南京航空航天大學

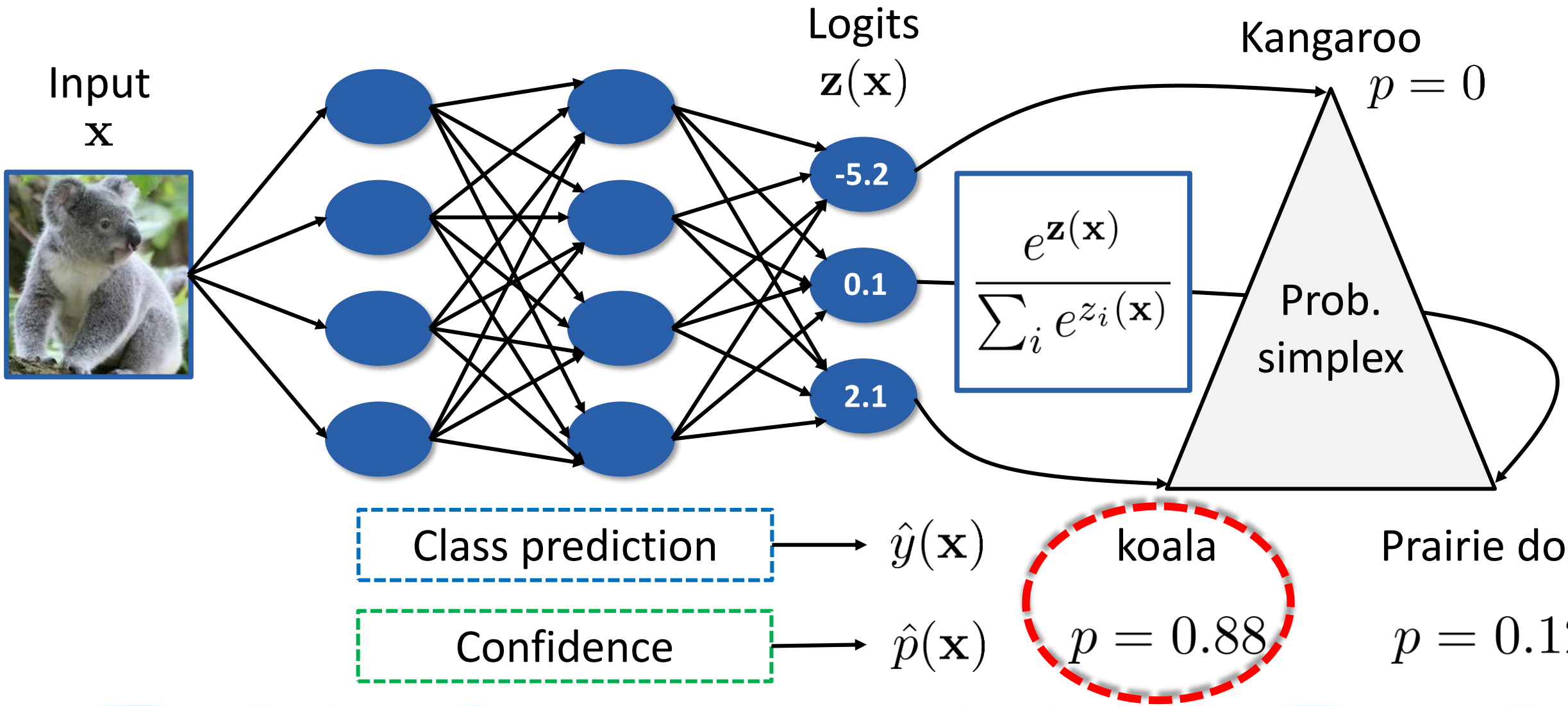
Nanjing University of Aeronautics and Astronautics

On Calibration of Modern Neural Networks

Chuan Guo [*1] **Geoff Pleiss** [*1] **Yu Sun** [*1] **Kilian Q. Weinberger** [1]

