ICML 2017

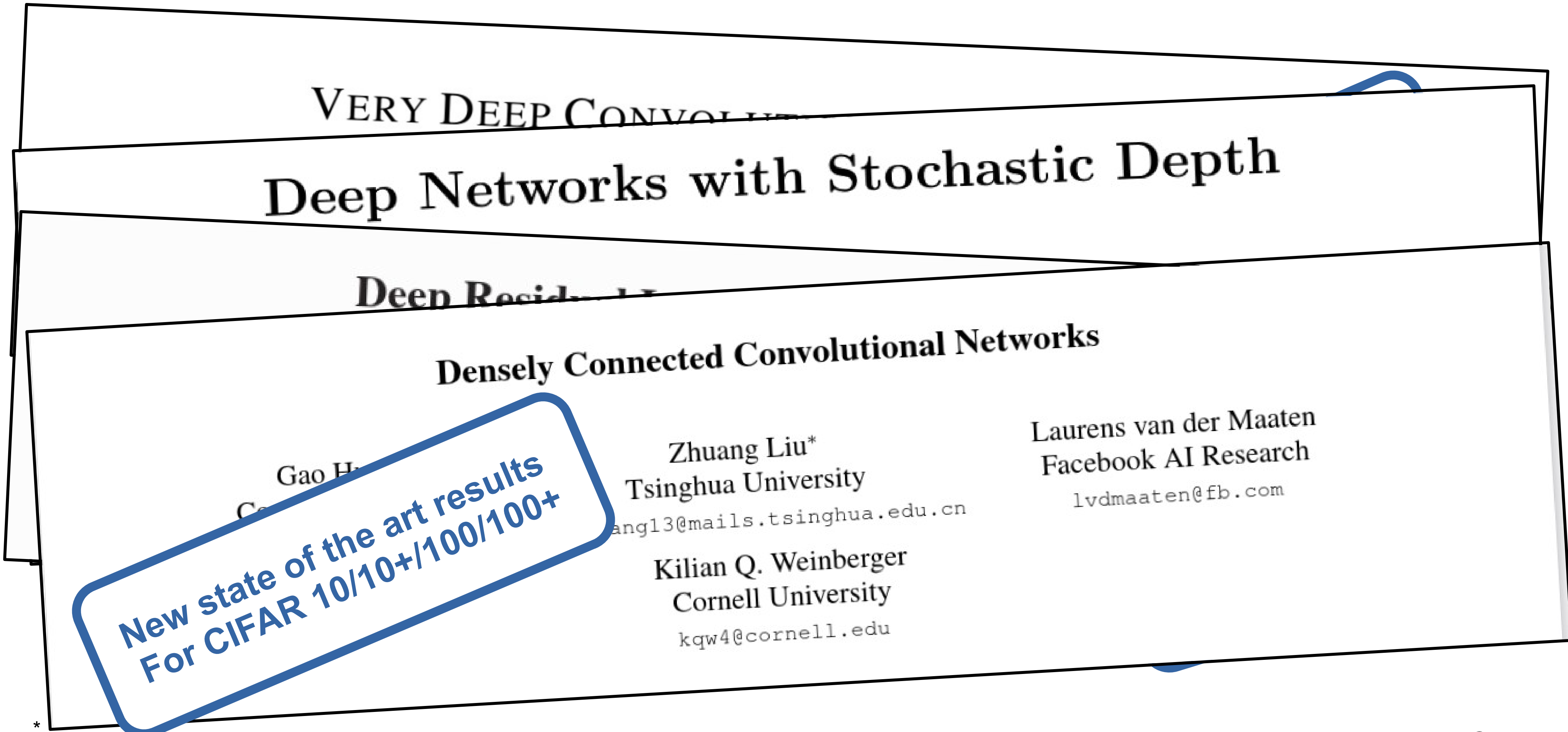
On Calibration of Modern Neural Networks

Chuan Guo Geoff Pleiss Yu Sun

Kilian Q. Weinberger

Presented by Lukas Fluri

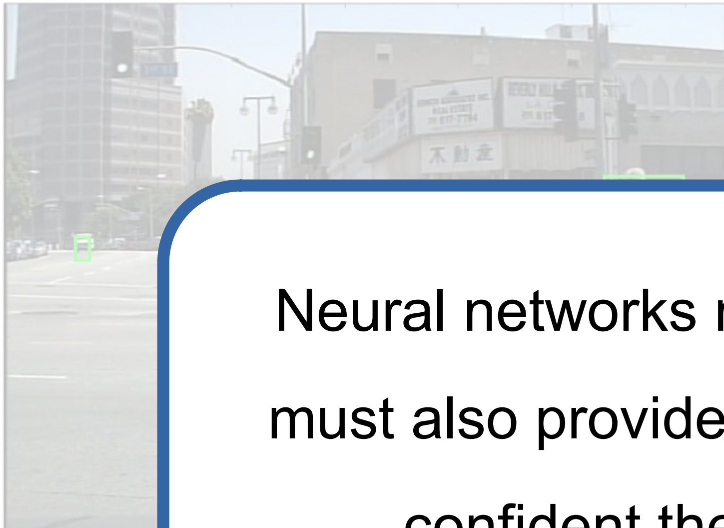
Introduction



* Under certain assumptions

** As of 2020

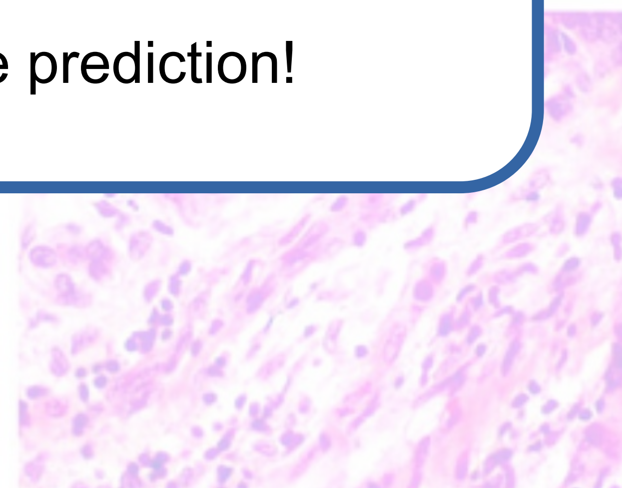
Introduction



Neural networks must not only be accurate, they must also provide a reliable estimation about how confident they are about the prediction!

Cancer detection

Survey: [Daoud Artif. Intell. Med. 2019]

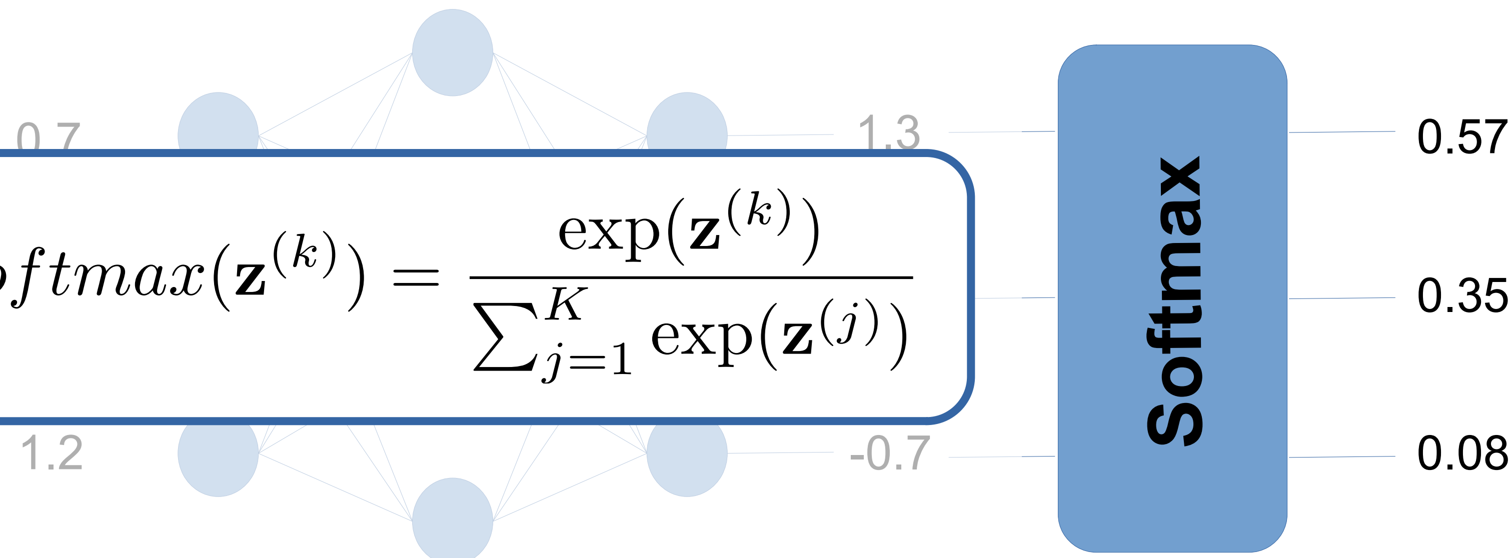


Source: [Spanhol et al. IEEE Trans Biomed 2016]

Overview

- **Introduction:** It's important for neural networks to be well-calibrated.
- **Definition:** How to measure model calibration?
- **Problem:** Modern neural networks are no longer calibrated!
- **Analysis:** Which factors might influence model calibration?
- **Mitigation:** How to calibrate neural networks?
- **Experiments:** Which calibration methods perform best?

