University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Calibration of Deep Learning Models

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

Multi-Class Classification

- Multi-Class Classification

$$h(X) = (\hat{Y}, \hat{P})$$

Multi-Class Classification

- Multi-Class Classification

$$h(X) = (\hat{Y}, \hat{P})$$

- Perfect Calibration

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

Multi-Class Classification

- Multi-Class Classification

$$h(X) = (\hat{Y}, \hat{P})$$

- Perfect Calibration

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

- Estimate the Expected Accuracy:
  - Group predictions into *M* interval bins
  - Each of the bins are of size 1/M
  - Calculate the accuracy of each bin

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

Multi-Class Classification

- Multi-Class Classification

$$h(X) = (\hat{Y}, \hat{P})$$

- Perfect Calibration

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

- Estimate the Expected Accuracy:
  - Group predictions into *M* interval bins
  - Each of the bins are of size 1/M
  - Calculate the accuracy of each bin

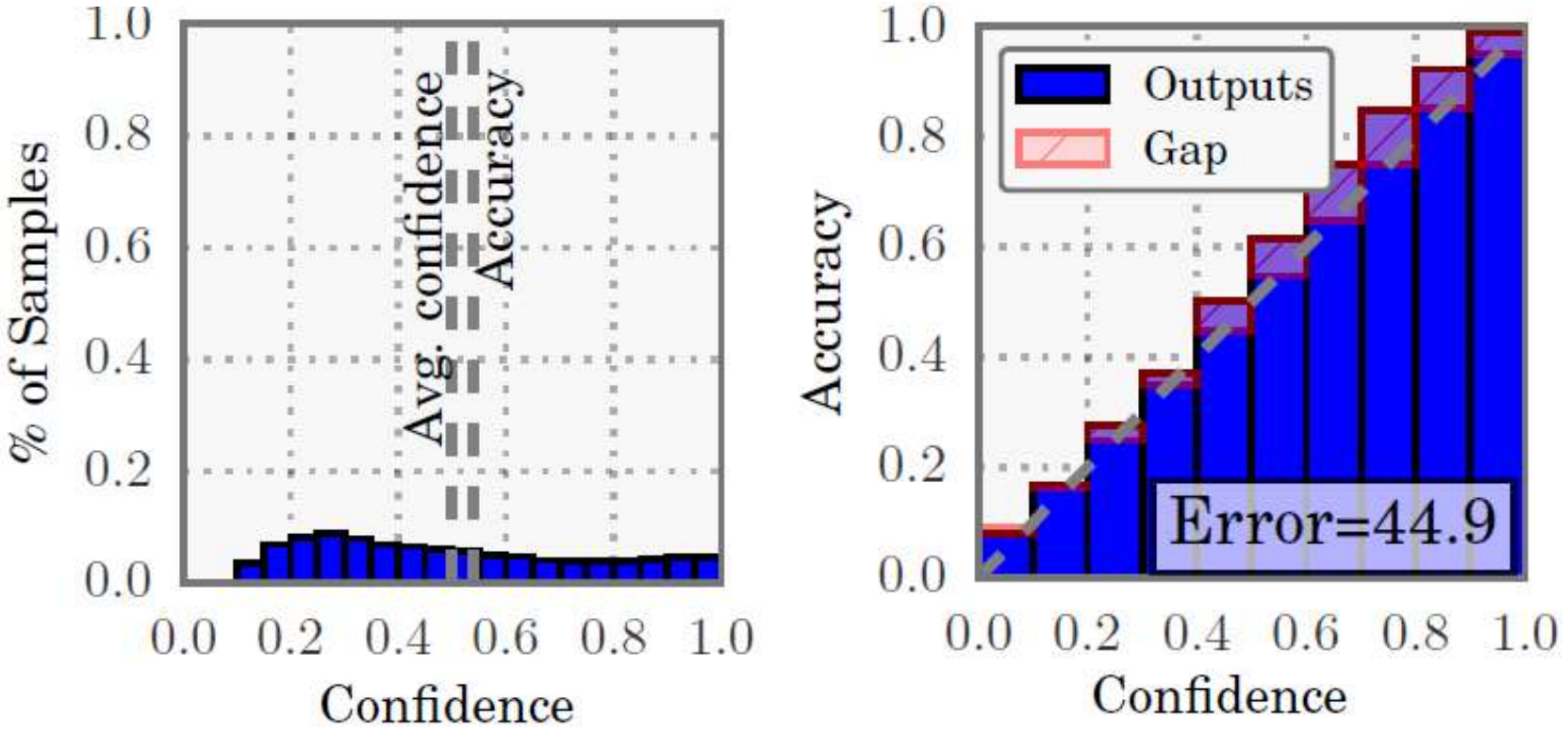
$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$$