ICML 2017

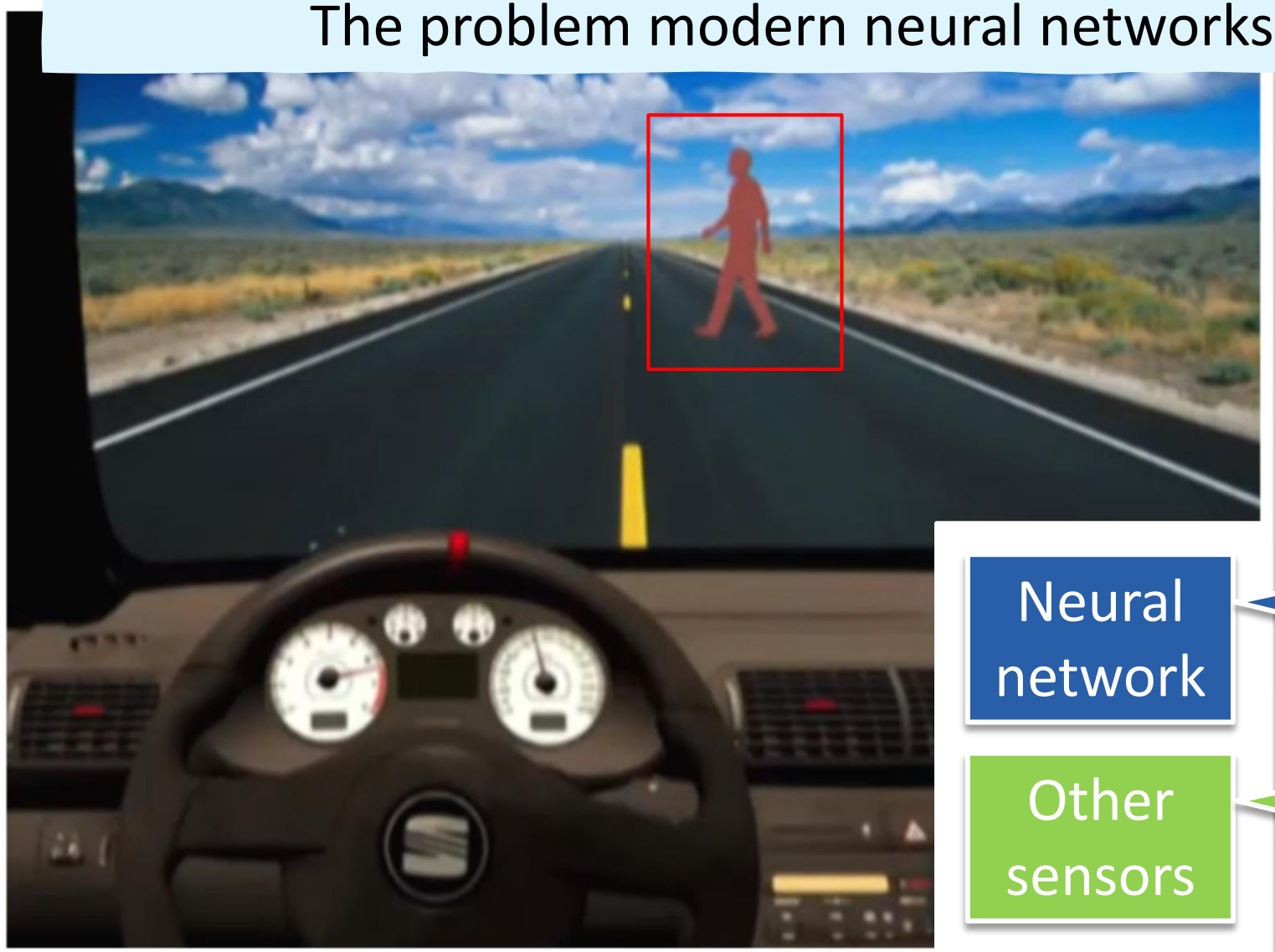
Notation



Confidence measures could be important in some real scenarios.

An illustration

The problem modern neural networks: overconfidence.



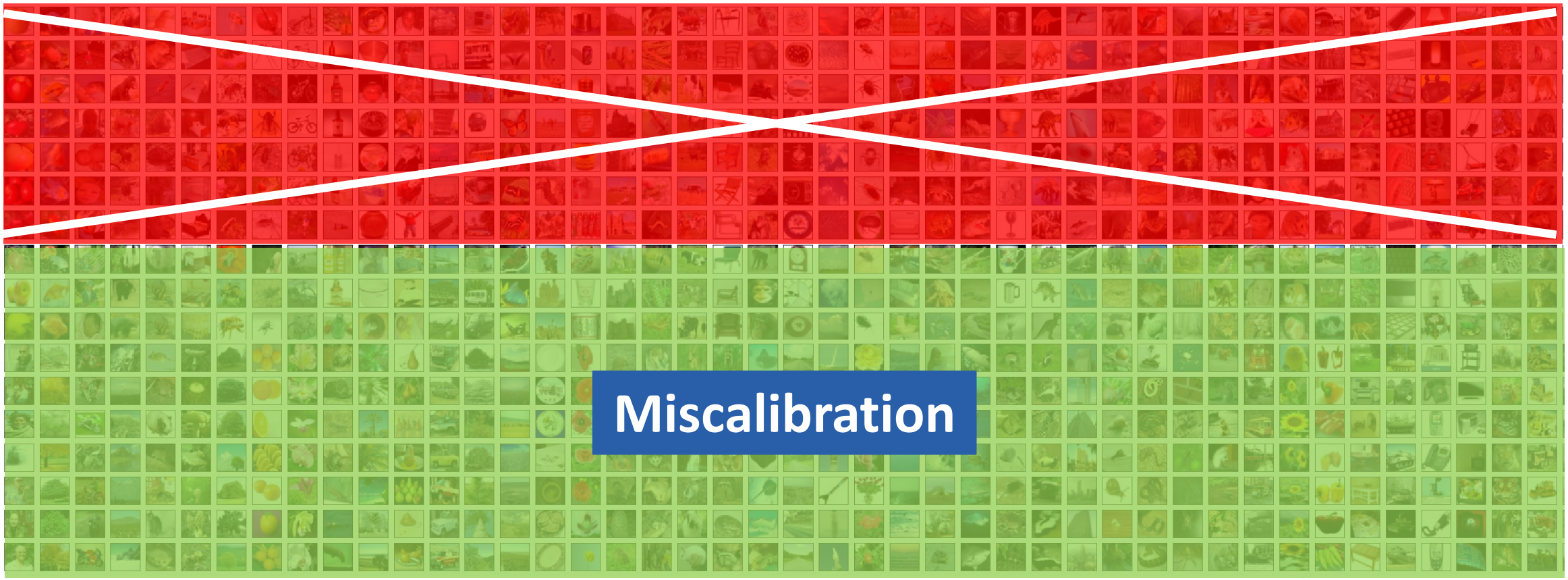
What happens if the confidence is 90% ?