How to create confidence estimates



$$softmax(\mathbf{z}^{(k)}) = \frac{\exp(\mathbf{z}^{(k)})}{\sum_{j=1}^{K} \exp(\mathbf{z}^{(j)})}$$

Input: $\mathbf{x}$

Output: $\mathbf{z} = NN_{\Theta}(\mathbf{x})$

Confidence: $\hat{\mathbf{p}} = softmax(\mathbf{z})$

How to interpret calibration

Input		Pred.	Conf.	
x_1	→	CAT	70%	✓
x_2	→	DOG	70%	✗
x_3	→	CAT	70%	✓
x_4	→	CAT	70%	✓
x_5	→	DOG	70%	✓
x_6	→	CAT	70%	✗
x_7	→	CAT	70%	✓
x_8	→	DOG	70%	✗
x_9	→	CAT	70%	✓
x_{10}	→	DOG	70%	✓

Different sources of error

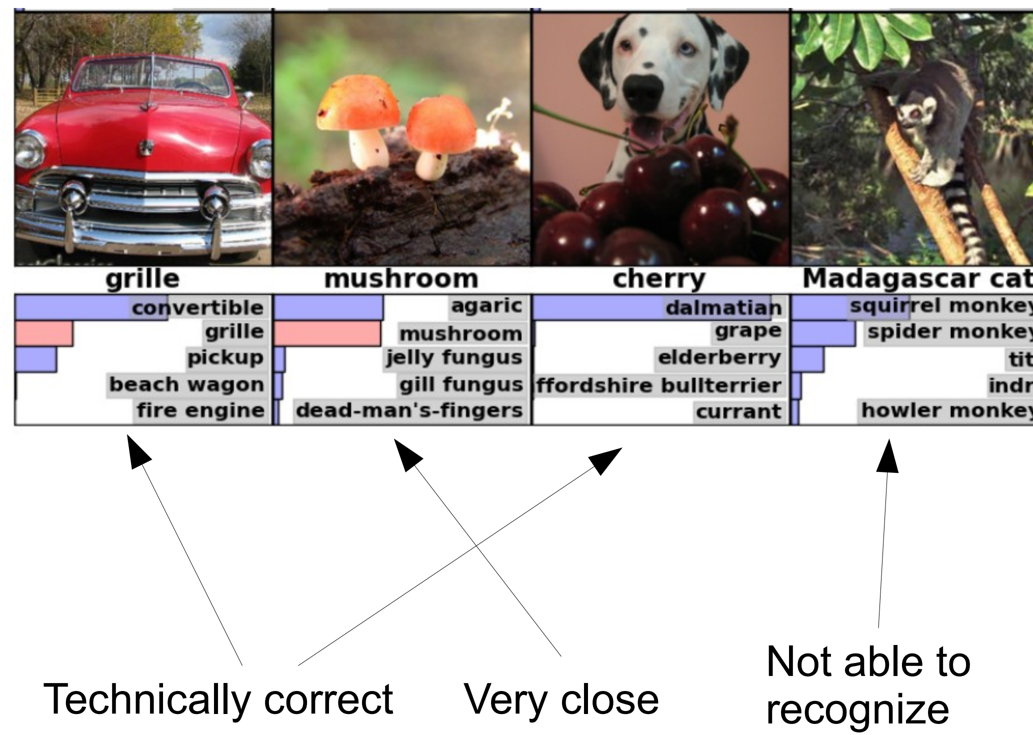


Image source: Krizhevsky et al. “ImageNet Classification...”, 2012

How to define model-calibration

Input		Pred.	Conf.	True	
x_1	$\longrightarrow$	CAT	70%	CAT	✓
x_2	$\longrightarrow$	DOG	70%	CAT	✗
x_3	$\longrightarrow$	CAT	70%	CAT	✓
x_4	$\longrightarrow$	CAT	70%	CAT	✓
x_5	$\longrightarrow$	DOG	70%	DOG	✓
x_6	$\longrightarrow$	CAT	70%	DOG	✗
x_7	$\longrightarrow$	CAT	70%	CAT	✓
$\dots$			$\dots$		
		$\hat{Y}$	$\hat{P}$	Y	

Perfect calibration

A neural network has *perfect calibration* if for all $p \in [0,1]$:

$$\mathbb{P}\left(\hat{Y} = Y \,|\, \hat{P} = p\right) = p$$

How to define model-calibration

- Perfect calibration: $\mathbb{P}\left(\hat{Y} = Y | \hat{P} = p\right) \;=\; p \quad \forall p \in [0,1]$

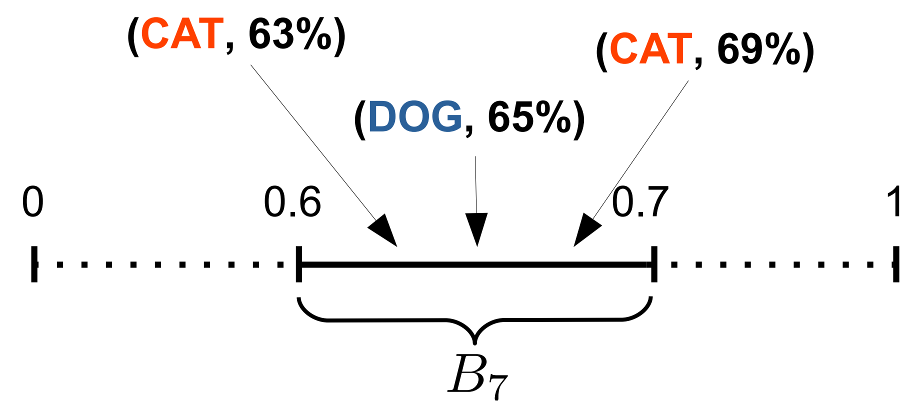
- Model calibration: $\mathbb{E}\left[\left|\mathbb{P}\left(\hat{Y} = Y | \hat{P} = p\right) - p\right|\right]$

Problem: In practice we only have finite data!
We need to approximate the model calibration

How to define model-calibration

- Expected Calibration Error (ECE):
  1. Train neural network on training data
  2. Create predictions and confidence estimates using the test data
  3. Group the predictions into M bins. Define bin B_m to be the set of all predictions $(\hat{y}_i, \hat{p}_i)$ for which it holds that

$$\hat{p}_i \in \left(\frac{m-1}{M}, \frac{m}{M} \right]$$



How to define model-calibration

- Expected Calibration Error (ECE):

4. Compute the accuracy and confidence of bin B_m as:

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

5. Compute the expected calibration error as:

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

How to define model-calibration

- Perfect calibration: $\mathbb{P}\left(\hat{Y} = Y | \hat{P} = p\right) \; = \; p \quad \forall p \in [0, 1]$

Diff. from perfect calibration

- Model calibration: $\mathbb{E}\left[\left|\mathbb{P}\left(\hat{Y} = Y | \hat{P} = p\right) - p\right|\right]$

Finite approximation

- Expected Calibration Error:

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

How to define model-calibration

- Expected Calibration Error:

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

Computes weighted average of mis-calibration

- Maximum Calibration Error: Useful for high risk applications

$$MCE = \max_{m \in \{1,\ldots,M\}} \left| acc(B_m) - conf(B_m) \right|$$

Computes maximum mis-calibration

Reliability Diagram

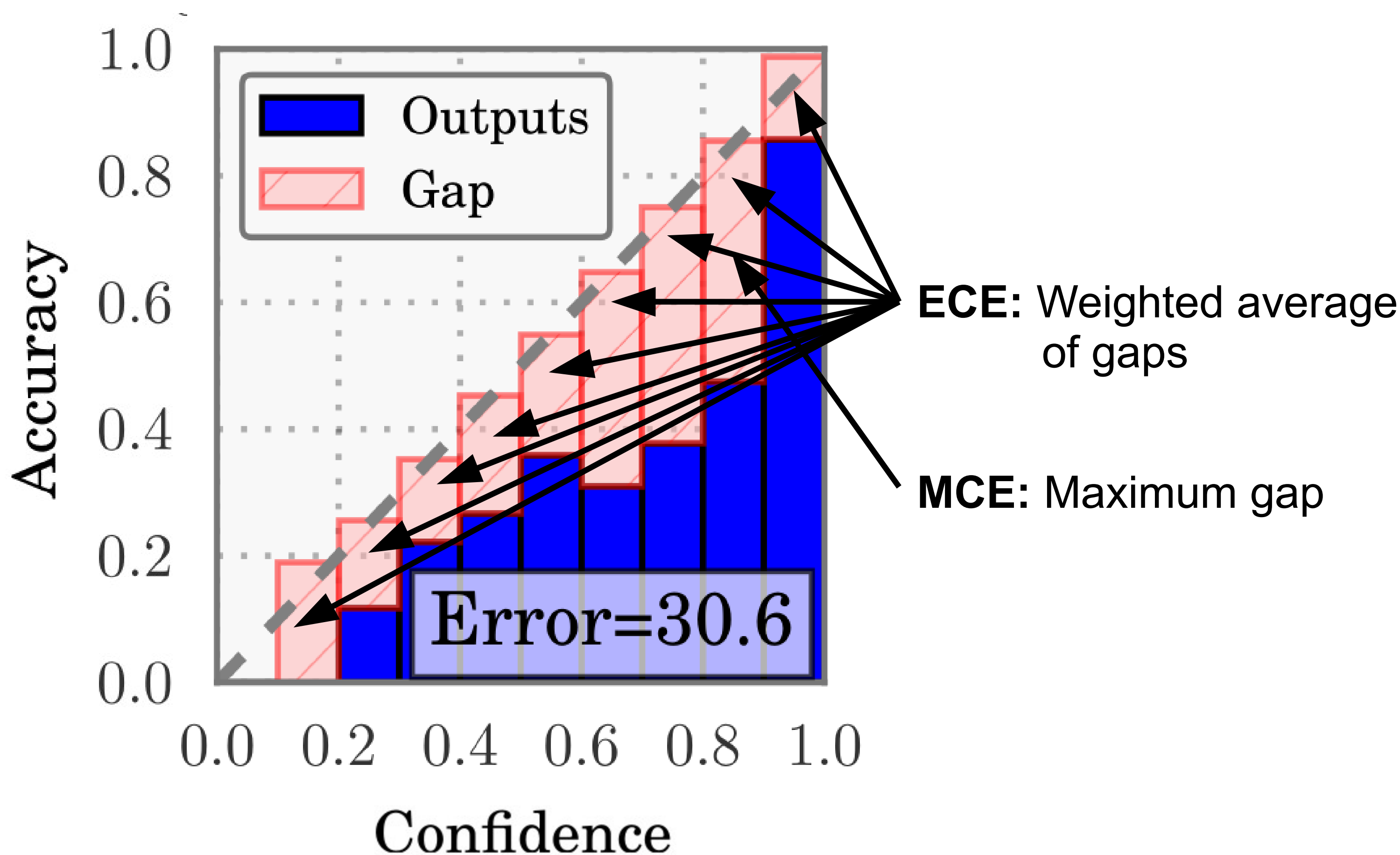


Image source: Guo, Pleiss et al. “On calibration...”, 2017

Problem

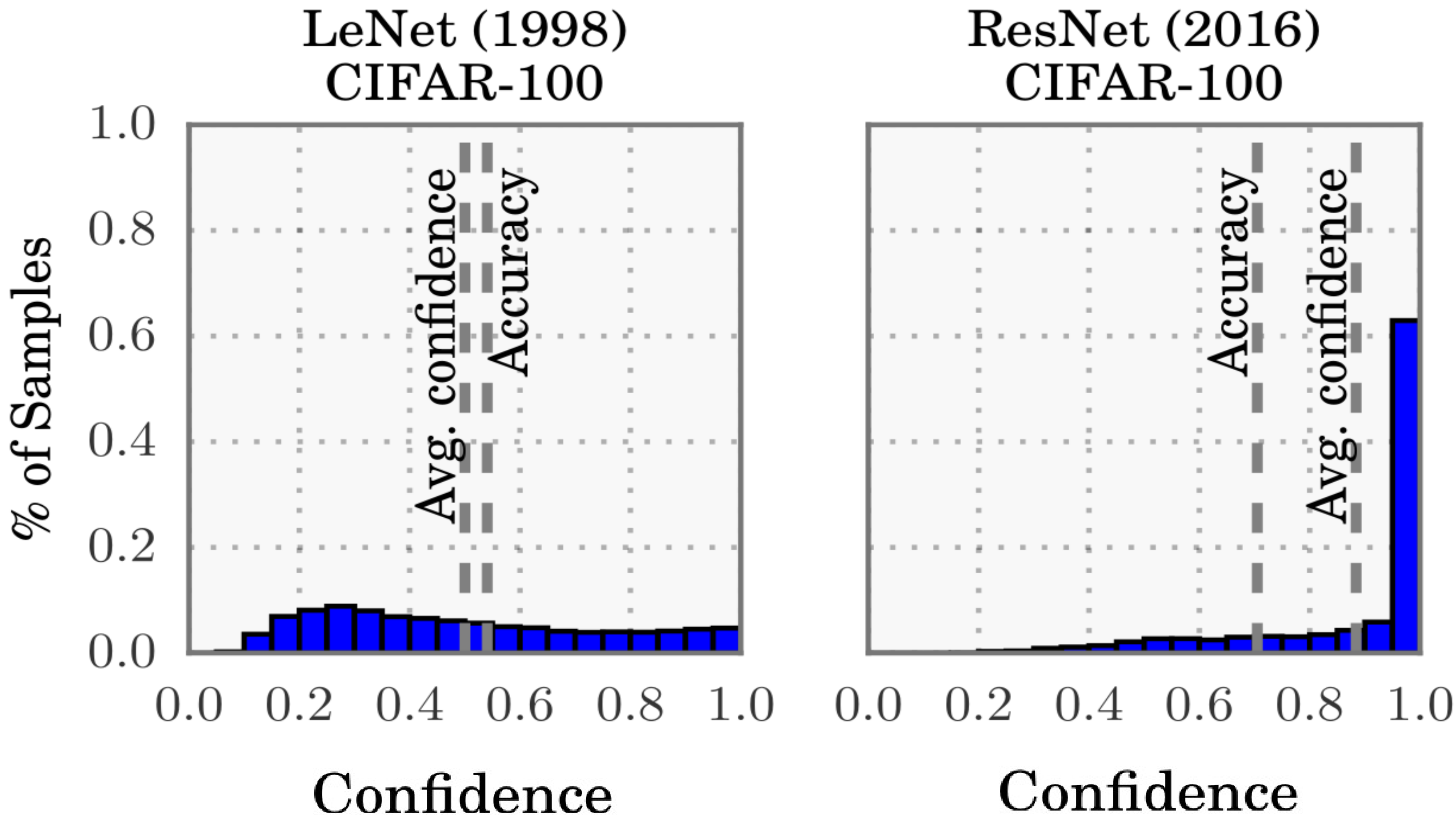


Image source: Guo, Pleiss et al. “On calibration...”, 2017

Problem

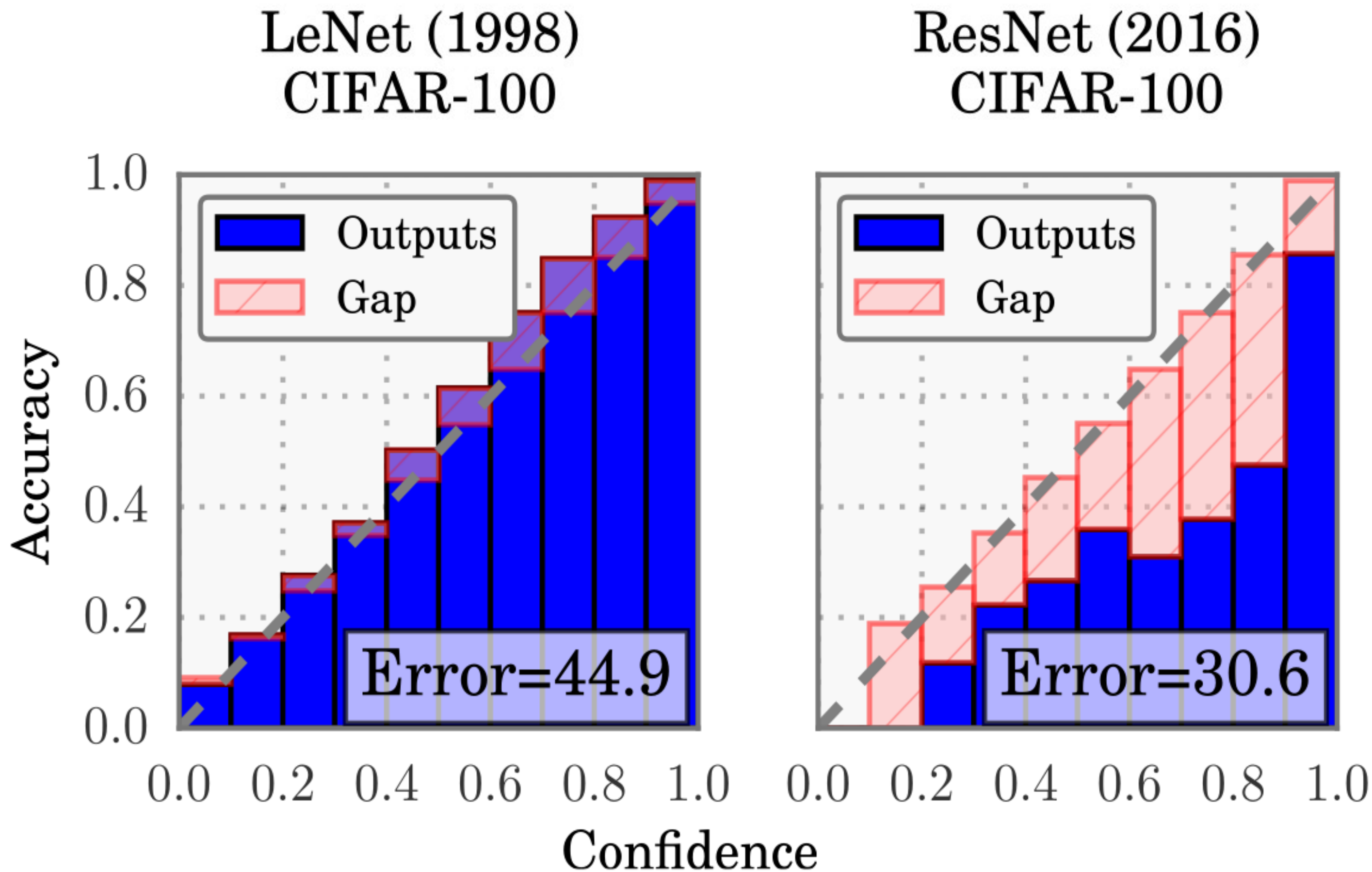


Image source: Guo, Pleiss et al. “On calibration...”, 2017

Goal

1) Understand why neural networks have become miscalibrated

2) Identify and compare methods to alleviate this problem

Overview

- **Introduction:** It's important for neural networks to be well-calibrated.

- **Definition:** How to measure model calibration?
  - ECE, MCE, Reliability diagrams

- **Problem:** Modern neural networks are no longer calibrated!

- **Analysis:** Which factors might influence model calibration?

- **Mitigation:** How to calibrate neural networks?

- **Experiments:** Which calibration methods perform best?

