Supplementary Materials for: On Calibration of Modern Neural Networks

S1. Further Information on Calibration Metrics

We can connect the ECE metric with our exact miscalibration definition, which is restated here:

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

Let $F_{\hat{P}}(p)$ be the cumulative distribution function of $\hat{P}$ so that $F_{\hat{P}}(b)-F_{\hat{P}}(a)=\mathbb{P}(\hat{P}\in[a,b])$. Using the Riemann-Stieltjes integral we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\hat{P}}\left[\left|\mathbb{P}\left(\hat{Y}=Y \mid \hat{P}=p\right)-p\right|\right] \\
&=\int_0^1\left|\mathbb{P}\left(\hat{Y}=Y \mid \hat{P}=p\right)-p\right| dF_{\hat{P}}(p) \\
&\approx \sum_{m=1}^{M}\left|\mathbb{P}(\hat{Y}=Y|\hat{P}=p_m)-p_m\right| \mathbb{P}(\hat{P}\in I_m)
\end{aligned}$$

where I_m represents the interval of bin B_m. $\left|\mathbb{P}(\hat{Y}=Y|\hat{P}=p_m)-p_m\right|$ is closely approximated by $|\text{acc}(B_m)-\hat{p}(B_m)|$ for n large. Hence ECE using M bins converges to the M-term Riemann-Stieltjes sum of $\mathbb{E}_{\hat{P}}\left[\left|\mathbb{P}\left(\hat{Y}=Y \mid \hat{P}=p\right)-p\right|\right]$.

S2. Further Information on Temperature Scaling

Here we derive the temperature scaling model using the entropy maximization principle with an appropriate balanced equation.

Claim 1. *Given n samples' logit vectors $\mathbf{z}_1,\ldots,\mathbf{z}_n$ and class labels $y_1,\ldots,y_n$, temperature scaling is the unique solution q to the following entropy maximization problem:*

$$\begin{aligned}
\max_{q} \quad & -\sum_{i=1}^{n}\sum_{k=1}^{K} q(\mathbf{z}_i)^{(k)} \log q(\mathbf{z}_i)^{(k)} \\
\textit{subject to} \quad & q(\mathbf{z}_i)^{(k)} \geq 0 \qquad \forall i,k \\
& \sum_{k=1}^{K} q(\mathbf{z}_i)^{(k)} = 1 \quad \forall i \\
& \sum_{i=1}^{n} z_i^{(y_i)} = \sum_{i=1}^{n}\sum_{k=1}^{K} z_i^{(k)} q(\mathbf{z}_i)^{(k)}.
\end{aligned}$$

The first two constraint ensure that q is a probability distribution, while the last constraint limits the scope of distributions. Intuitively, the constraint specifies that the average true class logit is equal to the average weighted logit.

Proof. We solve this constrained optimization problem using the Lagrangian. We first ignore the constraint $q(\mathbf{z}_i)^{(k)}$ and later show that the solution satisfies this condition. Let $\lambda,\beta_1,\ldots,\beta_n\in\mathbb{R}$ be the Lagrangian multipliers and define

$$\begin{aligned}
L = & -\sum_{i=1}^{n}\sum_{k=1}^{K} q(\mathbf{z}_i)^{(k)} \log q(\mathbf{z}_i)^{(k)} \\
& + \lambda \sum_{i=1}^{n}\left[\sum_{k=1}^{K} z_i^{(k)} q(\mathbf{z}_i)^{(k)} - z_i^{(y_i)}\right] \\
& + \sum_{i=1}^{n} \beta_i \sum_{k=1}^{K} (q(\mathbf{z}_i)^{(k)} - 1).
\end{aligned}$$

Taking the derivative with respect to $q(\mathbf{z}_i)^{(k)}$ gives

$$\frac{\partial}{\partial q(\mathbf{z}_i)^{(k)}} L = -nK - \log q(\mathbf{z}_i)^{(k)} + \lambda z_i^{(k)} + \beta_i.$$

Setting the gradient of the Lagrangian L to 0 and rearranging gives

$$q(\mathbf{z}_i)^{(k)} = e^{\lambda z_i^{(k)} + \beta_i - nK}.$$

Since $\sum_{k=1}^{K} q(\mathbf{z}_i)^{(k)} = 1$ for all i, we must have

$$q(\mathbf{z}_i)^{(k)} = \frac{e^{\lambda z_i^{(k)}}}{\sum_{j=1}^{K} e^{\lambda z_i^{(j)}}},$$

which recovers the temperature scaling model by setting $T=\frac{1}{\lambda}$. □

Figure S1 visualizes Claim 1. We see that, as training continues, the model begins to overfit with respect to NLL (red line). This results in a low-entropy softmax distribution over classes (blue line), which explains the model's overconfidence. Temperature scaling not only lowers the NLL but also raises the entropy of the distribution (green line).

