Can Calibration Improve Sample Prioritization?

Ganesh Tata*[1], Gautham Krishna Gudur*[2], Gopinath Chennupati[3], Mohammad Emtiyaz Khan[4]

University of Alberta[1], Global AI Accelerator, Ericsson[2], Amazon Alexa AI[3], RIKEN Center for Advanced Intelligence Project[4]

*Equal Contribution

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- Calibration can reduce overconfident predictions of deep neural networks, but **can calibration also accelerate training?**
- We show that performing **calibration during training** can –
  - Improve **the quality of subsets** when performing **sample prioritization**
  - Reduce the number of training samples per epoch (**by at least 70%**)
  - **Speed up** the overall training process
- **Calibrated pre-trained 'target' models** coupled with calibration during training can also guide sample prioritization.

Calibration (during training)

A technique that *curbs overconfident predictions* in deep neural networks, wherein the predicted (softmax) probabilities reflect true probabilities of correctness (better confidence estimates).

Label Smoothing

The one-hot encoded ground truth labels (y_k) are smoothened using a parameter α

$$y_k^{LS} = y_k(1-\alpha) + \alpha/K$$

where K is the number of classes. These smoothened targets y_k^{LS} and predicted outputs p_k are used to minimize the cross-entropy loss.

Mixup

A data augmentation method which is shown to output well-calibrated predictive scores.

$$\bar{x} = \lambda x_i + (1-\lambda)x_j$$
$$\bar{y} = \lambda y_i + (1-\lambda)y_j$$

where x_i and x_j are two randomly sampled input data points, and y_i and y_j are their respective one-hot encoded labels. Here, $\lambda \sim Beta(\alpha, \alpha)$ with $\lambda \in [0, 1]$.

Focal Loss

Calibrated probabilities are obtained by minimizing a regularized KL-divergence between the predicted and target distributions.

$$L_{Focal} = -(1-p)^{\gamma} logp$$

where p is the probability assigned by the model to the ground-truth correct class and γ is a hyperparameter.

Sample Prioritization

The process of selecting the most important samples/informative subsets during during training at each epoch.

Max Entropy: a de facto uncertainty sampling technique that selects the most informative samples (top-k) to maximize the predictive entropy.

$$\mathbb{H}[y|x, D_{train}] := -\sum_c p(y=c|x, D_{train}) \log p(y=c|x, D_{train})$$

Pre-trained Calibrated 'Target' Models

- Pre-trained models are used to obtain rich sample representations before training a downstream task.
- **Target** model – a pre-trained calibrated model with larger capacity
- **Current** model – model at hand which is being trained (with/without calibration)
- Sample prioritization with a pre-trained target model at each epoch **guides** the corresponding epochs of the current model's training process.
- Note – Sample prioritization with the calibrated target model performed in addition to calibrating the current model.

Experiments

- **Datasets** – CIFAR-10, CIFAR-100 (train: validation: test – 90: 10: 10)
- **Current** model => Resnet-34 (Label Smoothing, Mixup, Focal Loss)
- **Target** model => Resnet-50 with Mixup – CIFAR-10 (α=0.3), CIFAR-100 (α=0.25)
- During sample prioritization, start with *10 warm-up epochs* with all samples selected during training (no subset selection). Total training epochs – *200*.
- Then, select *n%* of total training samples in each epoch using the Max Entropy criterion. Subset sizes used for each epoch, *n – {10, 20, 30}*.
- **Evaluation Metrics** – Expected Calibration Error (ECE) and Accuracy
- SGD Optimizer; Learning rates – 0.01 (CIFAR-10) and 0.1 (CIFAR-100); Cosine annealing scheduler, Weight decay – 5e−4; Momentum – 0.9

Results

Table 1: Test Accuracies (%) and ECEs (%) across various calibration techniques and subset sizes with Resnet-34 as *current* model for both datasets.

Dataset	Calibration	100%		30%		20%		10%	
		Accuracy	ECE	Accuracy	ECE	Accuracy	ECE	Accuracy	ECE
CIFAR-10	**No Calibration** Cross-Entropy (Baseline)	94.1	4.1	93.6	5.33	**93.86**	4.01	**93.23**	5.2
	Label Smoothing 0.03/0.05/0.05/0.03	94	1.84	91.74	3.17	91.48	3.56	91.72	2.71
	Mixup 0.1/0.3/0.2/0.15	**95.1**	2.1	**94.39**	2.67	93.35	2.59	93.17	1.78
	Focal Loss 1/3/3/3	94.69	**1.71**	93.19	**1.2**	92.6	**1.25**	92.25	**1.42**
CIFAR-100	**No Calibration** Cross-Entropy (Baseline)	77.48	5.42	73.13	10.77	71.54	13.16	**69.65**	14.47
	Label Smoothing 0.03/0.03/0.03/0.09	77.05	4.88	72.21	3.45	70.93	5.75	68.63	5.67
	Mixup 0.15/0.15/0.15/0.35	**78.68**	3.59	**73.57**	**1.49**	**72.02**	**2.4**	69.1	**1.16**
	Focal Loss 1/3/3/5	78.59	**3.57**	71.86	1.67	70.61	3.25	65.81	1.82

Table 2: Test Accuracies (%) and ECEs (%) across various calibration techniques and subset sizes with Resnet-34 as *current* model for both datasets, and Resnet-50 (Mixup) as *target* model.

Dataset	Calibration	100%		30%		20%		10%	
		Accuracy	ECE	Accuracy	ECE	Accuracy	ECE	Accuracy	ECE
CIFAR-10	**No Calibration** Cross-Entropy (Baseline)	94.1	4.1	93.95	4.04	93.43	4.9	93.16	4.11
	Label Smoothing 0.03/0.05/0.05/0.03	94	1.84	93.62	2.93	93.3	3.32	**93.27**	1.9
	Mixup 0.1/0.3/0.15/0.15	**95.1**	2.1	**94.7**	2.88	**93.79**	2.73	93.22	2.16
	Focal Loss 1/2/2/1	94.69	**1.71**	93.15	**1.06**	92.65	**1.58**	92.84	**1.89**
CIFAR-100	**No Calibration** Cross-Entropy (Baseline)	77.48	5.42	75.38	9.36	75.04	9.39	71.07	9.27
	Label Smoothing 0.03/0.03/0.03/0.09	77.05	4.88	**76.06**	2.28	**75.27**	2.67	**72.59**	1.63
	Mixup 0.15/0.2/0.15/0.15	**78.68**	3.59	75.62	**0.86**	74.78	**1.43**	70.32	**0.86**
	Focal Loss 1/2/3/2	78.59	**3.57**	74.89	2.37	73.73	**1.43**	70.89	1.51

- **Calibration with sample prioritization => lower test ECEs across the board**
- **No significant trade-offs between accuracies and ECEs**
- Mixup consistently performs well (high accuracies, low ECEs), LS (least performance)
- **Performing calibration during training improves sample prioritization**
- **Target** – significant improvement over **current** (particularly for LS)

References

Guo et al. On Calibration of Modern Neural Networks. In ICML 2017.

Müller et al. When Does Label Smoothing Help? In NeurIPS 2019.

Thulasidasan et al. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks. In NeurIPS 2019.

Mukhoti et al. Calibrating Deep Neural Networks using Focal Loss. In NeurIPS 2020.

Hendrycks et al. Using Pre-Training Can Improve Model Robustness and Uncertainty. In ICML 2019.