Factors influencing calibration

1. Model capacity: Depth & Width of network

Factors influencing calibration

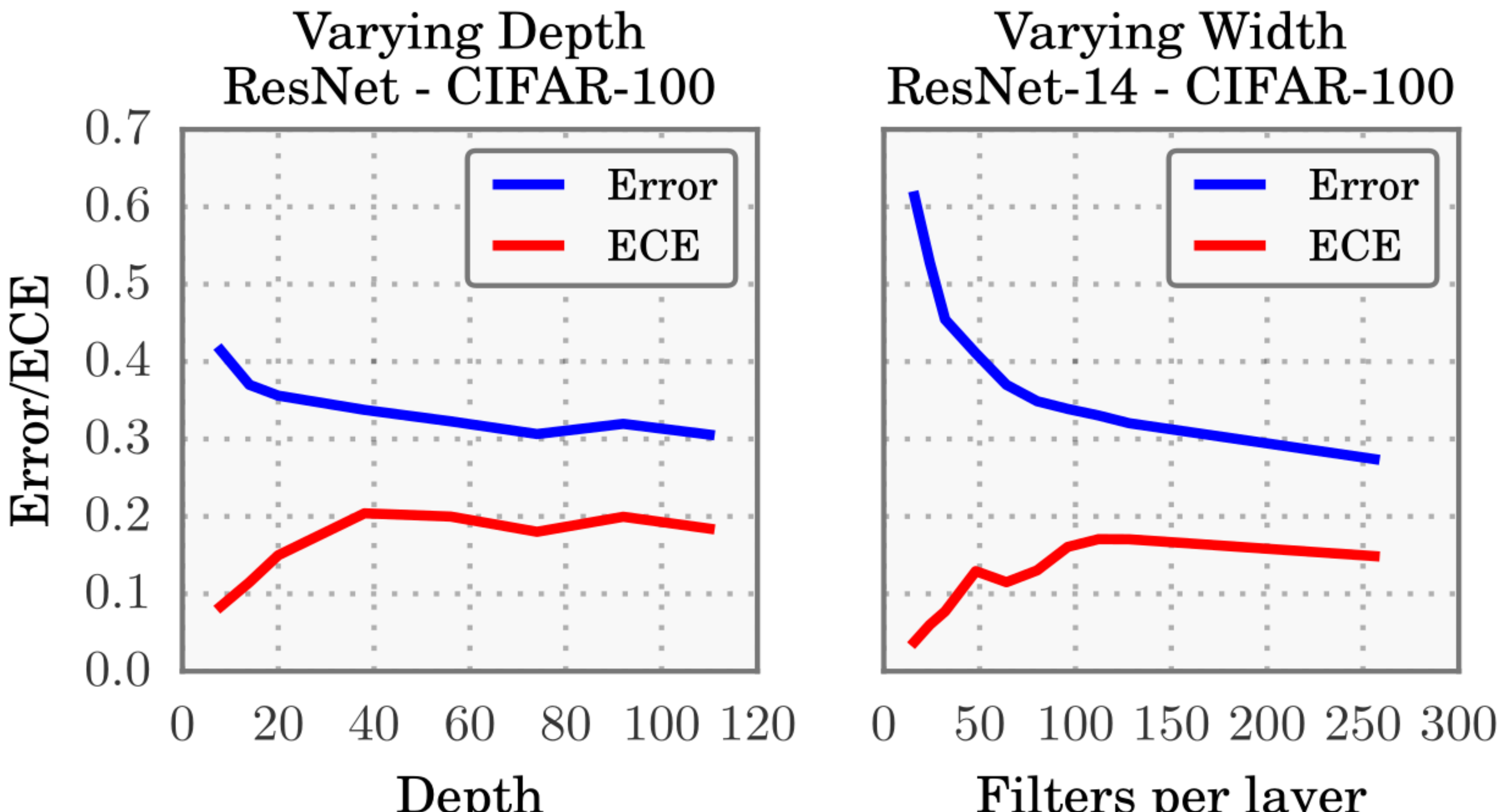


Image source: Guo, Pleiss et al. “On calibration...”, 2017

Factors influencing calibration

1. Model capacity: Depth & Width of network
2. Batch normalization
3. Weight decay

Factors influencing calibration

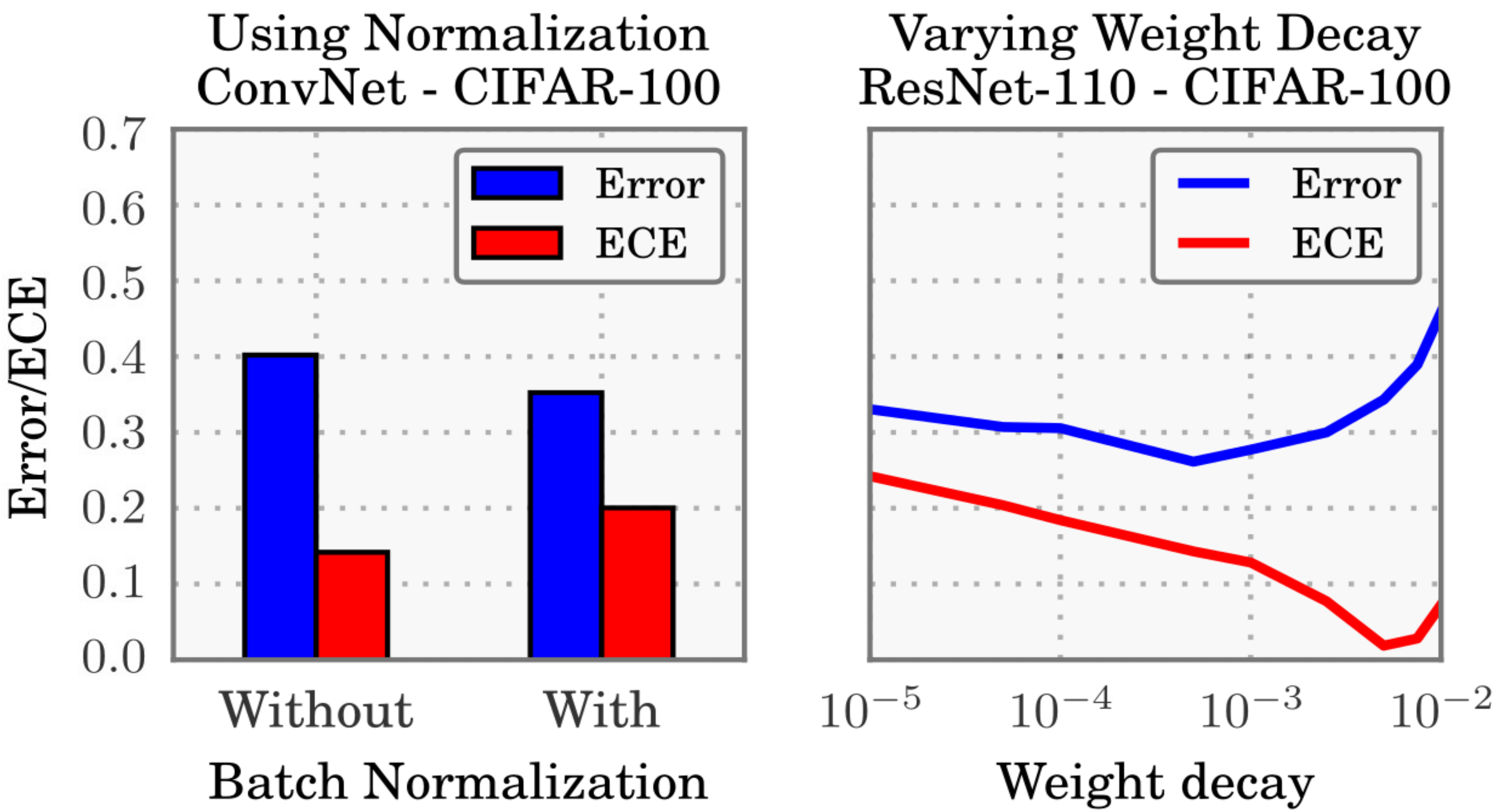


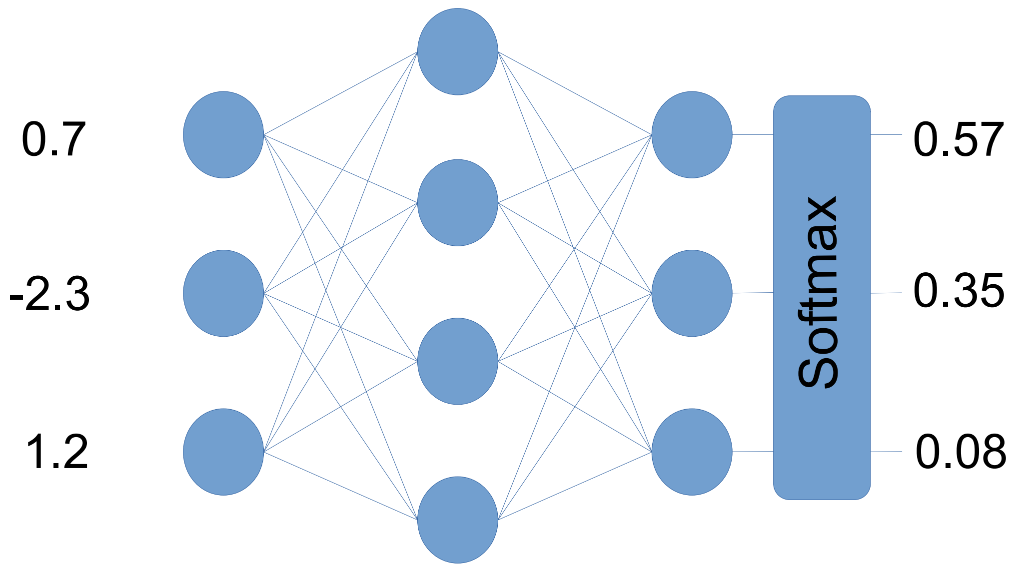
Image source: Guo, Pleiss et al. "On calibration...", 2017

Factors influencing calibration

1. Model capacity: Depth & Width of network
2. Batch normalization
3. Weight decay
4. Training using negative log-likelihood / cross-entropy loss

Factors influencing calibration

Training using negative log-likelihood / cross-entropy loss



$\hat{\pi}_\theta(\text{class } 1 \mid x)$

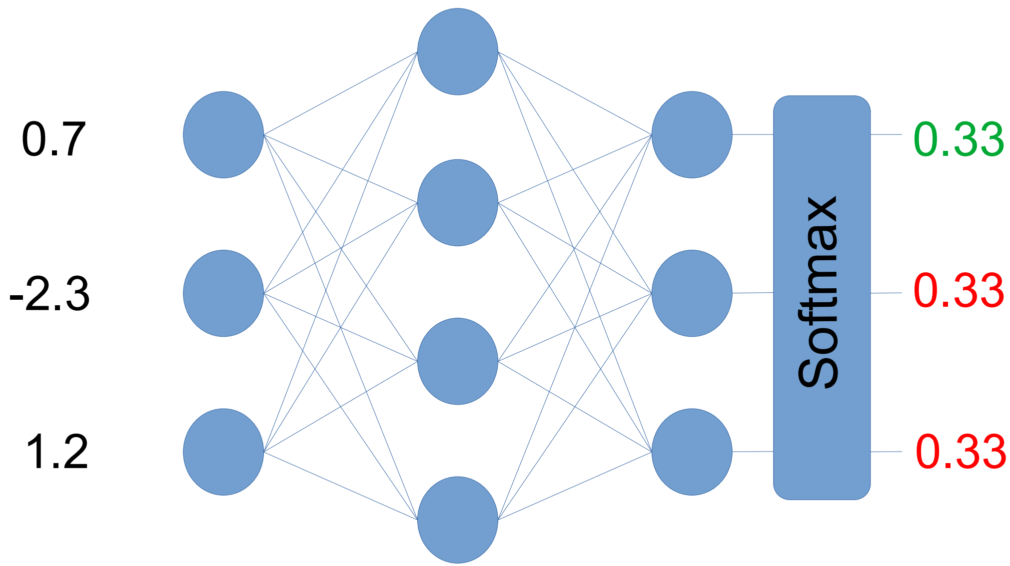
$\hat{\pi}_\theta(\text{class } 2 \mid x)$

