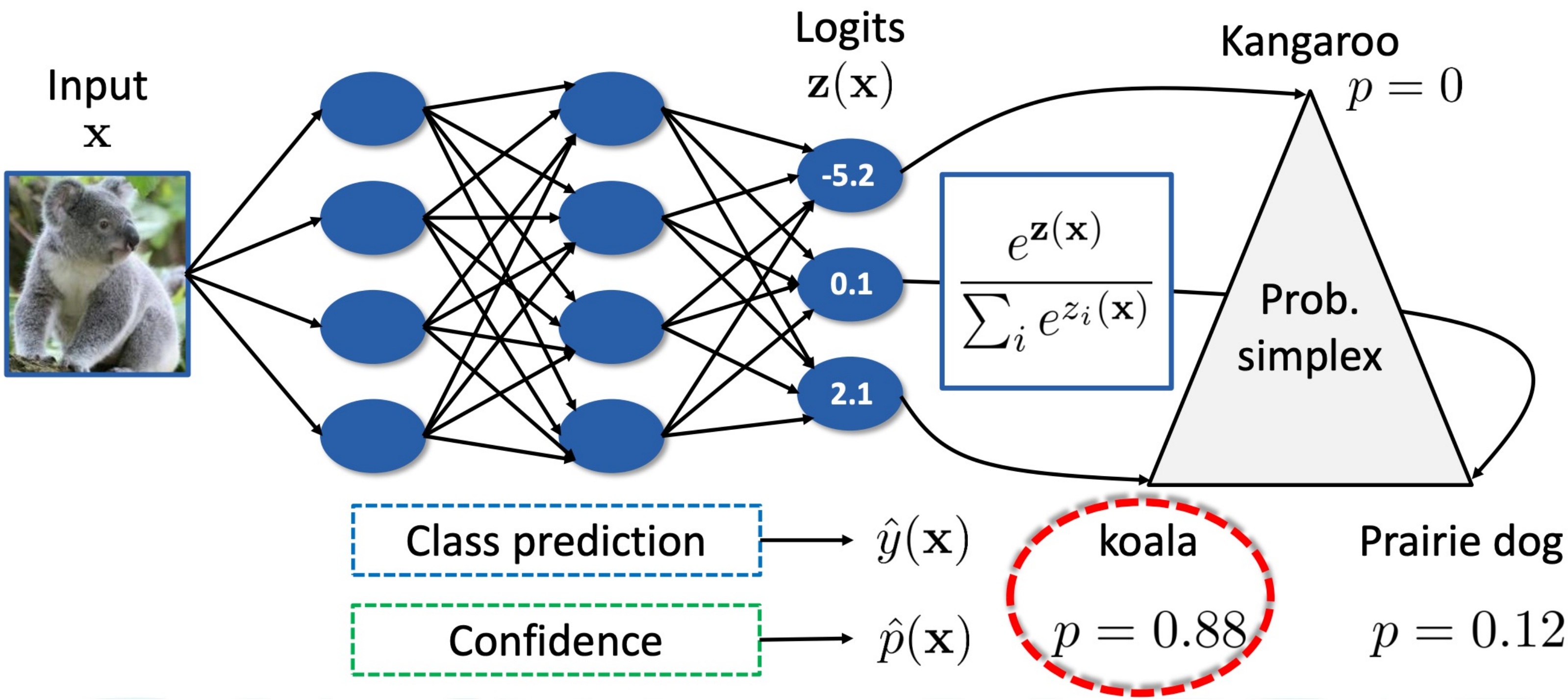
北京大學
PEKING UNIVERSITY

On Calibration of Modern Neural Networks

Citation 4294

ICML 2017

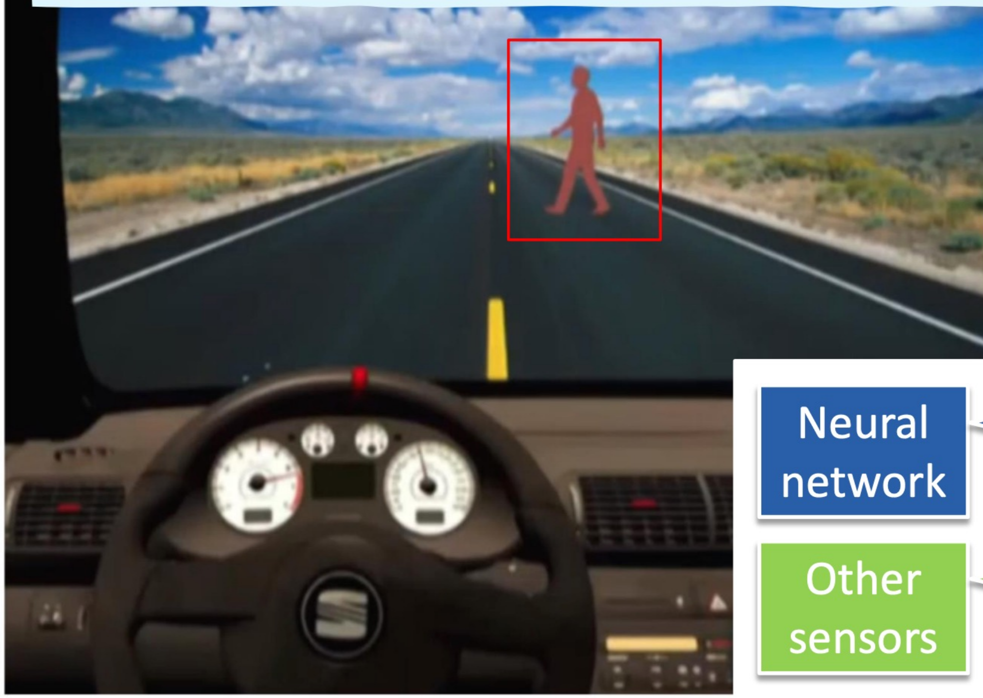
Motivation



In most cases, we only care about the accuracy without considering the confidence.

Motivation

The problem modern neural networks: overconfidence.



What happens if the confidence is 90% ?

Neural network → Plastic bag → 50% confidence

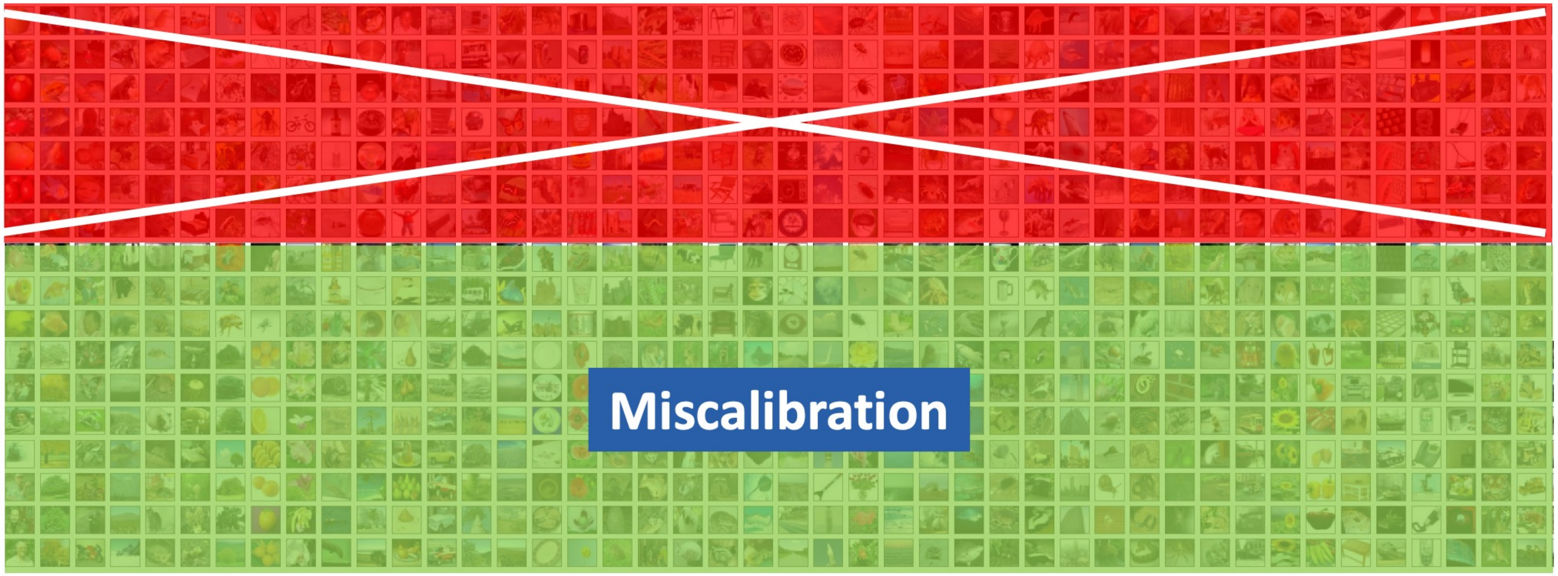
Other sensors → Person → 90% confidence

Confidence measures could be important in some real scenarios.

Motivation

The ResNet's accuracy is better but not match its confidence.

ResNet 101, Cifar 100
Samples with 80%-85% confidence



Overconfidence phenomenon occurs in DNNs.

Motivation

- Problems
  - Overconfidence (Miscalibration) phenomenon.
  - Confidence measures could be important in some real scenarios.

- Questions
  - How can we define/measure/visualize miscalibration?
  - What makes DNNs mis-calibrated?
  - How can we correct miscalibration?

The first question

How can we define/measure/visualize miscalibration?