S3. Additional Tables

Tables S1, S2, and S3 display the MCE, test error, and NLL for all the experimental settings outlined in Section 5.

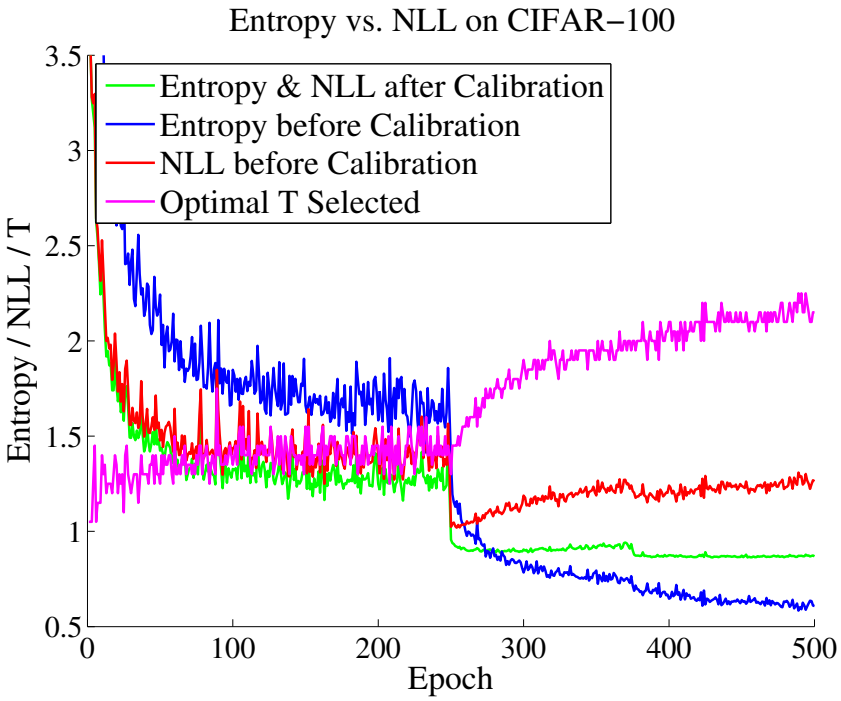

Figure S1. Entropy and NLL for CIFAR-100 before and after calibration. The optimal T selected by temperature scaling rises throughout optimization, as the pre-calibration entropy decreases steadily. The post-calibration entropy and NLL on the validation set coincide (which can be derived from the gradient optimality condition of T).

Dataset	Model	Uncalibrated	Hist. Binning	Isotonic	BBQ	Temp. Scaling	Vector Scaling	Matrix Scaling
Birds	ResNet 50	30.06%	25.35%	16.59%	11.72%	**9.08%**	9.81%	38.67%
Cars	ResNet 50	41.55%	**5.16%**	15.23%	9.31%	20.23%	8.59%	29.65%
CIFAR-10	ResNet 110	33.78%	26.87%	**7.8%**	72.64%	8.56%	27.39%	22.89%
CIFAR-10	ResNet 110 (SD)	34.52%	17.0%	16.45%	19.26%	15.45%	15.55%	**10.74%**
CIFAR-10	Wide ResNet 32	27.97%	12.19%	6.19%	9.22%	9.11%	**4.43%**	9.65%
CIFAR-10	DenseNet 40	22.44%	7.77%	19.54%	14.57%	4.58%	**3.17%**	4.36%
CIFAR-10	LeNet 5	8.02%	16.49%	18.34%	82.35%	**5.14%**	19.39%	16.89%
CIFAR-100	ResNet 110	35.5%	7.03%	10.36%	10.9%	4.74%	**2.5%**	45.62%
CIFAR-100	ResNet 110 (SD)	26.42%	9.12%	10.95%	9.12%	**8.85%**	**8.85%**	35.6%
CIFAR-100	Wide ResNet 32	33.11%	6.22%	14.87%	11.88%	**5.33%**	6.31%	44.73%
CIFAR-100	DenseNet 40	21.52%	9.36%	10.59%	**8.67%**	19.4%	8.82%	38.64%
CIFAR-100	LeNet 5	10.25%	18.61%	**3.64%**	9.96%	5.22%	8.65%	18.77%
ImageNet	DenseNet 161	14.07%	13.14%	11.57%	10.96%	12.29%	**9.61%**	-
ImageNet	ResNet 152	12.2%	14.57%	**8.74%**	8.85%	12.29%	9.61%	-
SVHN	ResNet 152 (SD)	19.36%	11.16%	18.67%	**9.09%**	18.05%	30.78%	18.76%
20 News	DAN 3	17.03%	10.47%	9.13%	**6.28%**	8.21%	8.24%	17.43%
Reuters	DAN 3	**14.01%**	16.78%	44.95%	36.18%	25.46%	18.88%	19.39%
SST Binary	TreeLSTM	21.66%	**3.22%**	13.91%	36.43%	6.03%	6.03%	6.03%
SST Fine Grained	TreeLSTM	27.85%	28.35%	19.0%	**8.67%**	44.75%	11.47%	11.78%

