The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration

Bingyuan Liu[1] Ismail Ben Ayed[1] Adrian Galdran[2] Jose Dolz[1]

[1]ÉTS Montreal, Canada [2]Universitat Pompeu Fabra, Spain

Introduction

Calibrating deep neural networks (DNNs) has been attracting an increased attention recently, which is critical to obtain trustworthy models. To address this issue, our contributions are as follows:

- Introduce a constrained-optimization perspective unifying previous calibration losses.
- Propose a simple and flexible generalization based on inequality constraints, which imposes a controllable margin on logit distances.
- Achieve state-of-the-art calibration performances over a variety of benchmarks, including standard/fine-grained image classification, semantic segmentation and text classification.

Background : calibration

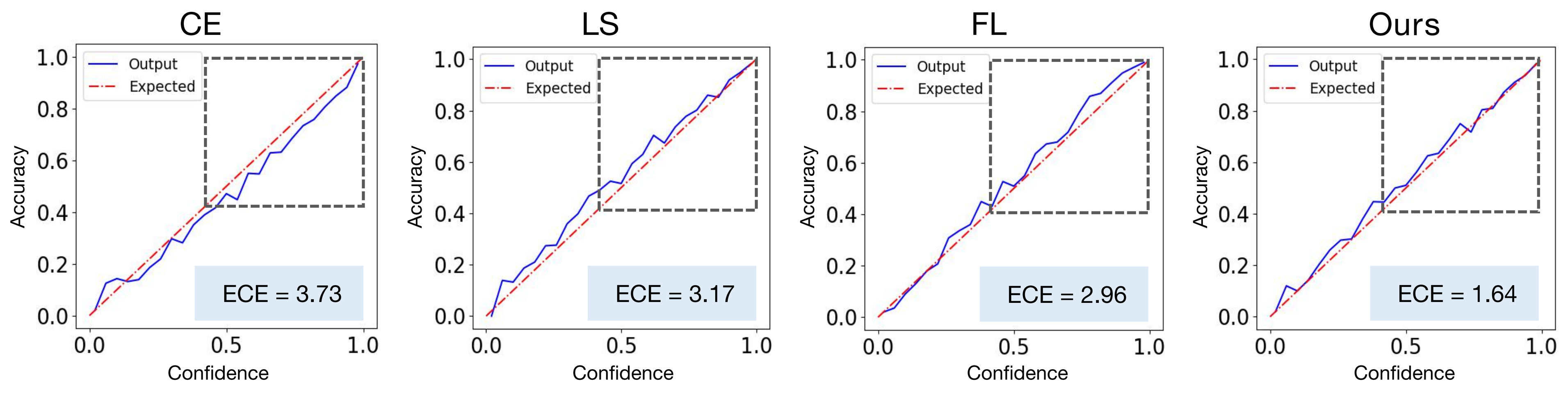

Figure: Calibration visualizations (reliability diagrams) and metrics (ECE) of different methods on Tiny-ImageNet.

Calibrated models. Perfectly calibrated models are those for which the predicted confidence for each sample is equal to the model accuracy : $\hat{p} = \mathbb{P}(\hat{y} = y|\hat{p})$.

Miscalibration of DNNs is mainly caused by overfitting due to the minimization of the cross-entropy (CE) during training, which implicitly pushes softmax vectors $\mathbf{s}$ towards the vertices of the simplex, thereby magnifying the distances between the largest logit $\max_k(l_k)$ and the rest of the logits.

A constrained-optimization perspective of calibration

Let us first define the vector of logit distances between the winner class and the rest as:

$$\mathbf{d}(\mathbf{l}) = (\max_j(l_j) - l_k)_{1 \le k \le K} \in \mathbb{R}^K \quad (1)$$

Previous state-of-the-art calibration losses, i.e., label smoothing (LS), focal loss (FL), and explicit confidence penalty (ECP), could be approximately viewed as **different soft penalty functions** for imposing the same logit-distance equality constraint on CE:

$$\mathbf{d}(\mathbf{l}) = \mathbf{0} \quad (2)$$

Clearly, this constraint is a trivial and non-informative solution.