Define miscalibration

- Perfectly calibrated model

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

- Expected Calibrated Error (ECE)

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

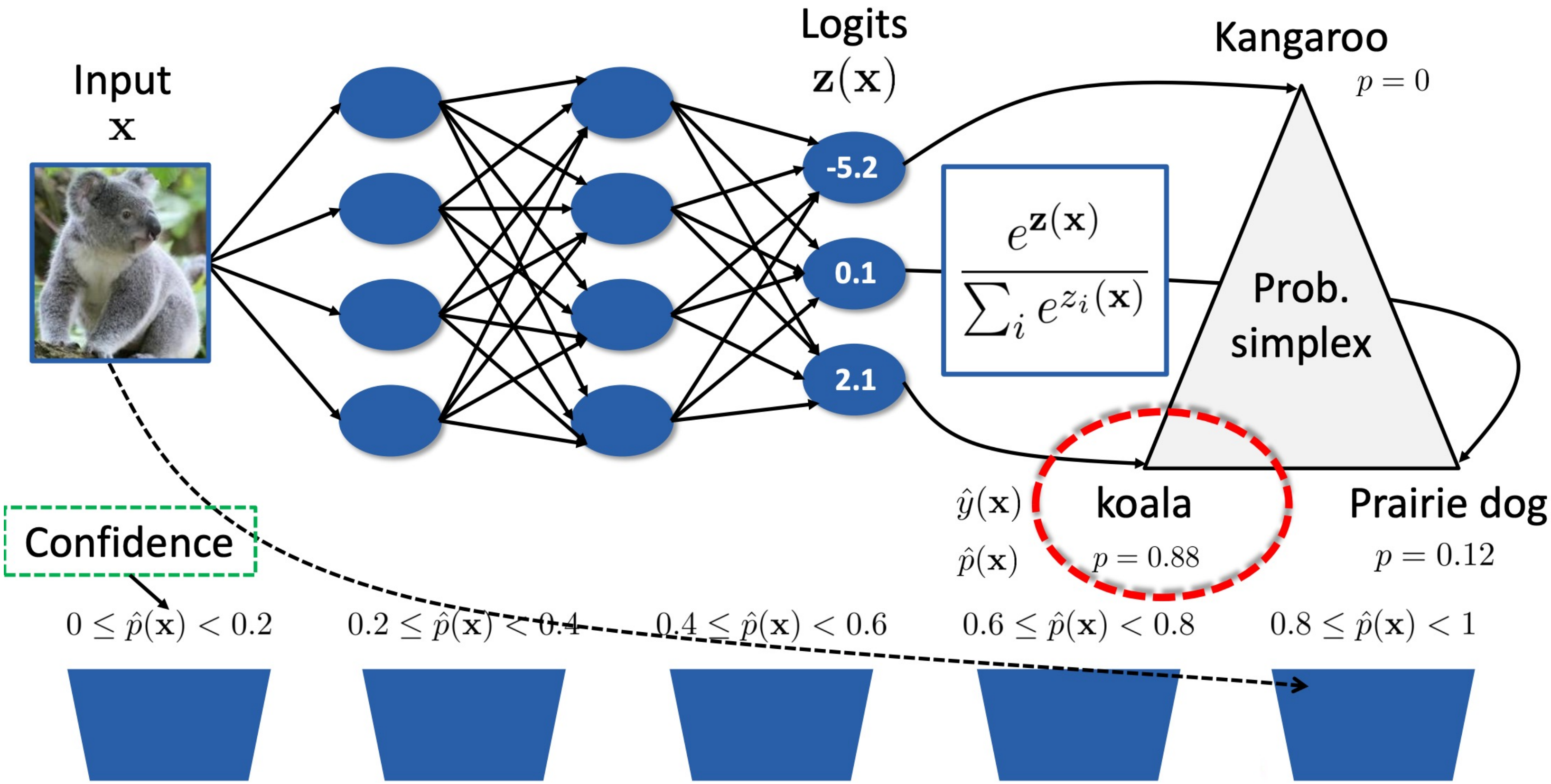
- ECE can be approximated by

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

where

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) \qquad \mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$

Measure miscalibration



Group by its prediction probability

Measure miscalibration

$0 \le \hat{p}(\mathbf{x}) < 0.2$	$0.2 \le \hat{p}(\mathbf{x}) < 0.4$	$0.4 \le \hat{p}(\mathbf{x}) < 0.6$	$0.6 \le \hat{p}(\mathbf{x}) < 0.8$	$0.8 \le \hat{p}(\mathbf{x}) < 1$

Avg conf:	0.55	0.74	0.86
— Accuracy:	0.50	0.67	0.71
Gap:	2 x 0.05	6 x 0.17	7 x 0.15