Neural network	Plastic bag	50% confidence
Other sensors	Person	90% confidence

Miscalibration

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

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



Outline

How can we measure/visualize miscalibration ?

What makes neural networks miscalibrated ?

How can we correct miscalibration ?

The first question

How can we measure/visualize miscalibration ?

Measure miscalibration

Input
$\mathbf{x}$

Logits
$\mathbf{z}(\mathbf{x})$

-5.2

0.1

2.1

$$\frac{e^{\mathbf{z}(\mathbf{x})}}{\sum_i e^{z_i(\mathbf{x})}}$$

Kangaroo
$p = 0$

Prob. simplex

$\hat{y}(\mathbf{x})$ koala

$\hat{p}(\mathbf{x})$ $p = 0.88$

Prairie dog
$p = 0.12$

Confidence

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

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

Expected Calibrated Error (ECE)

- Perfectly calibrated model:

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

- Miscalibration is difference between expected confidence and accuracy (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:

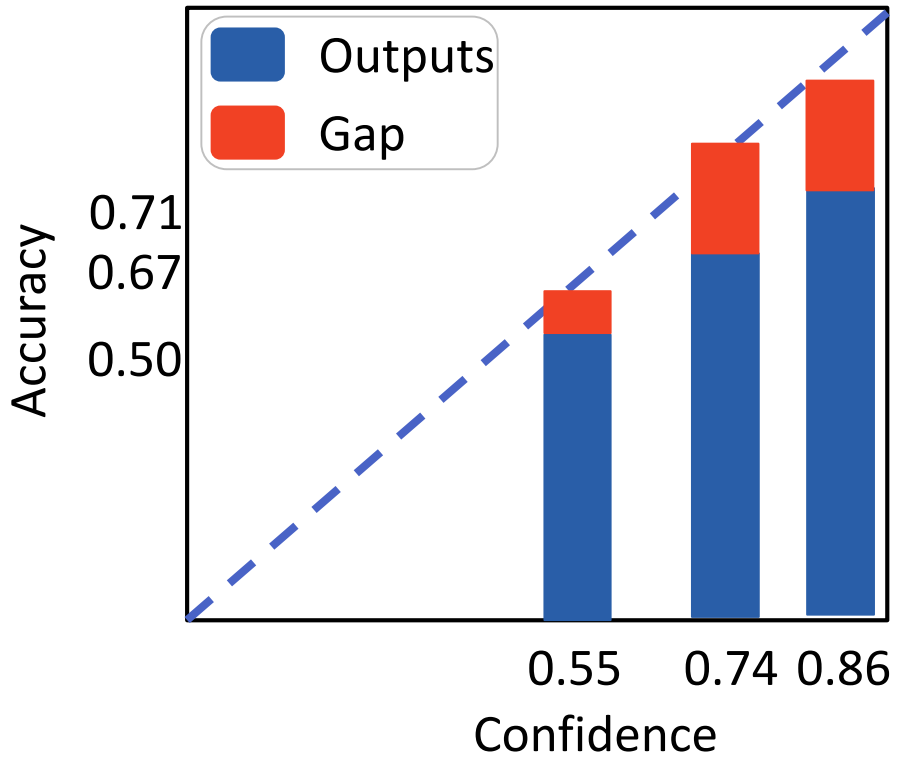
$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) \qquad \mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$

Naeini et al. Obtaining well calibrated probabilities using bayesian binning. AAAI, 2015