Table S1. MCE (%) (with $M = 15$ bins) on standard vision and NLP datasets before calibration and with various calibration methods. The number following a model's name denotes the network depth. MCE seems very sensitive to the binning scheme and is less suited for small test sets.

S4. Additional Reliability Diagrams

We include reliability diagrams for additional datasets: CIFAR-10 (Figure S2) and SST (Figure S3 and Figure S4). Note that, as mentioned in Section 2, the reliability diagrams do not represent the proportion of predictions that belong to a given bin.

Dataset	Model	Uncalibrated	Hist. Binning	Isotonic	BBQ	Temp. Scaling	Vector Scaling	Matrix Scaling
Birds	ResNet 50	**22.54%**	55.02%	23.37%	37.76%	**22.54%**	22.99%	29.51%
Cars	ResNet 50	14.28%	16.24%	14.9%	19.25%	14.28%	**14.15%**	17.98%
CIFAR-10	ResNet 110	**6.21%**	6.45%	6.36%	6.25%	**6.21%**	6.37%	6.42%
CIFAR-10	ResNet 110 (SD)	5.64%	5.59%	5.62%	**5.55%**	5.64%	5.62%	5.69%
CIFAR-10	Wide ResNet 32	**6.96%**	7.3%	7.01%	7.35%	**6.96%**	7.1%	7.27%
CIFAR-10	DenseNet 40	**5.91%**	6.12%	5.96%	6.0%	**5.91%**	5.96%	6.0%
CIFAR-10	LeNet 5	15.57%	15.63%	15.69%	15.64%	15.57%	**15.53%**	15.81%
CIFAR-100	ResNet 110	27.83%	34.78%	28.41%	28.56%	27.83%	**27.82%**	38.77%
CIFAR-100	ResNet 110 (SD)	**24.91%**	33.78%	25.42%	25.17%	**24.91%**	24.99%	35.09%
CIFAR-100	Wide ResNet 32	**28.0%**	34.29%	28.61%	29.08%	**28.0%**	28.45%	37.4%
CIFAR-100	DenseNet 40	26.45%	34.78%	26.73%	26.4%	26.45%	**26.25%**	36.14%
CIFAR-100	LeNet 5	**44.92%**	54.06%	45.77%	46.82%	**44.92%**	45.53%	52.44%
ImageNet	DenseNet 161	22.57%	48.32%	23.2%	47.58%	22.57%	**22.54%**	-
ImageNet	ResNet 152	**22.31%**	48.1%	22.94%	47.6%	**22.31%**	22.56%	-
SVHN	ResNet 152 (SD)	**1.98%**	2.06%	2.04%	2.04%	**1.98%**	2.0%	2.08%
20 News	DAN 3	20.06%	25.12%	20.29%	20.81%	20.06%	**19.89%**	22.0%
Reuters	DAN 3	2.97%	7.81%	3.52%	3.93%	2.97%	**2.83%**	3.52%
SST Binary	TreeLSTM	11.81%	12.08%	11.75%	**11.26%**	11.81%	11.81%	11.81%
SST Fine Grained	TreeLSTM	49.5%	49.91%	48.55%	49.86%	49.5%	49.77%	**48.51%**

Table S2. Test error (%) on standard vision and NLP datasets before calibration and with various calibration methods. The number following a model's name denotes the network depth. Error with temperature scaling is exactly the same as uncalibrated.