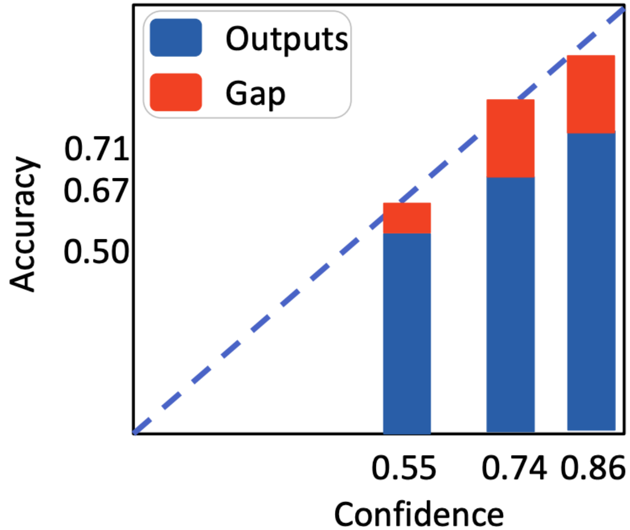
$$\frac{2 \times 0.05 + 6 \times 0.17 + 7 \times 0.15}{15}$$

Expected Calibrated Error (ECE) = 0.11

Visualize miscalibration

	0.55	0.74	0.86
Avg conf:	0.55	0.74	0.86
— Accuracy:	0.50	0.67	0.71
Gap:	0.05	0.17	0.15

Reliability Diagrams



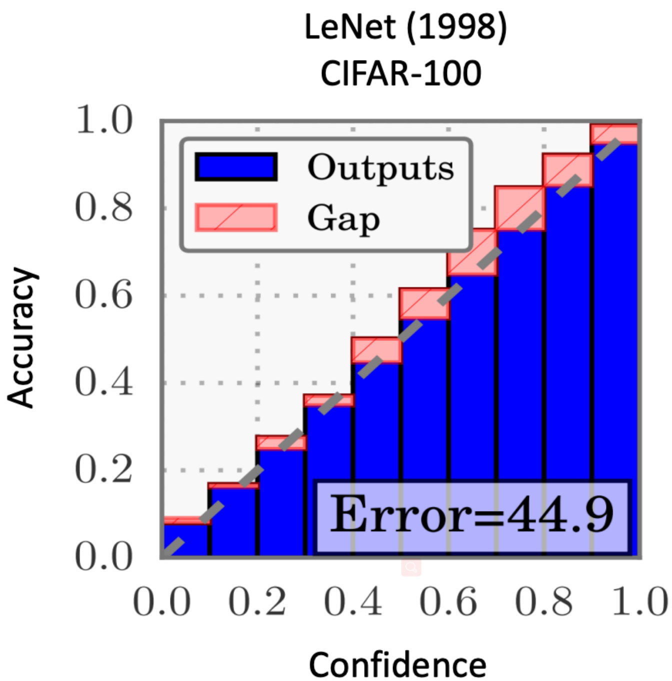
Niculescu-Mizil et al. Predicting good probabilities with supervised learning. ICML, 2005

The second question

What makes DNN mis-calibrated?

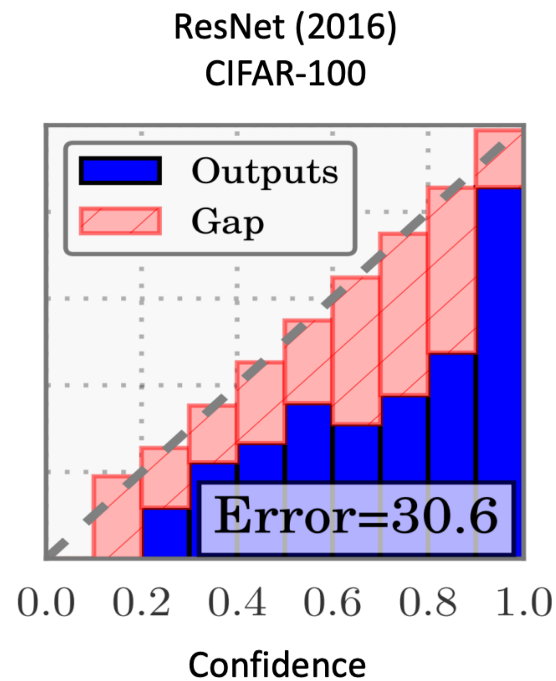
Neural network evaluation

2005: neural networks are calibrated.



Accuracy = 55.1%

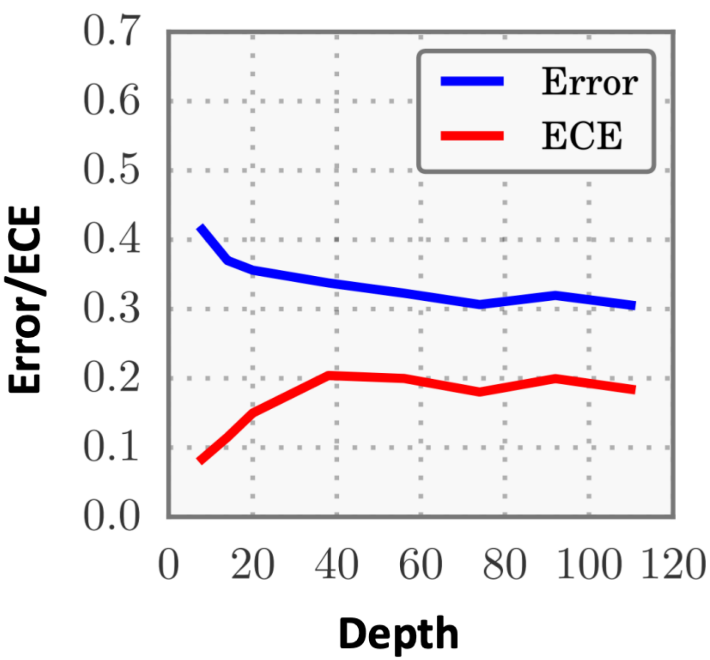
2017: neural networks are miscalibrated.