Visualize miscalibration

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

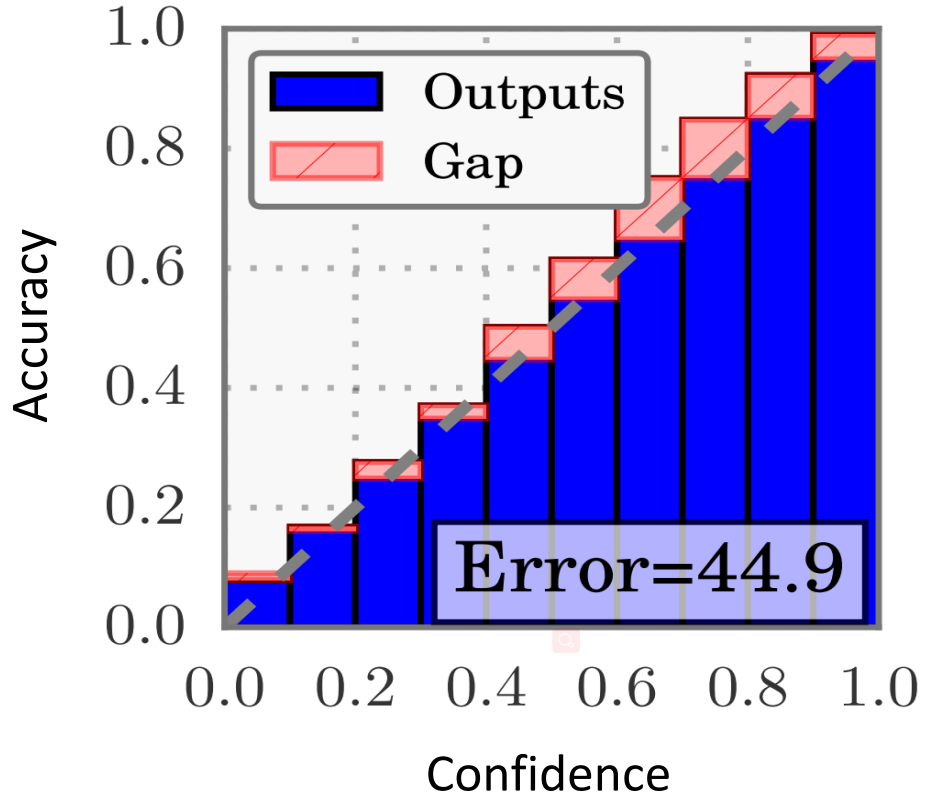
The second question

What makes neural networks miscalibrated ?

Neural network evaluation

2005: neural networks are calibrated.

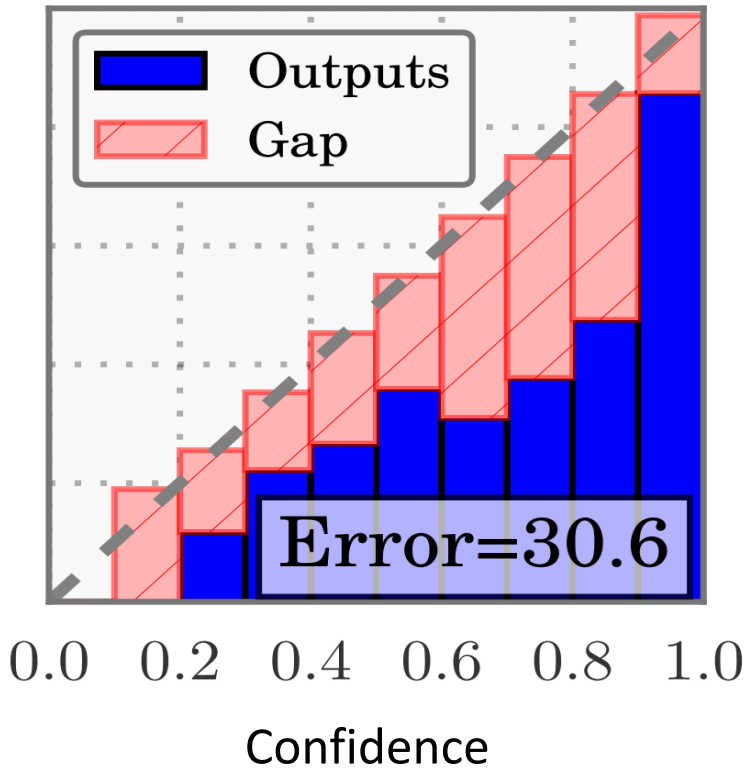
LeNet (1998)
CIFAR-100



Accuracy = 55.1%

2017: neural networks are miscalibrated.

