Improving Accuracy and Calibration via Differentiated Deep Mutual Learning

Supplementary Material

A. Definition of ECE and classwise-ECE

In this section, we adopt the notation from Section 3.

ECE is defined as the expected absolute difference between the model's confidence and its accuracy conditioned on confidence:

$$\mathbb{E}_{x\sim P_{data}}(|\mathbb{E}_{x'\sim P_{data}}(\mathbb{1}(\hat{y}_{x'}=y_{x'})|\hat{p}_{x'}=\hat{p}_x)-\hat{p}_x|), \tag{12}$$

where $\mathbb{E}$ stands for expectation and $\mathbb{1}$ is the indicator function. Since we only have finite samples, ECE cannot be directly calculated using the definition provided above. Therefore, in practical calculations, we replace the above definition with a discretized version of ECE in which the interval $[0,1]$ is divided into M equispaced bins. Let B_i denote the samples with confidences belonging to the i-th bin (i.e. $(\frac{i-1}{M},\frac{i}{M}]$). The accuracy of this bin is $A_i=\frac{1}{|B_i|}\sum_{x\in B_i}\mathbb{1}(\hat{y}_x=y_x)$. The average confidence of this bin is $C_i=\frac{1}{|B_i|}\sum_{x\in B_i}\hat{p}_x$. The discretized version of ECE is defined as

$$\text{ECE}=\sum_{i=1}^{M}\frac{|B_i|}{N}|A_i-C_i|, \tag{13}$$

where N is the number of samples in the dataset.

Similar to ECE for confidence calibration, the classwise-ECE [15] for classwise calibration is defined as

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{x\sim P_{data}}(|\mathbb{E}_{x'\sim P_{data}}(\mathbb{1}(y_{x'}=k)|p_{x'}^{(k)}=p_x^{(k)})-p_x^{(k)}|). \tag{14}$$

The discretized version of classwise-ECE is defined as

$$\text{classwise-ECE}=\sum_{k=1}^{K}\sum_{i=1}^{M}\frac{|B_{i,k}|}{NK}|A_{i,k}-C_{i,k}|, \tag{15}$$

where $B_{i,k}$ denotes the set of samples whose predicted probabilities of the k-th class lie in the i-th bin, $A_{i,k}=\frac{1}{|B_{i,k}|}\sum_{x\in B_{i,k}}\mathbb{1}(y_x=k)$ and $C_{i,k}=\frac{1}{|B_{i,k}|}\sum_{x\in B_{i,k}}p_x^{(k)}$.

B. Related work

B.1. Single-model calibration methods

The current single-model calibration methods can be roughly divided into train-time calibration methods and post-hoc calibration methods.

Train-time calibration: Train-time calibration methods focus on enhancing calibration during the training phase. These methods typically involve designing loss functions or incorporating additional regularization techniques. Entropy Regularization (ER) [26] is an approach that maximizes the entropy of the predicted probabilities while optimizing the cross-entropy loss. By encouraging the model to have more diverse and well-distributed predictions, it helps reduce overconfidence. Label Smoothing (LS) [22, 28] optimizes the cross-entropy loss between the predicted probabilities and smoothed labels, instead of using hard labels. This encourages the model to learn a more nuanced understanding of the data distribution and improves calibration. Focal Loss (FL) [21] addresses the issue of class imbalance by assigning more importance to minority class examples. By doing so, it improves calibration for both majority and minority classes. Maximum Mean Calibration Error (MMCE) [16] designs an auxiliary loss depended on the power of RKHS [7] functions induced by a universal kernel. Difference between Confidence and Accuracy (DCA) [19] proposes an auxiliary loss that encourages the model to minimize the discrepancy between predicted confidence and accuracy. Multi-class Difference in Confidence and Accuracy (MDCA) [10] extends the auxiliary loss introduced by DCA to calibrate the whole predicted probability distribution. This extension enhances the calibration performance of neural networks by ensuring accurate and reliable predictions across all confidence levels. Dynamic Train-time Data Pruning (DTDP) [25] achieves calibration by pruning low-confidence samples every few epochs.

Post-hoc calibration: Post-hoc calibration methods are applied after the model has been trained. Temperature Scaling (TS) [27] is a commonly used method that smooths the logits of a deep neural network to achieve calibration. By adjusting the temperature parameter, TS aligns the predicted probabilities with the expected confidence of the model. Dirichlet Scaling (DS) [15] extends the Beta-calibration [14] method from binary to multi-class classification. It models the predicted probabilities using a Dirichlet distribution and learns the parameters to calibrate the model.

B.2. Deep Ensembles

Deep Ensembles (DE) [5, 17] is a widely used model ensemble technique in deep learning. DE creates an ensemble by training multiple models independently with different initializations or training data subsets. During inference, the predictions from all models are combined, typically by averaging their predicted probabilities.