Accuracy = 69.4%

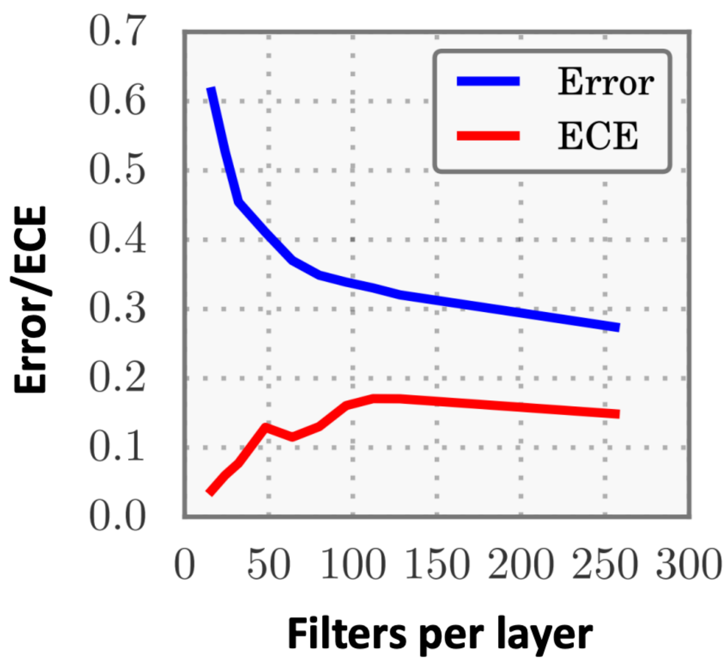
Increased network capacity

Varying Depth
ResNet - CIFAR100



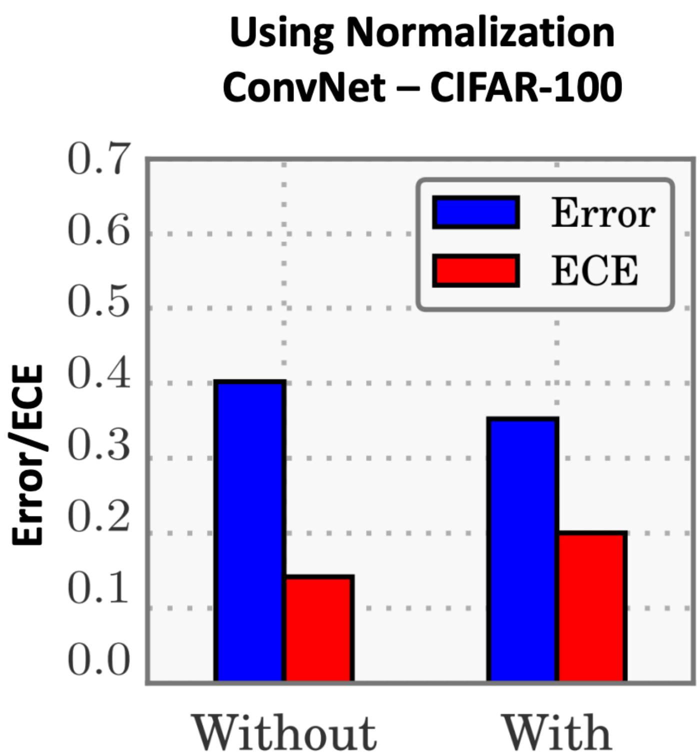
Fix filters per layer at 60