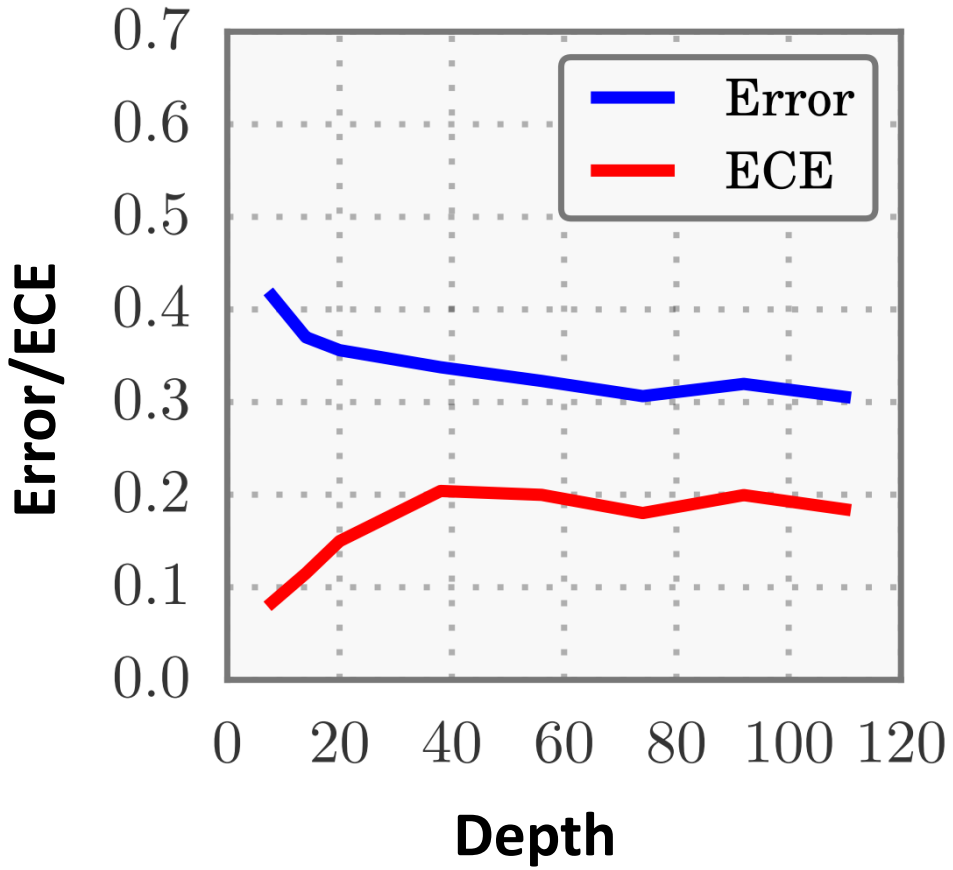
ResNet (2016)
CIFAR-100



Accuracy = 69.4%

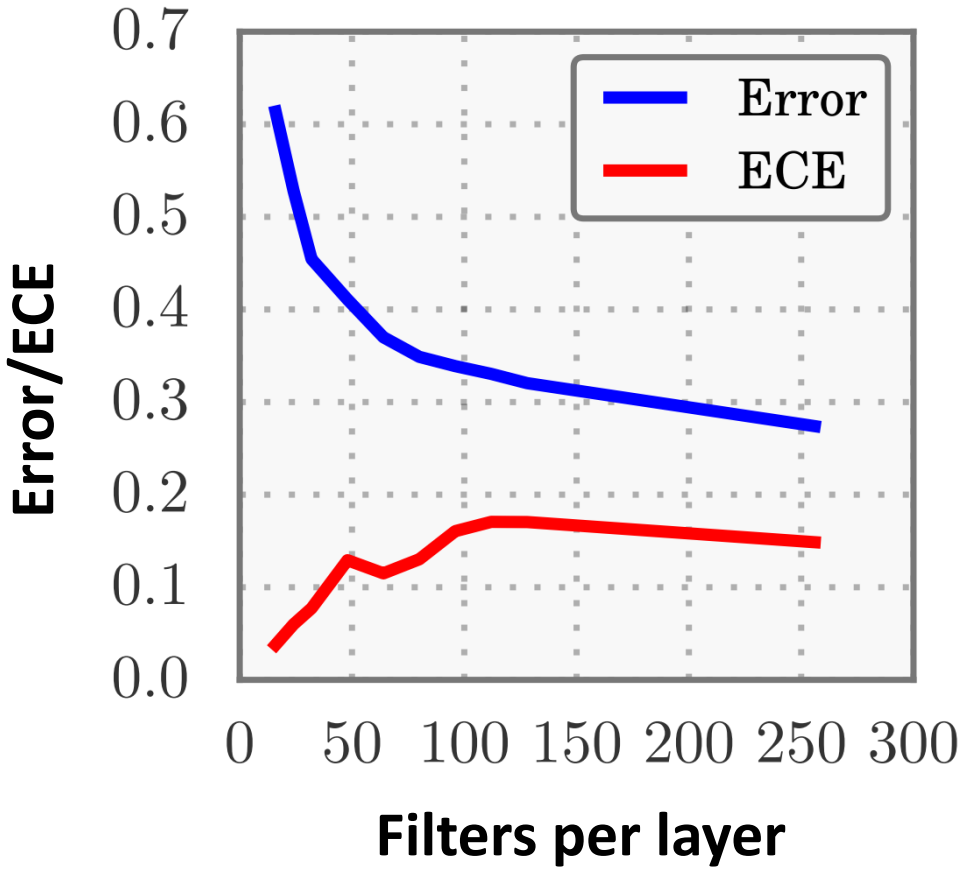
Increased network capacity

Varying Depth
ResNet - CIFAR100



Fix filters per layer at 60