Margin-based Label Smoothing

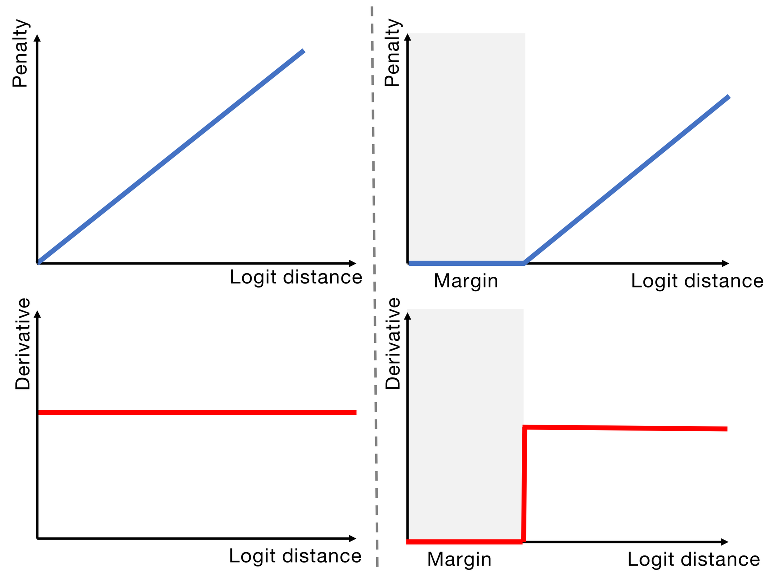

Figure: Illustration of the linear (left) and margin-based (right) penalties for imposing logit-distance constraints, along with the corresponding derivatives.

Though Eq. 2 is not reached in practice with soft penalties jointly with CE, it might prevent from reaching the best compromise between the discriminative performance and calibration.

To address this issue, we propose **a generalized inequality constraint with a positive and controllable margin**:

$$\min \quad \mathcal{L}_{\text{CE}} \quad \text{s.t.} \quad \mathbf{d}(\mathbf{l}) \le \mathbf{m}, \quad \mathbf{m} > 0 \quad (3)$$

The figure in the left illustrates the differences between the linear penalty for equality constraint in Eq. 2 and our margin-based inequality. The gradient of our method is back-propagated only on those logits where the distances are above the margin. In practice, we resort to a simpler unconstrained approximation with ReLU function:

$$\min \quad \mathcal{L}_{\text{CE}} + \lambda \sum_k \max(0, \max_j(l_j) - l_k - m) \quad (4)$$

Results

Datasets. Image classification: CIFAR-10 and Tiny-ImageNet; Fine-grained image classification: CUB-200-2011; Semantic segmentation: PASCAL VOC 2012; Text classification: 20 Newsgroups.

Metrics. Calibration: expected calibration error (ECE) and its variant, Adaptive ECE (AECE); Discrimination: accuracy (Acc) for classification and mean intersection over union (mIoU) for segmentation.

Table: Calibration (top) and classification (bottom) performances on two popular image classification benchmarks.

Dataset	Model	CE		ECP		LS		FL		FLSD		Ours (m=0)		Ours	
		ECE	AECE	ECE	AECE	ECE	AECE	ECE	AECE	ECE	AECE	ECE	AECE	ECE	AECE
Tiny-ImageNet	R-50	3.73	3.69	4.00	3.92	3.17	3.16	2.96	3.12	2.91	2.95	2.50	2.58	**1.64**	**1.73**
	R-101	4.97	4.97	4.68	4.66	2.20	2.21	2.55	2.44	4.91	4.91	1.89	1.95	**1.62**	**1.68**
CIFAR-10	R-50	5.85	5.84	3.01	**2.99**	2.79	3.85	3.90	3.86	3.84	3.60	3.72	4.29	**1.16**	3.18
	R-101	5.74	5.73	5.41	5.40	3.56	4.68	4.60	4.58	4.58	4.57	3.07	3.97	**1.38**	**3.25**

Dataset	Model	CE	ECP	LS	FL	FLSD	Ours (m=0)		Ours	
							Acc	Δ	Acc	Δ
Tiny-ImageNet	R-50	65.02	64.98	**65.78**	63.09	64.09	65.15	-0.63	64.74	-1.04
	R-101	65.62	65.69	**65.87**	62.97	62.96	65.72	-0.15	65.81	-0.06
CIFAR-10	R-50	93.20	94.75	94.87	94.82	94.77	94.76	-0.49	**95.25**	+0.38
	R-101	93.33	93.35	93.23	92.42	92.38	**95.36**	+0.23	95.13	-0.23

Results

Table: CUB-200-2011

Method	Acc	ECE
CE	73.09	6.75
ECP	73.51	5.55
LS	74.51	5.16
FL	72.87	8.41
Ours	74.56	2.78

Table: Pascal VOC 2012

Method	mIoU	ECE
CE	70.92	8.26
ECP	71.16	8.31
LS	71.00	9.35
FL	69.99	11.44
Ours	71.20	7.94

Table: 20 Newsgroups

Method	Acc	ECE
CE	67.01	22.75
ECP	66.48	22.97
LS	67.14	8.07
FL	66.08	10.80
Ours	67.89	5.40