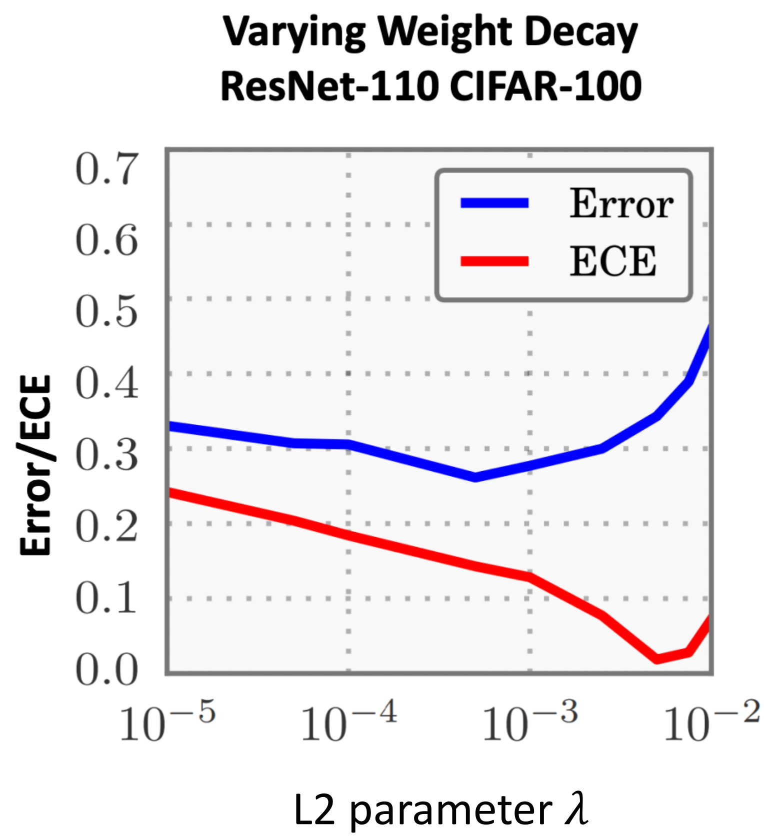
Varying Width
ResNet-14 - CIFAR100



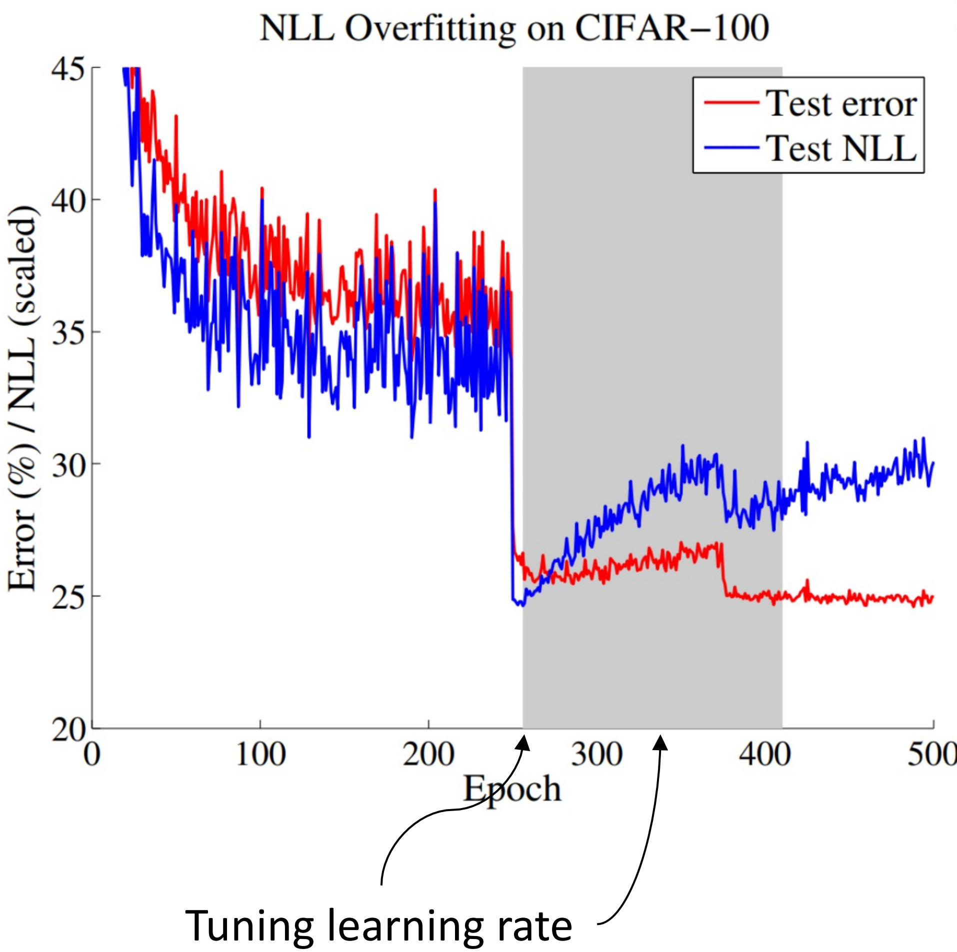
Fix the depth at 60

Batch normalization & Regularization





NLL overfitting



Negative log likelihood:

$$\mathcal{L} = -\sum_{i=1}^{n} \log(\hat{\pi}(y_i|\mathbf{x}_i))$$

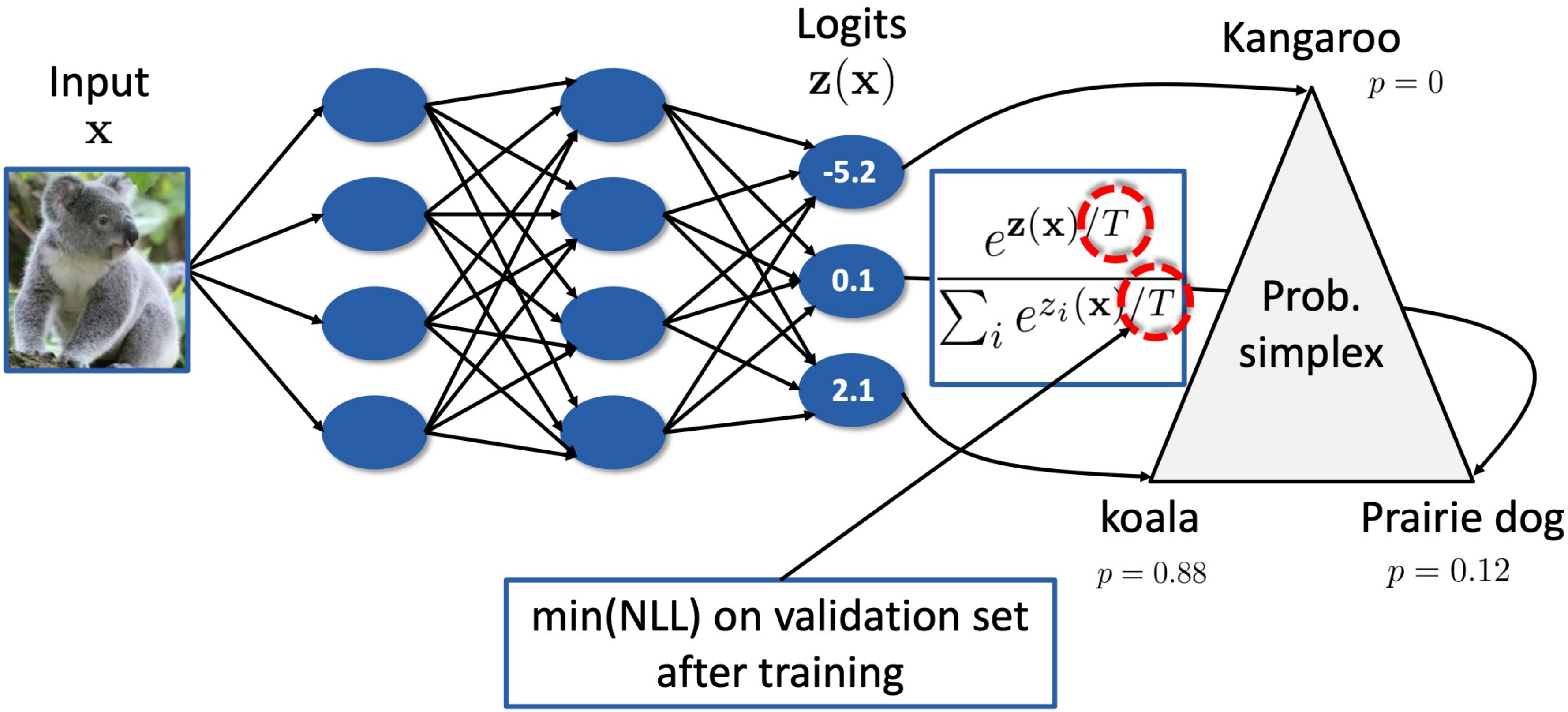
(x_i, y_i) *is sampling from the text set.*

The net work learns better classification accuracy at expense of well-modeled probabilities.

The third question

How can we correct miscalibration?

Temperature scaling



Temperature scaling does not affect the model's accuracy.