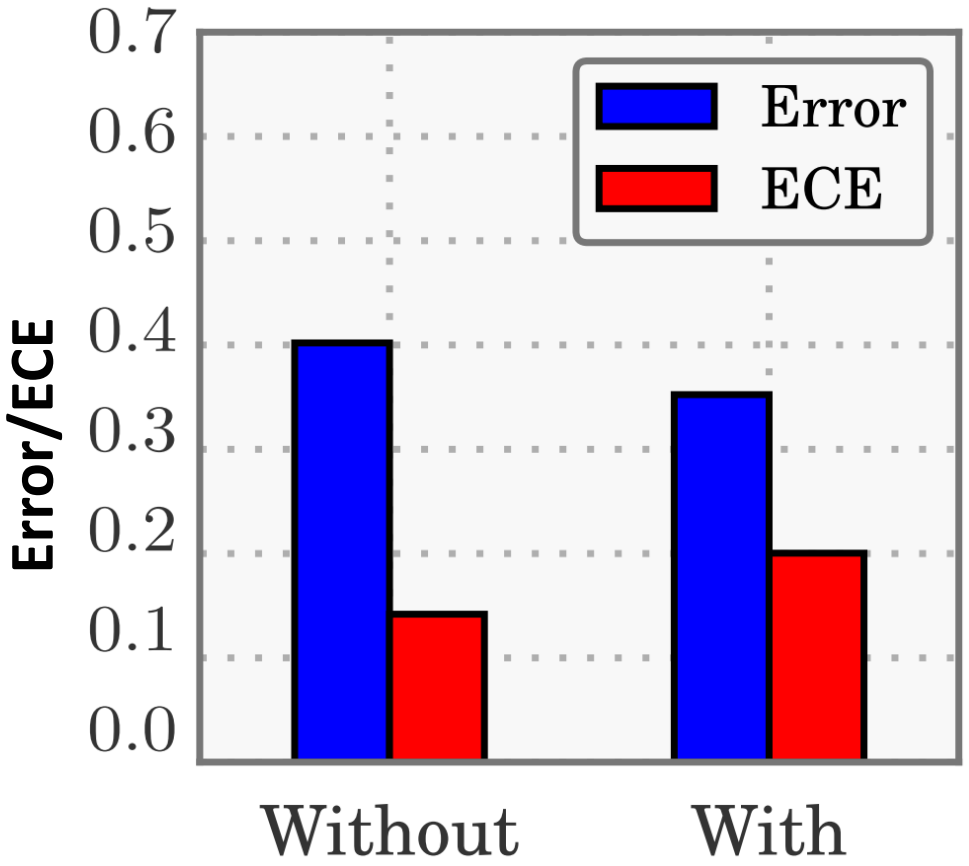
Varying Width
ResNet-14 - CIFAR100



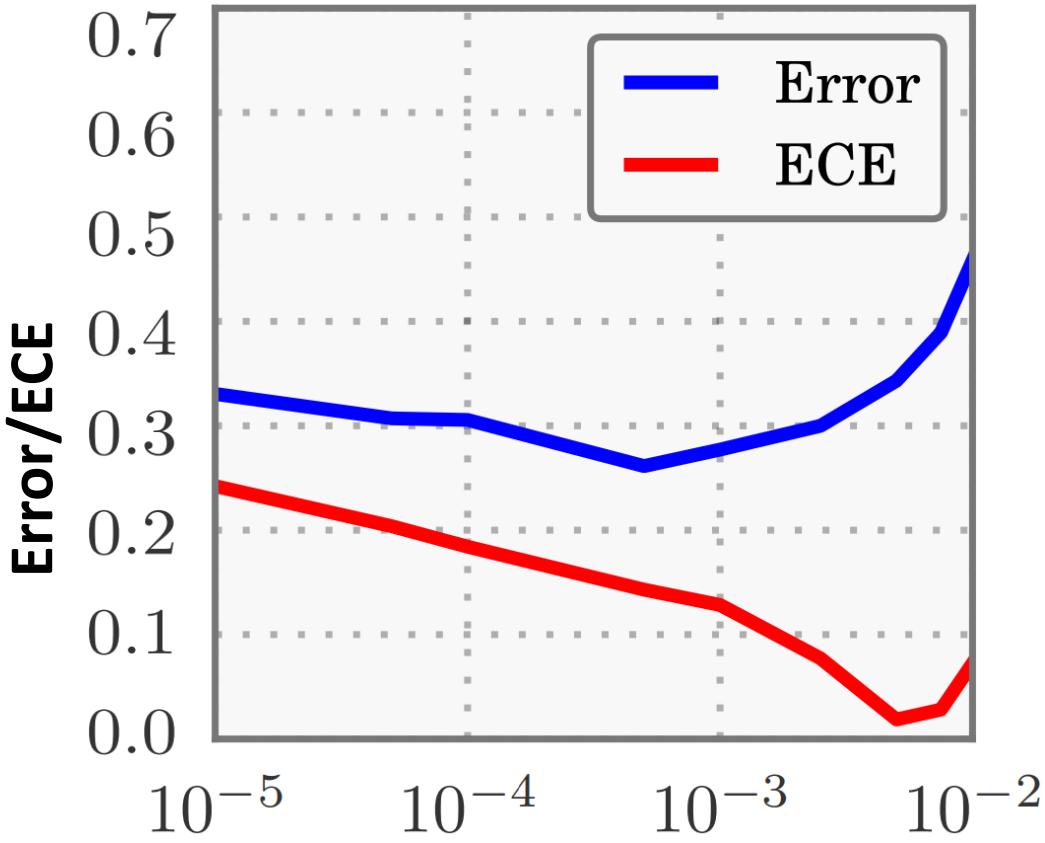
Fix the depth at 60

Batch normalization & Less regularization

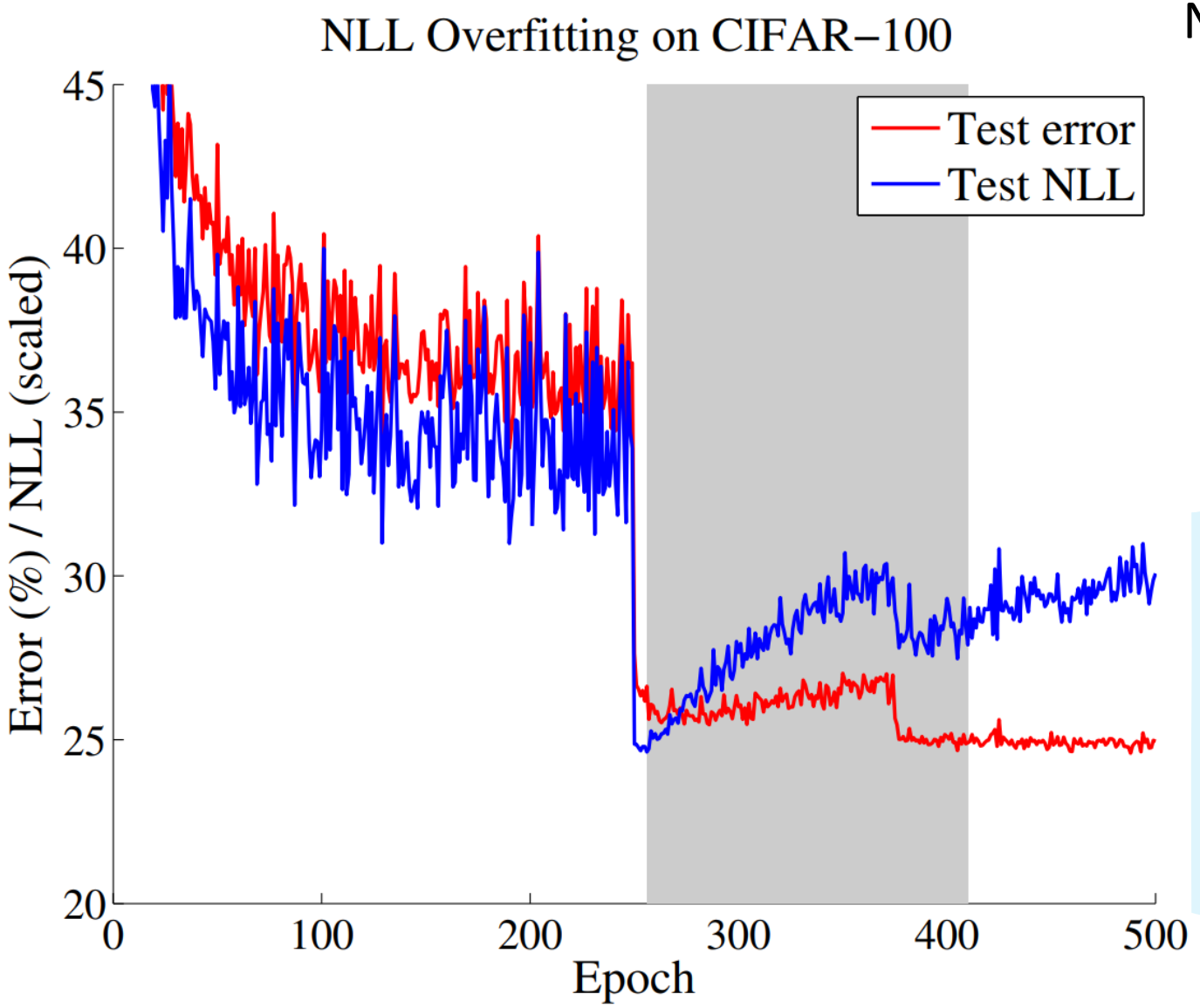
Using Normalization