DE has been shown to improve model performance and calibration by leveraging the diversity among independently trained models. Each model captures different aspects of the data and makes slightly different predictions, leading to more robust and well-calibrated ensemble predictions. While

DE has demonstrated effectiveness in improving model performance, it comes with additional computational overhead during inference.

C. Proof

Here we provide the proof of the proposition 1 presented in the main text (5).

Proposition 1. *Let $h_x(\underline{\theta}, \phi) = (1-\beta) f_x(\underline{\theta}) + \beta g_x(\phi)$ denote the ensemble predictive distribution of $f_x(\underline{\theta})$ and $g_x(\phi)$, $f_x(\underline{\theta})$ and $g_x(\phi)$. We have*

$$\frac{\partial}{\partial \theta} D_{KL}(h_x(\underline{\theta}, \phi) || f_x(\theta)) = \beta \frac{\partial}{\partial \theta} D_{KL}(g_x(\phi) || f_x(\theta)), \tag{16}$$

where $\beta \in [0, 1]$ is the weight of the distributions. Note that variables with underscores do not participate in gradient propagation.

Proof. Given a vector v, let $v^{(k)}$ denote the k-th dimension of v.

$$\begin{aligned}\frac{\partial}{\partial \theta} D_{KL}(h_x(\underline{\theta}, \phi) || f_x(\theta)) &= \frac{\partial}{\partial \theta} \sum_{k=1}^{K} (h_x^{(k)}(\underline{\theta}, \phi) \log \frac{h_x^{(k)}(\underline{\theta}, \phi)}{f_x^{(k)}(\theta)}) \\ &= \sum_{k=1}^{K} \frac{\partial}{\partial \theta} [h_x^{(k)}(\underline{\theta}, \phi)(\log h_x^{(k)}(\underline{\theta}, \phi) - \log f_x^{(k)}(\theta))].\end{aligned} \tag{17}$$

Since $\underline{\theta}$ does not participate in gradient backpropagation,

$$\begin{aligned}\frac{\partial}{\partial \theta} D_{KL}(h_x(\underline{\theta}, \phi) || f_x(\theta)) &= -\sum_{k=1}^{K} h_x^{(k)}(\underline{\theta}, \phi) \frac{\partial}{\partial \theta} \log f_x^{(k)}(\theta) \\ &= -(1-\beta) \sum_{k=1}^{K} f_x^{(k)}(\underline{\theta}) \frac{\partial}{\partial \theta} \log f_x^{(k)}(\theta) \\ &\quad -\beta \sum_{k=1}^{K} g_x^{(k)}(\phi) \frac{\partial}{\partial \theta} \log f_x^{(k)}(\theta).\end{aligned} \tag{18}$$

Let $z_x^{(i)}(\theta)$ denote the i-th dimension of the output logits of the primary model with input x. According to the chain rule of differentiation,

$$\begin{aligned}\sum_{k=1}^{K} f_x^{(k)}(\underline{\theta}) \frac{\partial}{\partial \theta} \log f_x^{(k)}(\theta) &= \sum_{k=1}^{K} \frac{\partial}{\partial \theta} f_x^{(k)}(\theta) \\ &= \sum_{i=1}^{K} \sum_{k=1}^{K} \frac{\partial f_x^{(k)}(\theta)}{\partial z_x^{(i)}(\theta)} \frac{\partial z_x^{(i)}(\theta)}{\partial \theta}.\end{aligned} \tag{19}$$

Since $f_x^{(k)}(\theta) = \exp(z_x^{(k)}(\theta)) / \sum_{j=1}^{K} \exp(z_x^{(j)}(\theta))$,

$$\frac{\partial f_x^{(k)}(\theta)}{\partial z_x^{(i)}(\theta)} = \begin{cases} f_x^{(i)}(\theta)(1 - f_x^{(i)}(\theta)), & \text{where } k = i \\ -f_x^{(i)}(\theta) f_x^{(k)}(\theta), & \text{where } k \neq i \end{cases} \tag{20}$$

So,

$$\begin{aligned}\sum_{k=1}^{K} \frac{\partial f_x^{(k)}(\theta)}{\partial z_x^{(i)}(\theta)} &= f_x^{(i)}(\theta)(1 - f_x^{(i)}(\theta)) - \sum_{k \neq i} f_x^{(i)}(\theta) f_x^{(k)}(\theta) \\ &= 0.\end{aligned} \tag{21}$$

Such that

$$\begin{aligned}\frac{\partial}{\partial \theta} D_{KL}(h_x(\underline{\theta}, \phi) || f_x(\theta)) &= -\beta \sum_{k=1}^{K} g_x^{(k)}(\phi) \frac{\partial}{\partial \theta} \log f_x^{(k)}(\theta) \\ &= \beta \frac{\partial}{\partial \theta} \sum_{k=1}^{K} g_x^{(k)}(\phi)(\log g_x^{(k)}(\phi) - \log f_x^{(k)}(\theta)) \\ &= \beta \frac{\partial}{\partial \theta} D_{KL}(g_x(\phi) || f_x(\theta)).\end{aligned} \tag{22}$$

Q.E.D.