Experiments

Dataset	Model	Uncalibrated	Hist. Binning	Isotonic	BBQ	Temp. Scaling	Vector Scaling	Matrix Scaling
Birds	ResNet 50	9.19%	4.34%	5.22%	4.12%	**1.85%**	3.0%	21.13%
Cars	ResNet 50	4.3%	**1.74%**	4.29%	1.84%	2.35%	2.37%	10.5%
CIFAR-10	ResNet 110	4.6%	0.58%	0.81%	**0.54%**	0.83%	0.88%	1.0%
CIFAR-10	ResNet 110 (SD)	4.12%	0.67%	1.11%	0.9%	**0.6%**	0.64%	0.72%
CIFAR-10	Wide ResNet 32	4.52%	0.72%	1.08%	0.74%	**0.54%**	0.6%	0.72%
CIFAR-10	DenseNet 40	3.28%	0.44%	0.61%	0.81%	**0.33%**	0.41%	0.41%
CIFAR-10	LeNet 5	3.02%	1.56%	1.85%	1.59%	**0.93%**	1.15%	1.16%
CIFAR-100	ResNet 110	16.53%	2.66%	4.99%	5.46%	**1.26%**	1.32%	25.49%
CIFAR-100	ResNet 110 (SD)	12.67%	2.46%	4.16%	3.58%	0.96%	**0.9%**	20.09%
CIFAR-100	Wide ResNet 32	15.0%	3.01%	5.85%	5.77%	**2.32%**	2.57%	24.44%
CIFAR-100	DenseNet 40	10.37%	2.68%	4.51%	3.59%	1.18%	**1.09%**	21.87%
CIFAR-100	LeNet 5	4.85%	6.48%	2.35%	3.77%	**2.02%**	2.09%	13.24%
ImageNet	DenseNet 161	6.28%	4.52%	5.18%	3.51%	**1.99%**	2.24%	-
ImageNet	ResNet 152	5.48%	4.36%	4.77%	3.56%	**1.86%**	2.23%	-
SVHN	ResNet 152 (SD)	0.44%	**0.14%**	0.28%	0.22%	0.17%	0.27%	0.17%
20 News	DAN 3	8.02%	**3.6%**	5.52%	4.98%	4.11%	4.61%	9.1%
Reuters	DAN 3	0.85%	1.75%	1.15%	0.97%	0.91%	**0.66%**	1.58%
SST Binary	TreeLSTM	6.63%	1.93%	**1.65%**	2.27%	1.84%	1.84%	1.84%
SST Fine Grained	TreeLSTM	6.71%	2.09%	**1.65%**	2.61%	2.56%	2.98%	2.39%

Table 1. ECE (%) (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.

Experiments

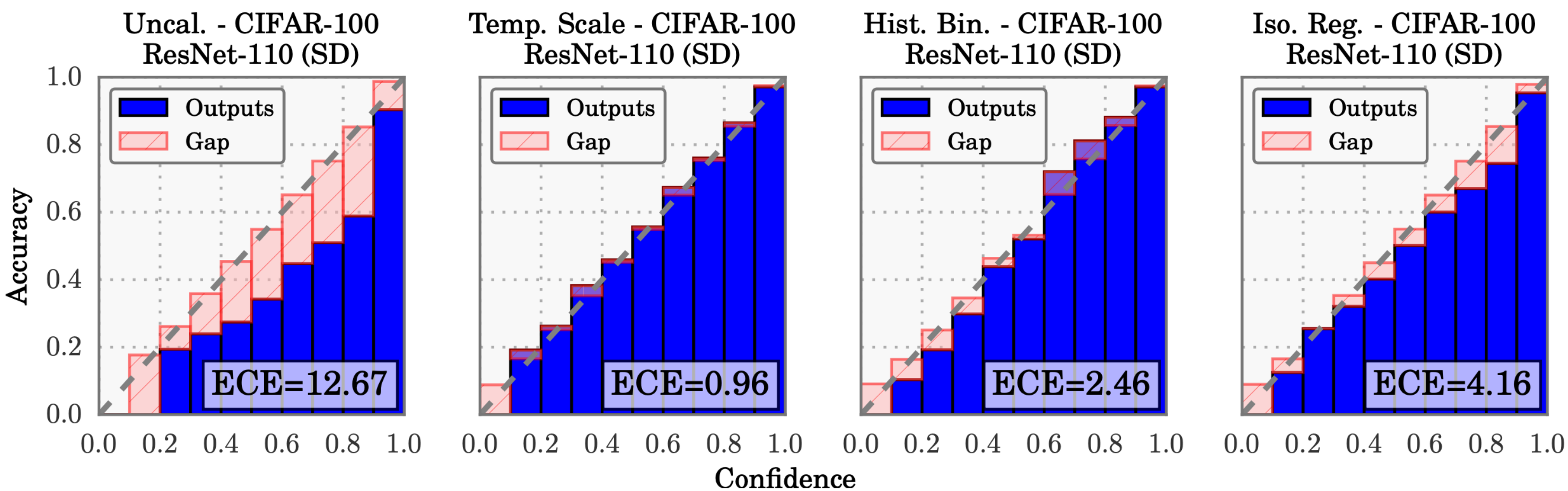


Figure 4. Reliability diagrams for CIFAR-100 before (far left) and after calibration (middle left, middle right, far right).

Awesome topics

- Traditional ML
  - ✓ When Easy and Hard Examples Meet Noisy Labels. (AAAI'24 Under review)
  - ◦ Aleatoric and epistemic uncertainty & Open-set Annotation. (CVPR'22 → TPAMI)
  - ◦ SBN & (Test-time Adaption) TTA & Active Learning & Noisy Labels …

- Data-centric AI
  - ◦ Active Learning for tuning LLMs. (→ IJCAI'24)
  - ◦ How to Generate the Best Prompts for Fine-Tuning.
  - ◦ ……

- LLMs-Attack
  - ◦ OOD & Hallucination attack. (→ ICLR'24)
  - ◦ Token & Sentence semantics attack. (→ ICLR'24)

- AI Generated Automation (AIGA)
  - ◦ DeMO: Large Decision Model. (……)
  - ◦ Thinking Hierarchy in LLMs. (…….)

Thanks