$\hat{\pi}_\theta(\text{class } 3 \mid x)$

$$NLL(\theta) \;=\; \arg\max_\theta \prod_{i=1}^{n} \hat{\pi}_\theta(y_i \mid x_i) \;=\; \arg\min_\theta -\sum_{i=1}^{n} \log(\hat{\pi}_\theta(y_i \mid x_i))$$

Factors influencing calibration

Training using negative log-likelihood / cross-entropy loss

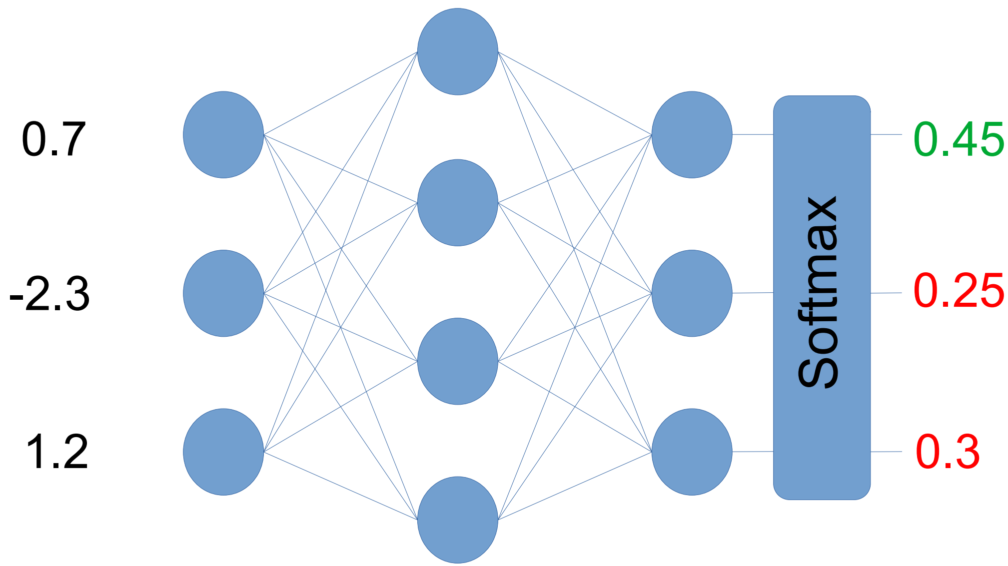


<u>How to minimize NLL:</u>

$$NLL(\theta) = \arg\max_{\theta} \prod_{i=1}^{n} \hat{\pi}_\theta(y_i \mid x_i) = \arg\min_{\theta} - \sum_{i=1}^{n} \log(\hat{\pi}_\theta(y_i \mid x_i))$$

Factors influencing calibration

Training using negative log-likelihood / cross-entropy loss



<u>How to minimize NLL:</u>

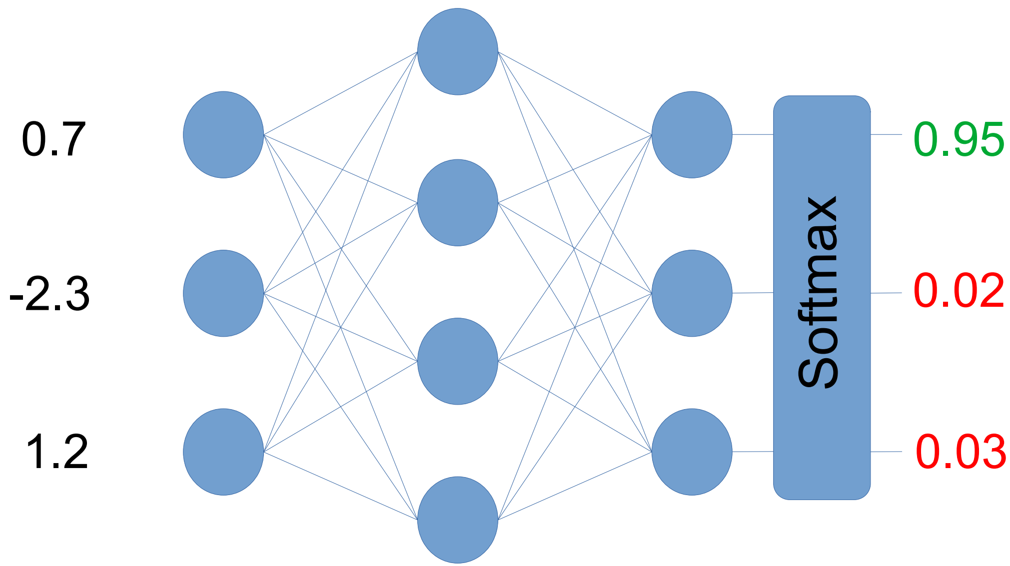
1) Predict the correct classes:

$$\hat{\pi}_\theta(y_i \mid x_i) \;\geq\; \hat{\pi}_\theta(y' \mid x_i) \quad \forall y' \in \mathcal{Y}$$

$$NLL(\theta) \;=\; \arg\max_\theta \prod_{i=1}^{n} \hat{\pi}_\theta(y_i \mid x_i) \;=\; \arg\min_\theta - \sum_{i=1}^{n} \log(\hat{\pi}_\theta(y_i \mid x_i))$$

Factors influencing calibration

Training using negative log-likelihood / cross-entropy loss



<u>How to minimize NLL:</u>

1) Predict the correct classes:

$$\hat{\pi}_\theta(y_i \mid x_i) \;\geq\; \hat{\pi}_\theta(y' \mid x_i) \quad \forall y' \in \mathcal{Y}$$

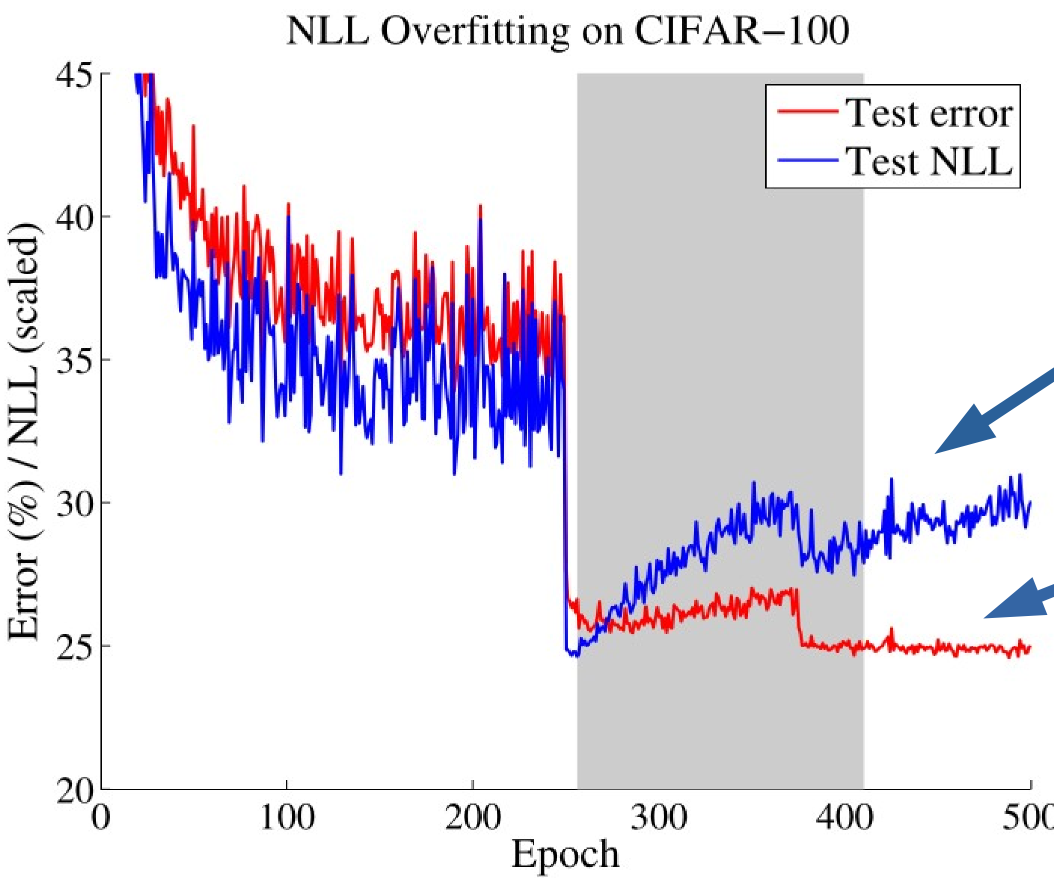
2) Increase confidence in correct classes!

$$NLL(\theta) \;=\; \arg\max_\theta \prod_{i=1}^{n} \hat{\pi}_\theta(y_i \mid x_i) \;=\; \text{ar}$$

Overfitting to NLL!

