.0097**	0.8671	**0.8841**
DDSM	AlexNet	0.2164	0.0658	**0.0591**	**0.6766**	0.6291
	ResNet	0.1844	**0.0307**	0.0798	**0.7195**	0.6987
	DenseNet	0.1798	**0.0337**	0.0754	0.7076	**0.7106**
	SqueezeNet	0.2173	**0.0458**	0.0805	**0.6853**	0.6771
Mendeley	AlexNet	0.1693	0.0396	**0.0273**	0.8585	**0.8785**
	ResNet	0.1475	0.0475	**0.0291**	0.8520	**0.8767**
	DenseNet	0.1136	0.0746	**0.0285**	0.8331	**0.8796**
	SqueezeNet	0.1871	0.0468	**0.0252**	0.8742	**0.8750**
Average		0.1242	0.0454	**0.0372**	0.8016	**0.8074**

[1] The temperature scaling method has the same accuracy as the uncalibrated models.

Table 3. Expected Calibration Error and Accuracy for Multi-Class Classification

Dataset	**Model**	**ECE** (smaller is better)			**Accuracy**[1] (larger is better)	
		Unca.	**Temp.**	**DCA**	**Unca.**	**DCA**
Kather	AlexNet	0.0279	0.0344	**0.0243**	**0.9062**	0.9052
	ResNet	**0.0248**	0.0318	0.0304	**0.9355**	0.9229
	DenseNet	0.0302	0.0286	**0.0237**	0.9385	**0.9410**
	SqueezeNet	0.0372	0.0439	**0.0269**	0.8932	**0.9038**
Average		0.0300	0.0347	**0.0263**	**0.9184**	0.9182

[1] The temperature scaling method has the same accuracy as the uncalibrated models.

5. Conclusion

We proposed a novel approach to neural network calibration that maintains the overall classification accuracy while significantly reducing model calibration error. We evaluated our approach across various architectures and datasets. The results show that our approach reduces calibration error significantly and comes closer to recovering the true probability than other approaches. The proposed method can be easily integrated into any classification tasks as an auxiliary loss term, thus not requiring an explicit training round for calibration. We believe this simple, fast, and straightforward method can serve as a strong baseline for future researchers.

References

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Gretton, A. Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, 2013.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heath, M., Bowyer, K., Kopans, D., Moore, R., and Kegelmeyer, W. P. The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, pp. 212–218. Medical Physics Publishing, 2000.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2011.

Kather, J. N., Weis, C.-A., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., Marx, A., and Zöllner, F. G. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:27988, 2016.

Kermany, D. and Goldbaum, M. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley Data*, 2, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2810–2819, 2018.

Mihail, R. P., Liang, G., and Jacobs, N. Automatic hand skeletal shape estimation from radiographs. *IEEE Transactions on NanoBioscience*, 18(3):296–305, July 2019. ISSN 1536-1241. doi: 10.1109/TNB.2019.2911026.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., and Csabai, I. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):4165, 2018.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241. Springer, 2015.

RSNA. Rsna intracranial hemorrhage detection, 2019. URL https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/overview.

Widmann, D., Lindsten, F., and Zachariah, D. Calibration tests in multi-class classification: A unifying framework. In *Advances in Neural Information Processing Systems*, pp. 12236–12246, 2019.

Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M. K., Zhang, Y., Sun, L., and Wang, G. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, Y., Wang, X., Blanton, H., Liang, G., Xing, X., and Jacobs, N. 2d convolutional neural networks for 3d digital breast tomosynthesis classification. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1013–1017. IEEE, 2019.