D. Tables of comparison results on CIFAR-10 and Tiny-ImageNet

See Tables S1 and S2.

E. Reliability diagrams and confidence histograms of different calibration methods

Figure S1 displays the reliability diagrams and confidence histograms of CE, FL+MDCA, and Diff-DML on the CIFAR-100 test set. In the reliability diagrams, the red bars represent the differences between confidence and accuracy within the current probability interval. It can be observed from the reliability diagrams that, compared to CE and FL+MDCA, the length of most of the red bars in Diff-DML is shorter, indicating that the confidence predictions output by Diff-DML better aligned with its accuracy. From the confidence histograms, it is evident that, compared to CE, the number of samples with confidence higher than 90% in the predicted distribution decreased noticeably with Diff-DML training. Additionally, we present the average confidence and model accuracy for all samples in the confidence histograms, indicated respectively with the green dashed line and the red dashed line. It can be observed that the red and green dashed lines with Diff-DML training are closer, indicating better calibration performance of Diff-DML.

F. Table for selection of auxiliary model

See Table S4.

Method	Acc(%)↑		ECE(%)↓		cw-ECE(10^{-3})↓	
	ResNet34	ResNet50	ResNet34	ResNet50	ResNet34	ResNet50
Baseline	95.03	94.91	3.23	2.95	6.66	6.42
CE	94.53	94.70	2.26	2.68	5.35	6.15
Post-hoc calibration methods						
CE+TS	94.53	94.70	**0.51**	**0.34**	2.99	3.19
CE+VS	94.63	94.73	0.55	0.58	3.03	2.96
CE+DS	94.57	94.70	0.61	0.52	3.14	3.17
Regularization based calibration methods						
DCA	95.22	94.91	3.28	3.22	7.01	6.68
MMCE	94.68	94.34	3.22	3.29	6.54	6.73
FL	94.88	94.31	1.21	1.07	3.36	4.06
FL+DTDP	93.59	93.88	2.46	2.35	5.72	5.69
FL+MDCA	94.75	94.16	0.87	0.68	3.04	3.72
DML-based calibration methods						
DML	94.84	94.10	1.26	1.06	3.32	3.06
Diff-DML(ours)	**95.39**	**95.20**	0.68	0.90	**2.43**	**2.60**

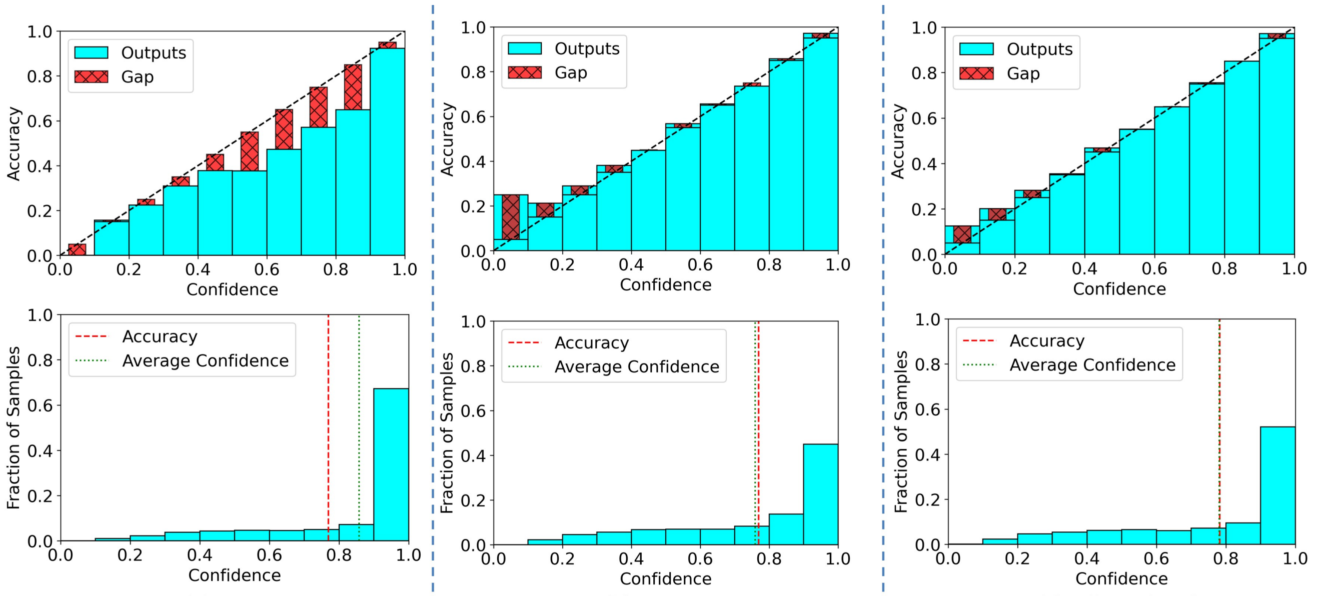
Table S1. Different calibration methods' accuracy, ECE, and cw-ECE on the CIFAR-10 dataset.