Factors influencing calibration

Training using negative log-likelihood / cross-entropy loss



Overfitting to NLL

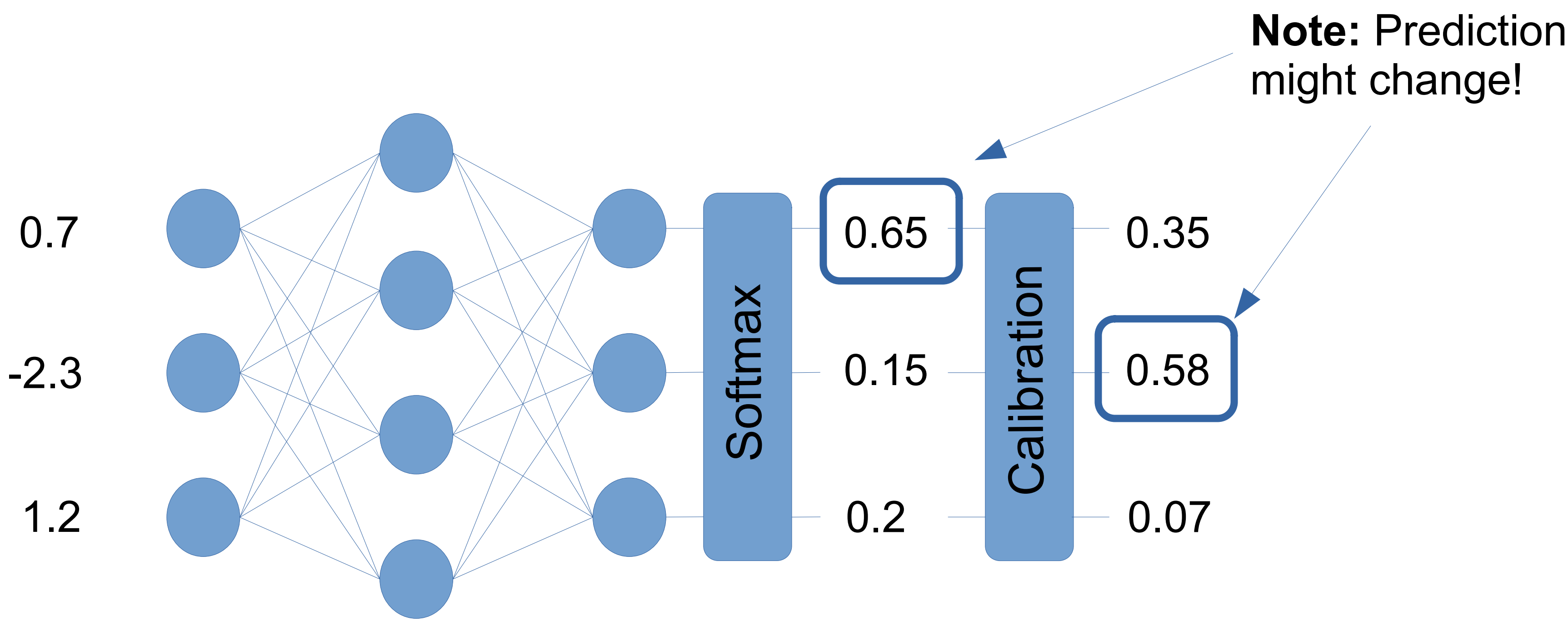
No overfitting to accuracy!

Image source: Guo, Pleiss et al. “On calibration...”, 2017

Overview

- **Introduction:** It’s important for neural networks to be well-calibrated.
- **Definition:** How to measure model calibration?
  - ECE, MCE, Reliability diagrams
- **Problem:** Modern neural networks are no longer calibrated!
- **Analysis:** Which factors might influence model calibration?
  - Model capacity, Normalization, Regularization, NLL
- **Mitigation:** How to calibrate neural networks?
- **Experiments:** Which calibration methods perform best?

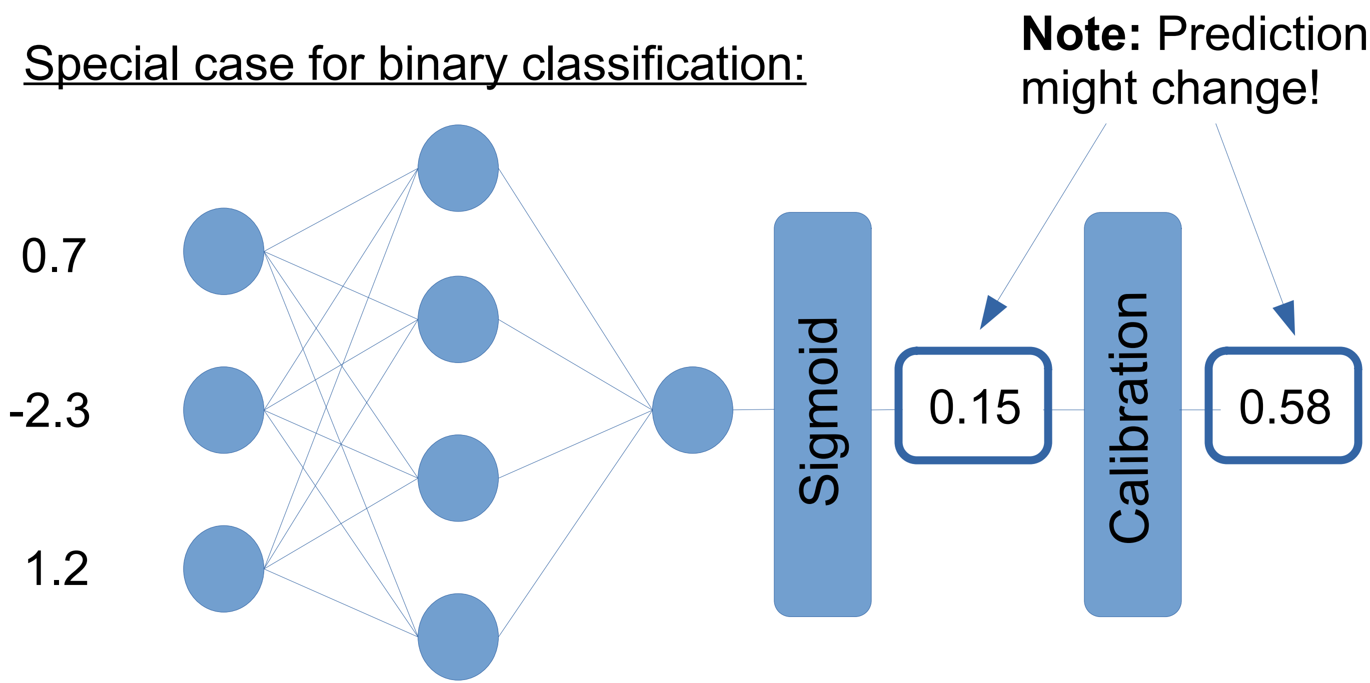
Calibration of neural networks



Input: $\mathbf{x}$ → Output: $\mathbf{z} = NN_\Theta(\mathbf{x})$ → Confidence: $\hat{\mathbf{p}} = softmax(\mathbf{z})$ → Calibrated Confidence: $\hat{\mathbf{q}} = calibration(\hat{\mathbf{p}})$

Calibration of neural networks

Special case for binary classification:



Input: $\mathbf{x}$ → Output: $\mathbf{z} = NN_{\Theta}(\mathbf{x})$ → Confidence: $\hat{\mathbf{p}} = \sigma(\mathbf{z})$ → Calibrated Confidence: $\hat{\mathbf{q}} = calibration(\hat{\mathbf{p}})$

Histogram Binning [Zadrozny et al. ICML 2001]

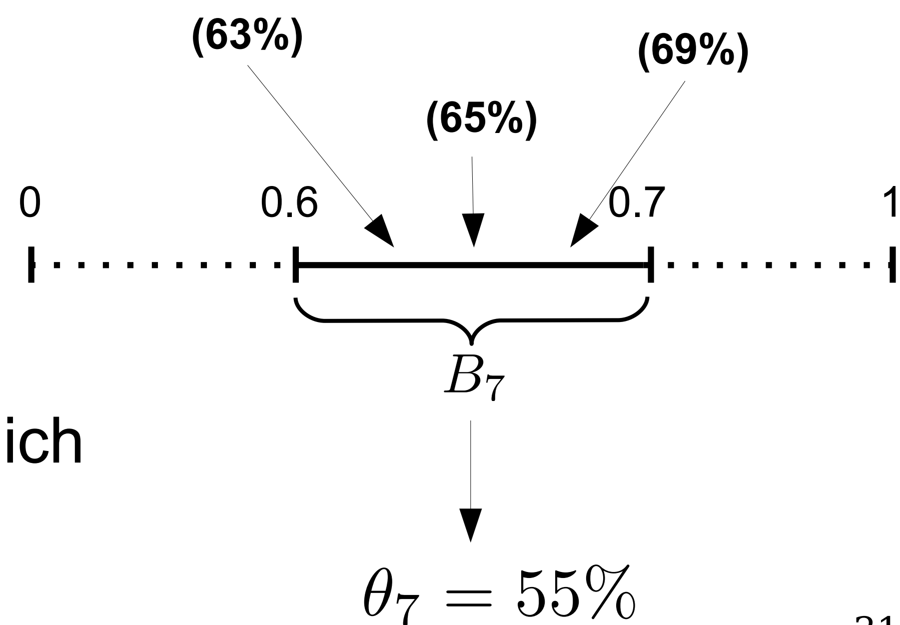
1. Group the predictions into M bins. Define bin B_m to be the set of all predictions $(\hat{y}_i, \hat{p}_i)$ for which it holds that:

$$\hat{p}_i \in \left(\frac{m-1}{M}, \frac{m}{M} \right]$$

2. For all predictions in bin B_m output the probability θ_m

3. For each bin B_m find θ_m which minimizes

$$\sum_{y_i : \hat{p}_i \in B_m} (y_i - \theta_m)^2$$

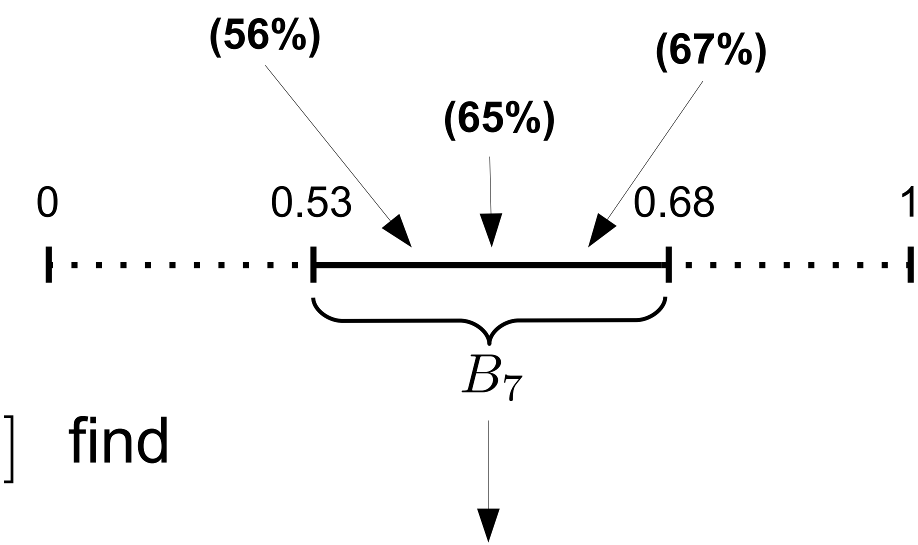


Isotonic Regression [Zadrozny et al. KDD 2002]