- Estimate average confidence within each bin:

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$

Reliability in LeNet (CIFAR-100)

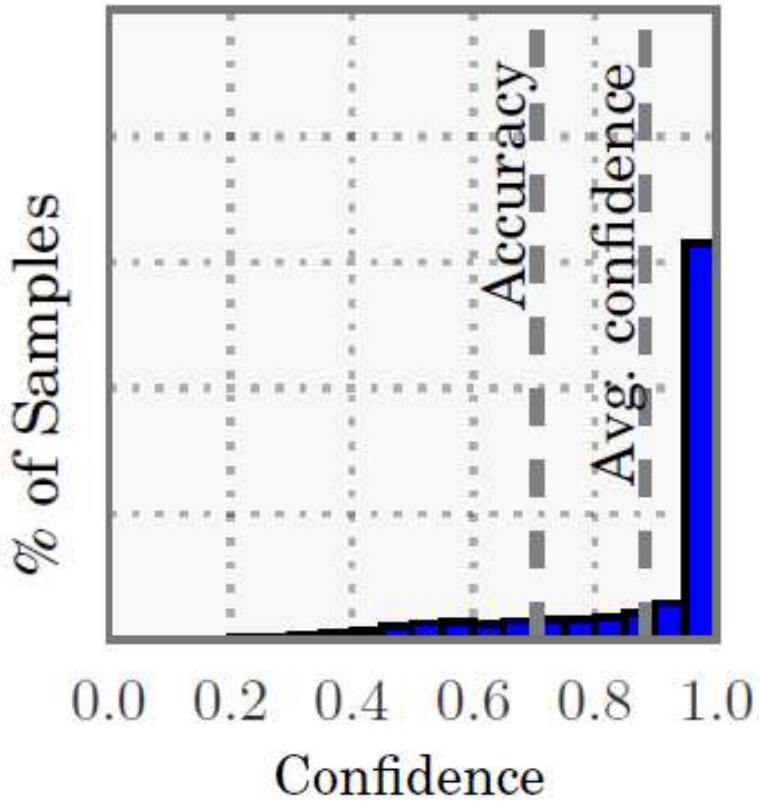
Well Calibrated

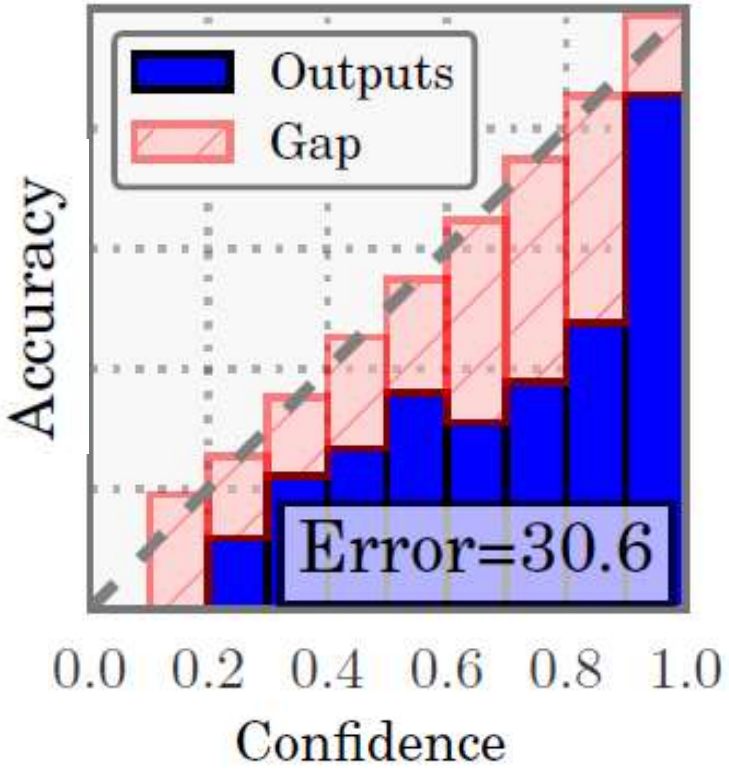


Guo et al. 'On Calibration of Modern Neural Networks', 2017

Reliability in ResNet (CIFAR-100)

Poorly Calibrated





Guo et al. 'On Calibration of Modern Neural Networks', 2017

Challenges in DNNs Calibration

- Why DNNs are more mis-calibrated?
- Which architectures tend to cause more mis-calibration?
- How to post-process DNNs models to improve calibration?