ConvNet – CIFAR-100



Varying Weight Decay
ResNet-110 CIFAR-100



NLL overfitting



Negative log likelihood:

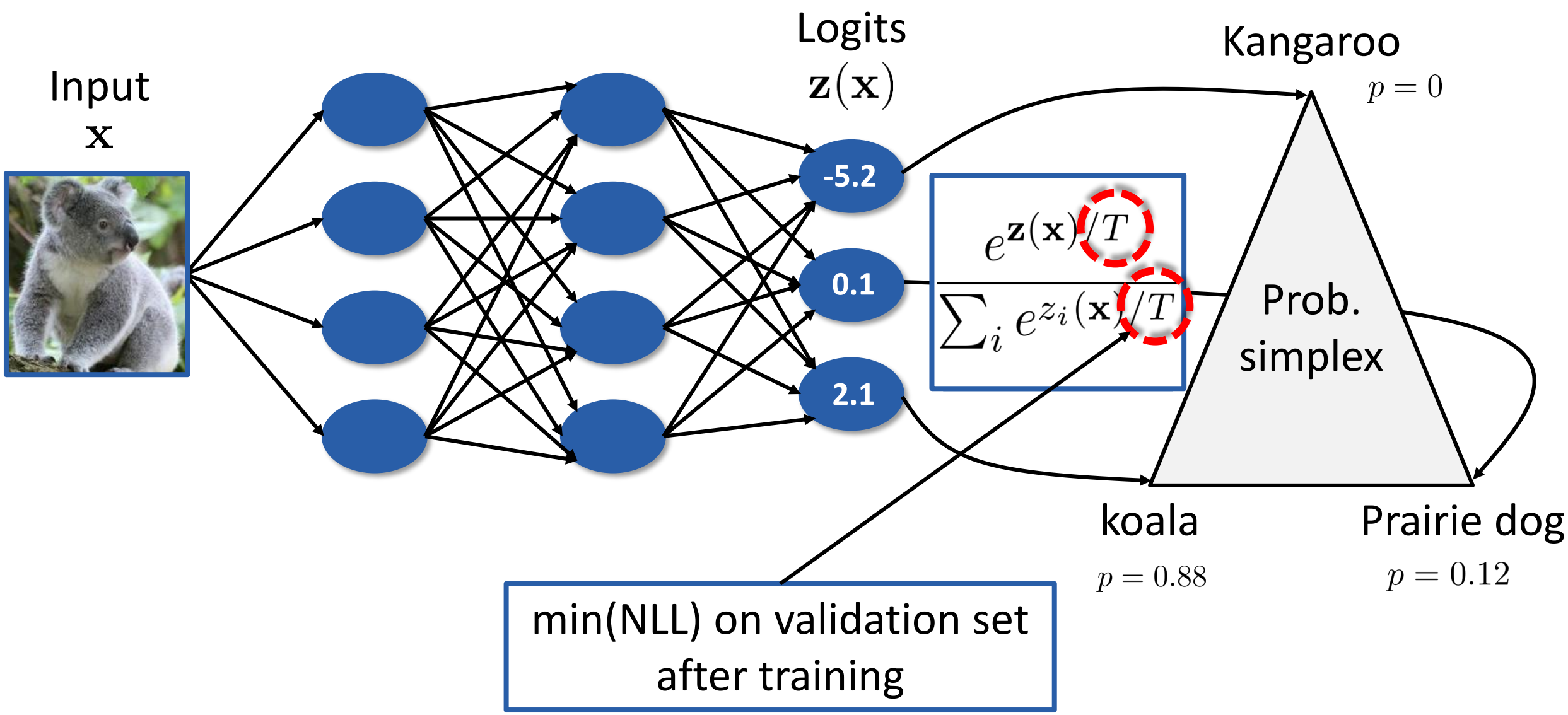
$$\mathcal{L} = -\sum_{i=1}^{n} \log(\hat{\pi}(y_i|\mathbf{x}_i))$$