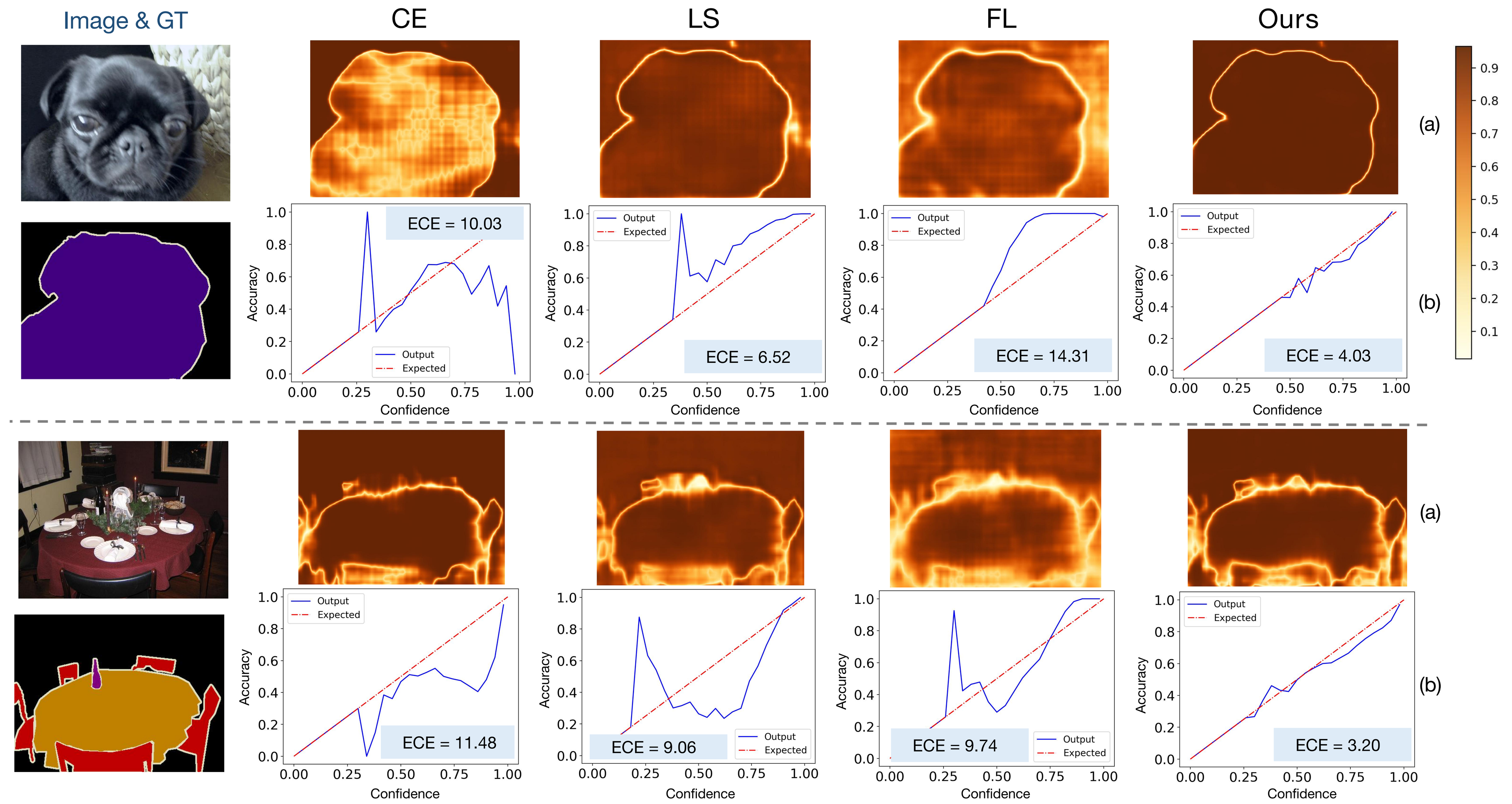

Figure: **Visual results on semantic segmentation.** In the left, we give the original image with ground-truth (GT), then we present the confidence map (a) and the reliability diagram (b) with the ECE (%) score for each method. The value of confidence map represent the predicted confidence, i.e., the element of the soft-max probability for the winner class. It is noted that deeper color denotes higher confidence in the map, as shown in the legend at the upper right corner.

Conclusion

- We introduce a constrained-optimization perspective unifying previous calibration losses and then propose the margin-based label smoothing method.
- Unlike previous losses, our method always push the model to a non-trivial and informative solution, thus achieving better compromise between discriminative performance and calibration.
- Future works include comprehensive studies on data/domain distributional shift, and improving the optimization algorithm.