Expected Calibration Error

- Difference in Expectation between confidence and accuracy

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

- Partition predictions into M equally-spaced bins
- Estimate the Expected Calibration Error as a weighted average

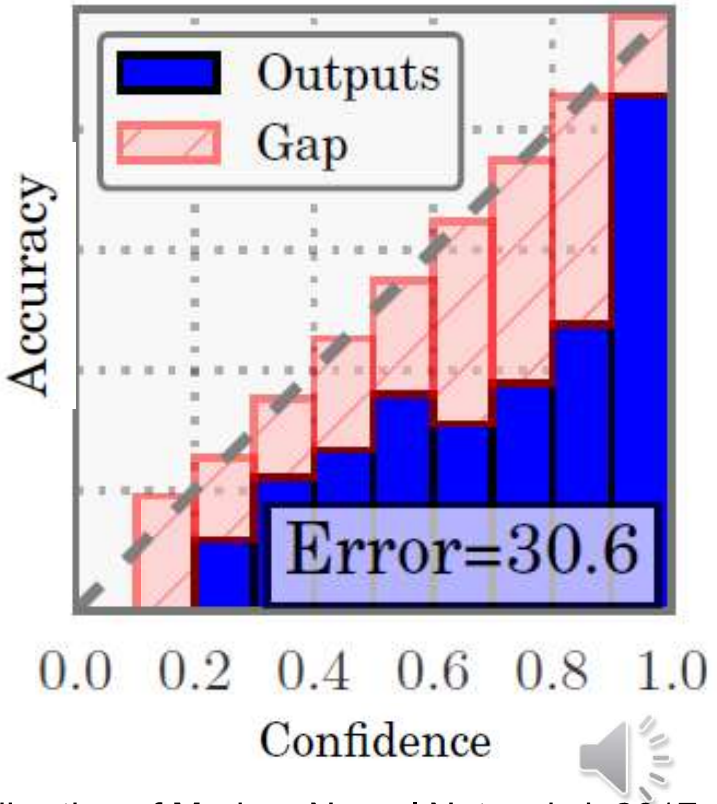
$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

Maximum Calibration Error

- Worst-case deviation between confidence and accuracy

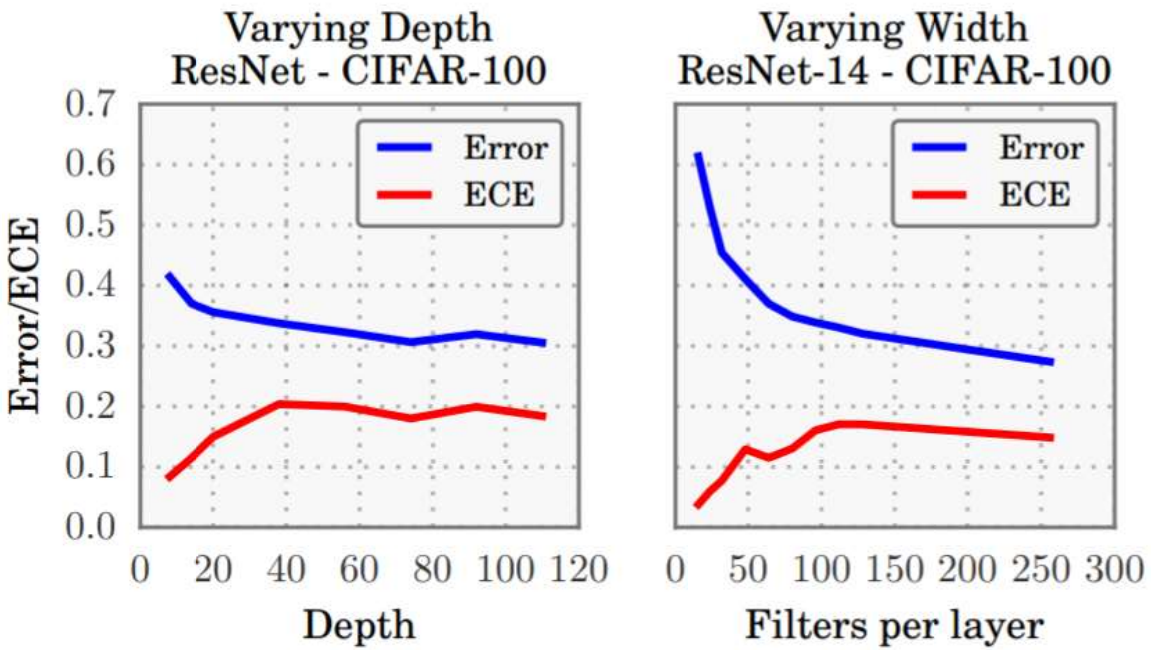
$$MCE = max_{m \in \{1,\cdots,M\}} |acc(B_m) - conf(B_m)|$$