1. Group the predictions into M bins. Define bin B_m to be the set of all predictions $(\hat{y}_i, \hat{p}_i)$ for which it holds that:

$$\hat{p}_i \in \left(\frac{m-1}{M}, \frac{m}{M} \right]$$

2. For all predictions in bin B_m output the probability θ_m



3. For each bin $B_m = (a_m, a_{m+1}]$ find $\boxed{(\theta_m, a_m, a_{m+1})}$ which minimize

$$\sum_{y_i : \hat{p}_i \in B_m} (y_i - \theta_m)^2$$

$$\theta_7 = 55\%$$

Bayesian Binning into Quantiles (BBQ)

[Naeini et al. AAAI 2015]

- Look at all possible binning schemes at the same time!
- For a given validation set $\mathcal{D}$ let $\mathcal{S}$ be the set of all possible binning schemes for this data set.
- Previous models: Fix one binning scheme $s \in \mathcal{S}$ and compute opt… θ for each … data sample … predict $\hat{q}_i =$ …
- Bayesian Binning into Quantiles:

Prediction $\hat{q}_i$ under the model s

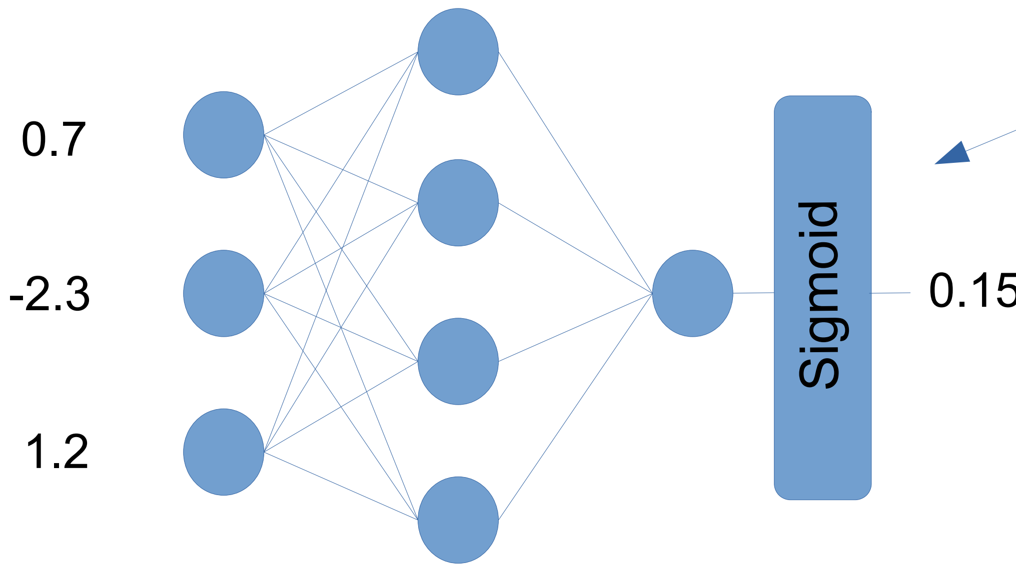
Sum over all binning schemes

How probable the model s is given the data $\mathcal{D}$

$$\hat{q}_i = \mathbb{P}(y_i = 1 \mid \hat{p}_i) = \sum_{s \in \mathcal{S}} \mathbb{P}(y_i = 1 \mid \hat{p}_i, S = s, \mathcal{D}) \cdot \mathbb{P}(S = s \mid \mathcal{D})$$

Platt scaling

[Platt et al. Advances in large margin classifiers 1999]

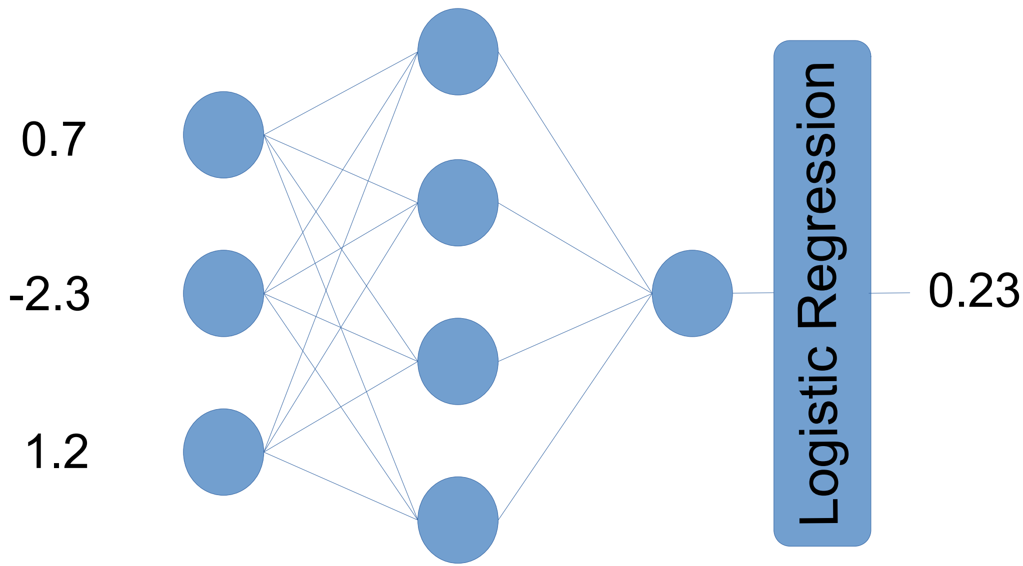


Parametrize sigmoid to adapt to mis-calibration!

Input:	Output:	Confidence:
$\mathbf{x}$	$\mathbf{z} = NN_{\Theta}(\mathbf{x})$	$\hat{\mathbf{p}} = \sigma(\mathbf{z})$

Platt scaling

[Platt et al. Advances in large margin classifiers 1999]

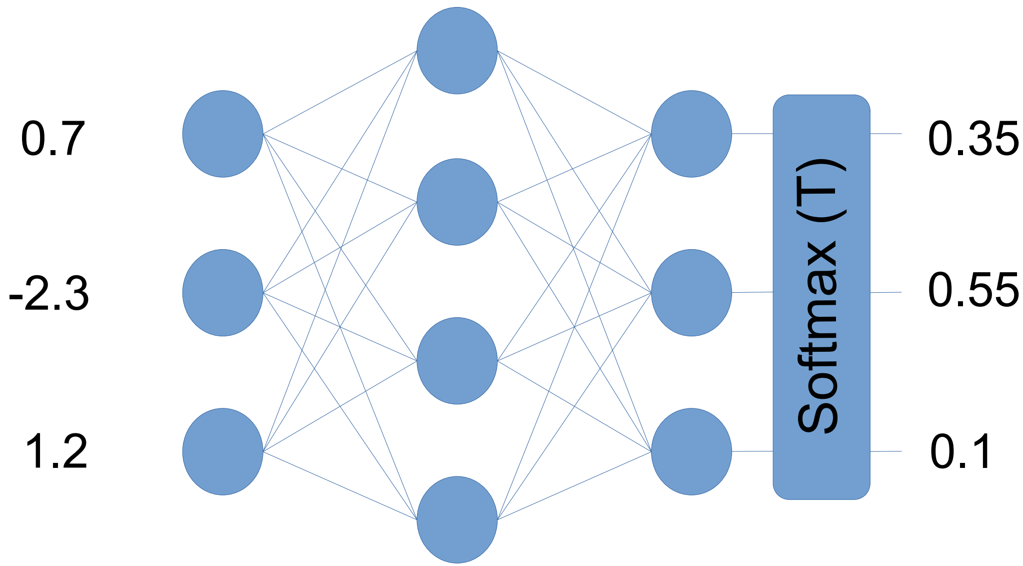


Input: $\mathbf{x}$

Output: $\mathbf{z} = NN_{\Theta}(\mathbf{x})$

Confidence: $\hat{\mathbf{q}} = \sigma(a \cdot \mathbf{z} + b) \quad a, b \in \mathbb{R}$

Temperature scaling



Temperature scaling

- Number of parameters is constant!
- This method doesn't change the predictions! => Accuracy stays the same
- Very easy to implement
- Fast to compute

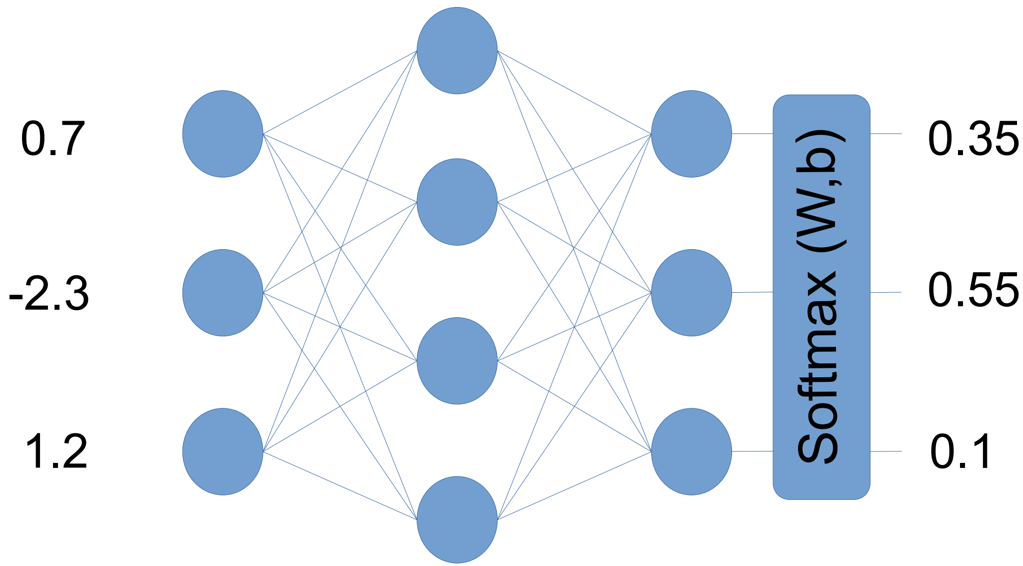
Input: $\mathbf{x}$

Output: $\mathbf{z} = NN_{\Theta}(\mathbf{x})$

Confidence: $\hat{\mathbf{q}} = softmax(\mathbf{z}/T)$

$T \in \mathbb{R}$

Matrix and Vector scaling



Matrix scaling

- No restrictions on W
- Number of parameters grows quadratically!