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

- Class prediction: $\hat{y}_i = \text{argmax}_k \, z_i^{(k)}$
- Confidence score: $\hat{p}_i = \max_k \, \sigma_{\text{SM}}(\mathbf{z}_i)^{(k)} \quad \sigma_{\text{SM}}(\mathbf{z}_i)^{(k)} = \frac{\exp(z_i^{(k)})}{\sum_{j=1}^{K} \exp(z_i^{(j)})}$
- Calibrated confidence: $\hat{q}_i = \max_k \, \sigma_{\text{SM}}(\mathbf{z}_i / T)^{(k)}$

Temperature

min(NLL) on validation set after training

- Temperature scaling does not affect the model's accuracy.

Calibration results

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.

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%

Calibration results

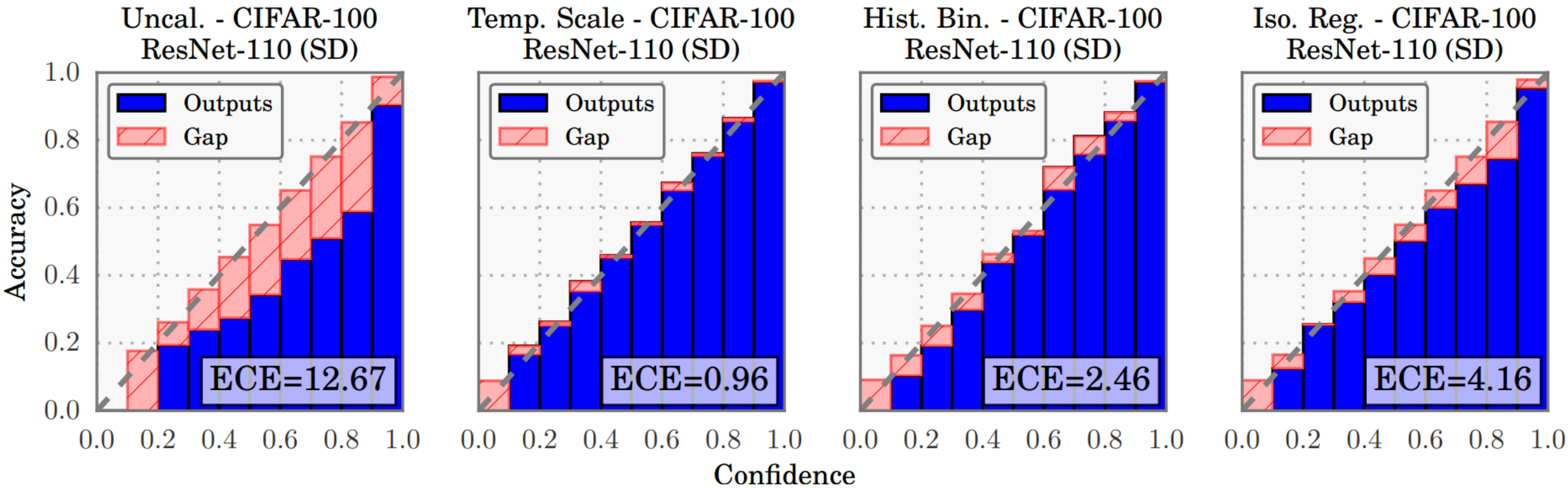


Figure 4. Reliability diagrams for CIFAR-100 before (far left) and after calibration (middle left, middle right, far right).

Conclusion

- Modern neural network tend to be miscalibrated.
- Factors: model capacity, batch normalization, weight decay.
- Simple solution: temperature scaling.

THANKS