Dataset	Model	Uncalibrated	Hist. Binning	Isotonic	BBQ	Temp. Scaling	Vector Scaling	Matrix Scaling
Birds	ResNet 50	0.9786	1.6226	1.4128	1.2539	**0.8792**	0.9021	2.334
Cars	ResNet 50	0.5488	0.7977	0.8793	0.6986	0.5311	**0.5299**	1.0206
CIFAR-10	ResNet 110	0.3285	0.2532	0.2237	0.263	0.2102	0.2088	**0.2048**
CIFAR-10	ResNet 110 (SD)	0.2959	0.2027	0.1867	0.2159	0.1718	**0.1709**	0.1766
CIFAR-10	Wide ResNet 32	0.3293	0.2778	0.2428	0.2774	0.2283	0.2275	**0.2229**
CIFAR-10	DenseNet 40	0.2228	0.212	0.1969	0.2087	**0.1750**	0.1757	0.176
CIFAR-10	LeNet 5	0.4688	0.529	0.4757	0.4984	0.459	**0.4568**	0.4607
CIFAR-100	ResNet 110	1.4978	1.4379	1.207	1.5466	**1.0442**	1.0485	2.5637
CIFAR-100	ResNet 110 (SD)	1.1157	1.1985	1.0317	1.1982	**0.8613**	0.8655	1.8182
CIFAR-100	Wide ResNet 32	1.3434	1.4499	1.2086	1.459	**1.0565**	1.0648	2.5507
CIFAR-100	DenseNet 40	1.0134	1.2156	1.0615	1.1572	0.9026	**0.9011**	1.9639
CIFAR-100	LeNet 5	1.6639	2.2574	1.8173	1.9893	**1.6560**	1.6648	2.1405
ImageNet	DenseNet 161	0.9338	1.4716	1.1912	1.4272	0.8885	**0.8879**	-
ImageNet	ResNet 152	0.8961	1.4507	1.1859	1.3987	**0.8657**	0.8742	-
SVHN	ResNet 152 (SD)	0.0842	0.1137	0.095	0.1062	**0.0821**	0.0844	0.0924
20 News	DAN 3	0.7949	1.0499	0.8968	0.9519	0.7387	**0.7296**	0.9089
Reuters	DAN 3	0.102	0.2403	0.1475	0.1167	0.0994	**0.0990**	0.1491
SST Binary	TreeLSTM	0.3367	0.2842	0.2908	0.2778	**0.2739**	**0.2739**	**0.2739**
SST Fine Grained	TreeLSTM	1.1475	1.1717	1.1661	1.149	1.1168	**1.1085**	1.1112

Table S3. NLL (%) on standard vision and NLP datasets before calibration and with various calibration methods. The number following a model's name denotes the network depth. To summarize, NLL roughly follows the trends of ECE.

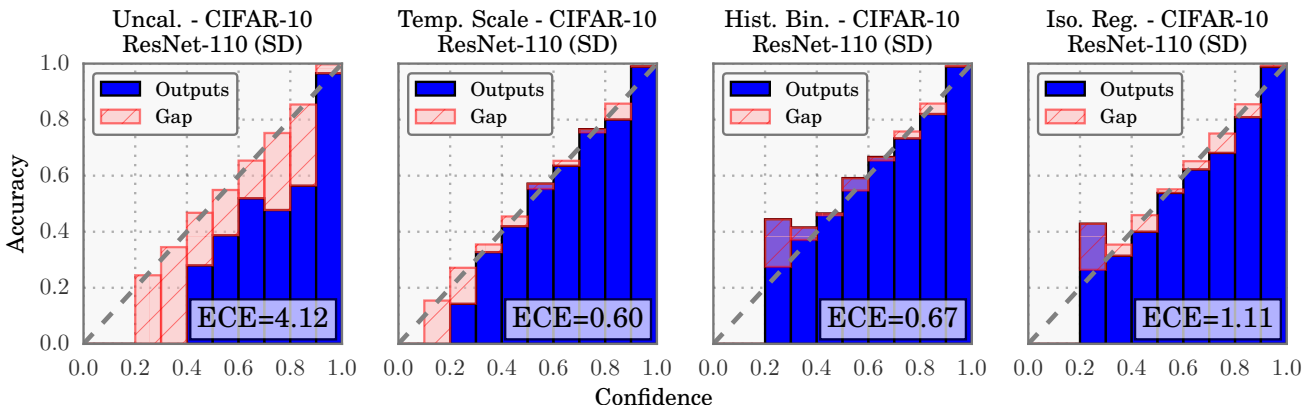

Figure S2. Reliability diagrams for CIFAR-10 before (far left) and after calibration (middle left, middle right, far right).