Vector scaling

- Restrict W to be a diagonal matrix
- Number of parameters grows linearly

Input: $\mathbf{x}$ → Output: $\mathbf{z} = NN_{\Theta}(\mathbf{x})$ → Confidence: $\hat{\mathbf{q}} = softmax(\mathbf{W} \cdot \mathbf{z} + \mathbf{b})$

$$\mathbf{W} \in \mathbb{R}^{k \times k}$$

$$\mathbf{b} \in \mathbb{R}^{k}$$

Overview

- **Introduction:** It's important for neural networks to be well-calibrated.
- **Definition:** How to measure model calibration?
  - ECE, MCE, Reliability diagrams
- **Problem:** Modern neural networks are no longer calibrated!
- **Analysis:** Which factors might influence model calibration?
  - Model capacity, Normalization, Regularization, NLL
- **Mitigation:** How to calibrate neural networks?
  - Binning, Platt- Matrix/Vector-, Temperature- scaling
- **Experiments:** Which calibration methods perform best?

Experiments: Results ECE

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%

Image source: Guo, Pleiss et al. “On calibration...”, 2017

Experiments: Results Error

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%**

Image source: Guo, Pleiss et al. “On calibration...”, 2017

Results

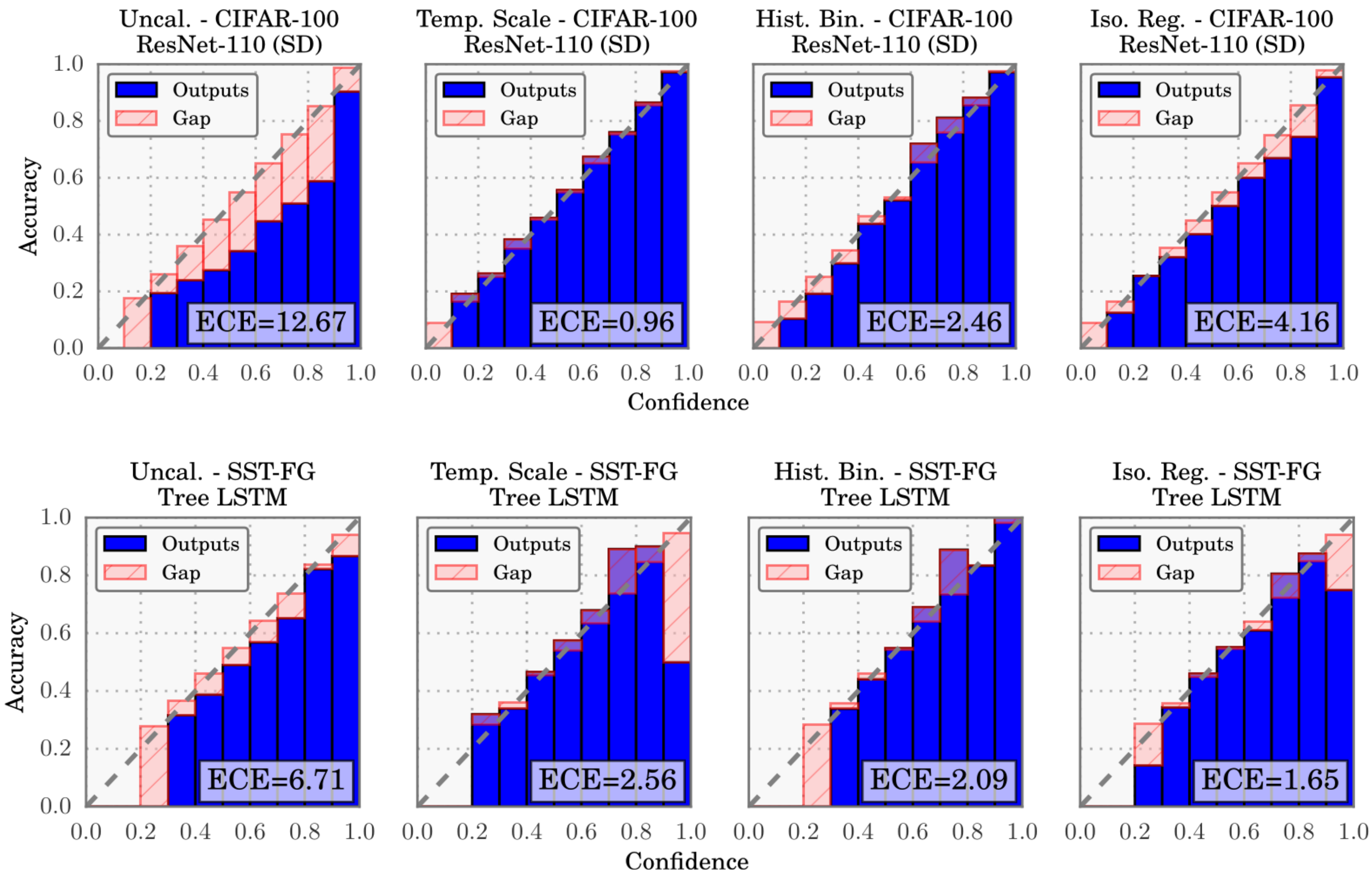


Image source: Guo, Pleiss et al. “On calibration...”, 2017

Overview

- **Introduction:** It's important for neural networks to be well-calibrated.

- **Definition:** How to measure model calibration?
  - ECE, MCE, Reliability diagrams

- **Problem:** Modern neural networks are no longer calibrated!

- **Analysis:** Which factors might influence model calibration?
  - Model capacity, Normalization, Regularization, NLL

- **Mitigation:** How to calibrate neural networks?
  - Binning, Platt- Matrix/Vector-, Temperature- scaling

- **Experiments:** Which calibration methods perform best?
  - Temperature scaling

My Take

- Interesting paper
- Well-written
- More data to show correlation between optimization techniques and ECE would have been appreciated

Takeaways

- **Fact:** Neural Networks are increasingly used in high risk decision making applications
- **Problem:** Modern neural networks are miscalibrated
- **Solution:** Performing Post-processing like for example temperature scaling to adjust confidence estimates helps to mitigate the problem

Appendix

Experiments: Datasets

Table	Description	# of classes	Train/Validation/Test
Caltech-UCSD	Bird images	200	5,994 / 2,897 / 2,897
Stanford Cars	Car images	196	8,041 / 4,020 / 4,020
ImageNet 2012	Natural scene images	1000	1.3M /25,000 / 25,000
CIFAR-10/CIFAR-100	Color images	10 / 100	45,000 / 5,000 / 10,000
Street View House Numbers (SVHN)	House number images	10	598,388 / 6,000 / 26,032
20 News	News articles	20	9,034 / 2,259 / 7,528
Reuters	News articles	8	4,388 / 1,097 / 2,189
Stanford Sentiment Treebank	Movie reviews	2 / 5	6,920 / 872 / 1,821 544 / 1,101 / 2,210

Experiments: Networks

- Image classification tasks:
  - ResNets [He et al. CVPR 2016]
  - ResNets with stochastic depth [Huang et al. ECCV 2016]
  - Wide ResNets [Zagoruyko et al. BMVC 2016]
  - DenseNets [Huang et al. CVPR 2017]
- Document classification tasks:
  - Deep Averaging Networks [Iyyer et al. ACL 2015]
  - TreeLSTMs [Tai et al. ACL 2015]

Experiments: Results MCE

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%

Image source: Guo, Pleiss et al. “On calibration...”, 2017

Paper Impact

On calibration of modern neural networks

C Guo, G Pleiss, Y Sun… - … Conference on Machine …, 2017 - proceedings.mlr.press

Confidence calibration–the problem of predicting probability estimates representative of the true correctness likelihood–is important for classification models in many applications. We discover that modern neural networks, unlike those from a decade ago, are poorly calibrated. Through extensive experiments, we observe that depth, width, weight decay, and Batch Normalization are important factors influencing calibration. We evaluate the performance of various post-processing calibration methods on state-of-the-art architectures …

☆ 99 $\boxed{\text{Cited by 1220}}$ Related articles All 7 versions »

Paper impact

- **Confidence of out-of-distribution samples:**
  - Enhancing the reliability of out-of-distribution image detection in NNs: https://arxiv.org/pdf/1706.02690.pdf
  - Training Confidence-calibrated classifiers for detecting out-of-distribution samples: https://arxiv.org/pdf/1711.09325.pdf
  - Learning Confidence for Out-of-Distribution Detection in Neural Networks: https://arxiv.org/pdf/1802.04865.pdf
  - Deep anomaly detection with outlier exposure: https://arxiv.org/pdf/1812.04606.pdf
  - Why ReLU networks yield high-confidence predictions far away fromthe training data and how to mitigate the problem: https://openaccess.thecvf.com/content_CVPR_2019/papers/Hein_Why_ReLU_Networks_Yield_High-Confidence_Predictions_Far_Away_From_the_CVPR_2019_paper.pdf

- **Application of paper:**
  - A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6722431/
  - Deep k-Nearest Neighbors: Towards Confident,Interpretable and Robust Deep Learning: https://arxiv.org/pdf/1803.04765.pdf?fbclid=IwAR2D5gqQf9SL0xRWBctEVrUCL9uUiIf9lZrpPN83YZYbiCGdLAlMlhhaVns
- **Comparison and Critique:**
  - Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift: https://arxiv.org/pdf/1906.02530.pdf
  - Measuring calibration in deep learning: https://openaccess.thecvf.com/content_CVPRW_2019/papers/Uncertainty%20and%20Robustness%20in%20Deep%20Visual%20Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.pdf
- **Calibration and fairness:**
  - On fairness and calibration: https://arxiv.org/pdf/1709.02012.pdf