Method	Acc(%)↑		ECE(%)↓		cw-ECE(10^{-3})↓	
	ResNet34	ResNet50	ResNet34	ResNet50	ResNet34	ResNet50
Baseline	54.88	54.84	9.03	7.53	1.75	1.74
CE	54.73	54.38	7.63	6.14	1.71	1.64
Post-hoc calibration methods						
CE+TS	54.73	54.38	1.92	1.62	1.53	1.53
CE+VS	55.34	54.02	1.46	1.52	1.52	1.43
CE+DS	52.54	49.74	7.19	10.08	2.13	2.44
Regularization based calibration methods						
DCA	54.36	54.74	8.62	7.92	1.84	1.81
MMCE	54.50	54.12	6.55	4.37	1.67	1.56
FL	54.16	53.38	2.72	2.13	1.61	1.63
FL+DTDP	55.70	57.92	**1.31**	1.76	1.59	1.57
FL+MDCA	53.92	52.98	3.00	3.14	1.62	1.66
DML-based calibration methods						
DML	56.44	56.50	3.78	2.52	1.61	1.54
Diff-DML(ours)	**57.50**	**58.02**	1.33	**1.20**	**1.52**	**1.48**

Table S2. Different calibration methods' accuracy, ECE, and cw-ECE on the Tiny-ImageNet dataset.

Method	Acc(%)↑		ECE(%)↓		cw-ECE(10^{-3})↓	
	ResNet34	ResNet50	ResNet34	ResNet50	ResNet34	ResNet50
CIFAR-10						
DE-3	**95.79**	**95.63**	1.19	0.77	3.64	3.01
DML	94.84	94.10	1.26	1.06	3.32	3.06
DML-2	94.52	94.66	1.17	1.06	3.31	3.27
Diff-DML(ours)	95.39	95.20	0.68	0.90	**2.43**	2.60
Diff-DML-2(ours)	95.29	95.23	**0.51**	**0.63**	2.48	**2.48**
Tiny-ImageNet						
DE-3	**58.78**	57.60	1.95	3.32	1.58	1.59
DML	56.44	56.50	3.78	2.52	1.61	1.54
DML-2	57.46	55.72	3.24	2.33	1.58	1.56
Diff-DML(ours)	57.50	58.02	1.33	1.20	1.52	**1.48**
Diff-DML-2(ours)	58.60	**58.08**	**1.08**	**1.10**	**1.49**	1.53

Table S3. Comparison of DML and Diff-DML trained with more auxiliary models against DE on CIFAR-10 and Tiny-ImageNet dataset.



(a) CE

(b) FL+MDCA

(c) Diff-DML(ours)

Figure S1. Reliability diagrams (top) and confidence histograms (bottom) of (a) CE, (b) FL+MDCA and (c) Diff-DML on CIFAR-100. In the reliability diagrams, blue bars depict the accuracy of model-predicted samples within various confidence intervals, while red bars signify the disparities between confidence and accuracy within the current probability interval. Ideally, a perfectly calibrated model would exhibit all blue bars aligned on the diagonal, implying the absence of red bars. Confidence histograms illustrate confidence distribution, with the green dashed line indicating average confidence and the red dashed line representing prediction accuracy.

f	ResNet34			ResNet50		
g	Acc(%)↑	ECE(%)↓	cw-ECE(10^{-3})↓	Acc(%)↑	ECE(%)↓	cw-ECE(10^{-3})↓
CIFAR-10						
CNN	94.35	8.00	16.8	94.10	7.65	16.14
ResNet18	95.39	0.68	**2.34**	95.20	0.90	2.60
ResNet34	**95.48**	0.65	2.35	95.16	**0.59**	2.59
ResNet50	**95.48**	**0.57**	2.56	**95.25**	0.64	**2.36**
CIFAR-100						
CNN	76.09	5.31	2.03	75.36	7.67	2.42
ResNet18	78.20	**0.65**	1.43	**79.05**	0.86	1.40
ResNet34	**79.46**	1.17	**1.37**	78.74	0.91	**1.36**
ResNet50	79.02	0.78	1.38	79.01	**0.85**	1.37

Table S4. Performance of Diff-DML using different backbones as auxiliary models