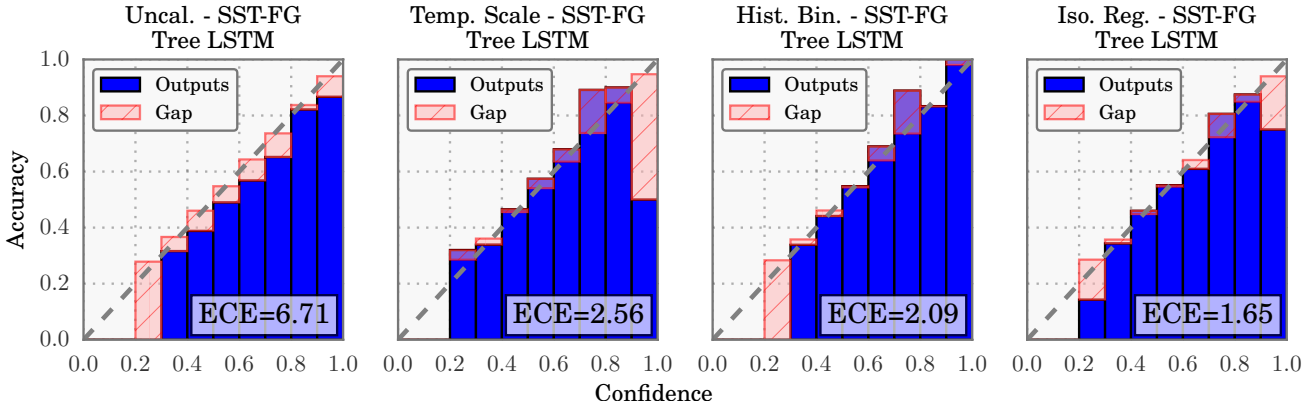

Figure S3. Reliability diagrams for SST Binary and SST Fine Grained before (far left) and after calibration (middle left, middle right, far right).

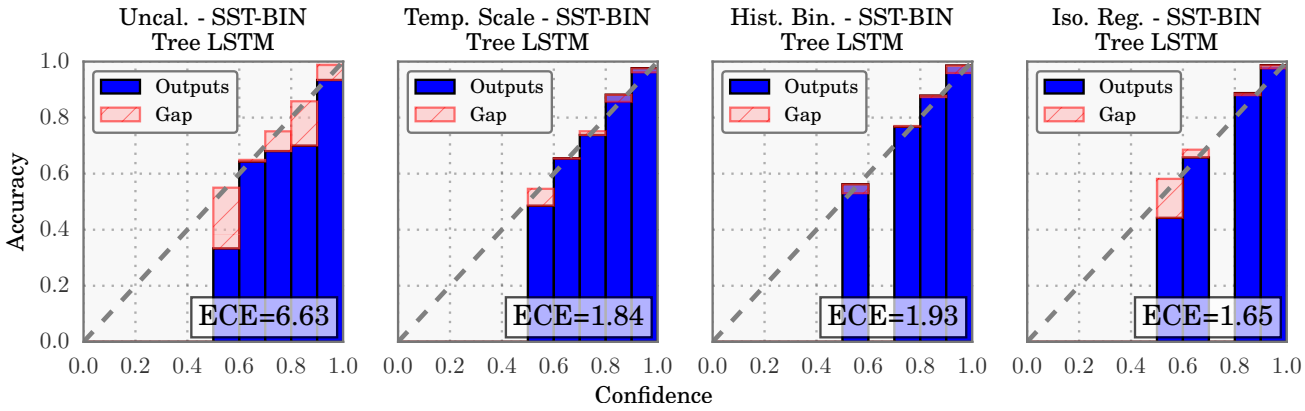

Figure S4. Reliability diagrams for SST Binary and SST Fine Grained before (far left) and after calibration (middle left, middle right, far right).