- MCE is the largest calibration gap across all bins on reliability diagrams



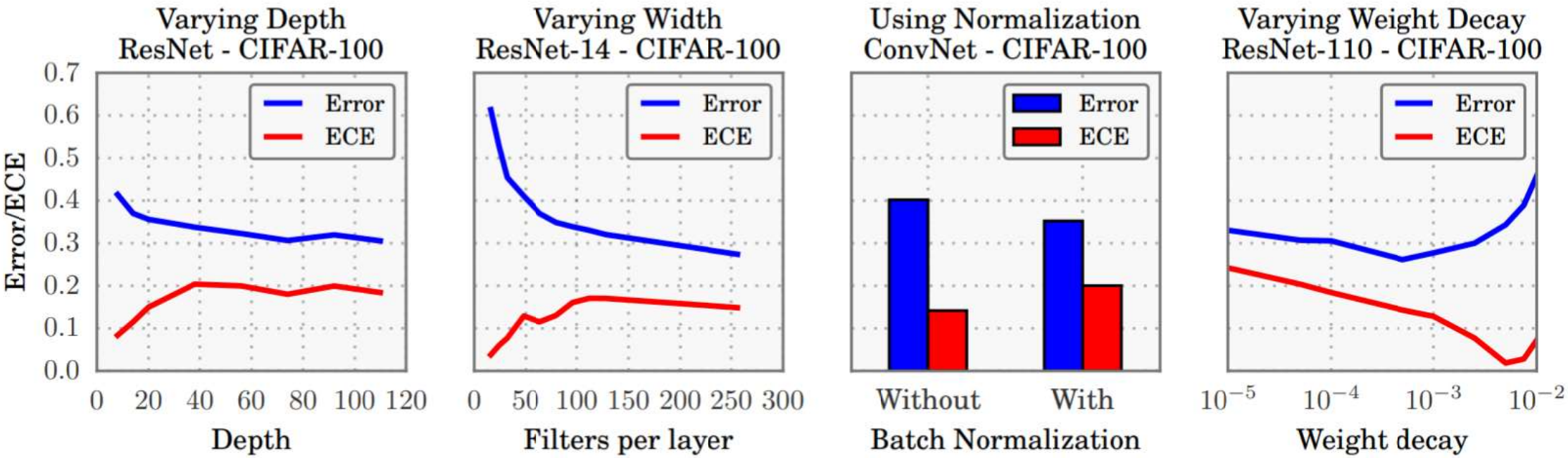
Guo et al. 'On Calibration of Modern Neural Networks', 2017

Mis-Calibration DNN architectures



Guo et al. 'On Calibration of Modern Neural Networks', 2017

Mis-Calibration DNN architectures



Guo et al. 'On Calibration of Modern Neural Networks', 2017

Criticism of Expected Calibration Error

- Ambiguity in how to implement the binning procedure
  - Trade of between bias and variance
  - Fixed calibration ranges
  - Static binning schemes
- Ambiguity in how to implement the measure for multi-class paradigms
  - Calibration error is not uniform across classes

Summary

- Calibration measures the reliability of confidence
- In healthcare and critical application calibration is important
- Deep Neural Network appear to have low calibration scores, which reflect overfitting, even when their binary classification scores are generalizable

References

- Guo et al. ‘On Calibration of Modern Neural Networks’, International Conference on Machine Learning, 2017.
- Nixon et al. ‘Measuring Calibration in Deep Learning’, CVPR, 2019.