FROM BLACK-BOX TO WHITE-BOX: EXAMINING CONFIDENCE CALIBRATION UNDER DIFFERENT CONDITIONS

FRANZISKA SCHWAIGER, MAXIMILIAN HENNE, FABIAN KUEPPERS, FELIPPE SCHMOELLER ROZA, KARSTEN ROSCHER, ANSELM HASELHOFF
THE PERCEPTION CHALLENGE

Source: Towards a Framework to Manage Perceptual Uncertainty for Safe Automated Driving, Krzysztof Czarnecki and Rick Salay
Modern neural networks are overconfident [1]

**Measurement of Miscalibration:** Expected Calibration Error (ECE):

\[ ECE = \sum_{n=1}^{N} \frac{|I(n)|}{|D|} \cdot |\text{acc}(n) - \text{conf}(n)| \]

Fig. 1: Reliability Diagram of a 110-layer ResNet on CIFAR-100 [1]
OBJECT DETECTION PIPELINE

A Bochkovskiy et. al, Yolov4: Optimal speed and accuracy of object detection
Modern object detectors are also not well calibrated [2]

**Measurement of Miscalibration:**
Detection Expected Calibration Error (D-ECE):

\[
D - ECE_K = \sum_{n=1}^{N_{total}} \frac{|l(n)|}{|D|} \cdot |prec(n) - conf(n)|
\]
BACKGROUND: NON-MAXIMUM-SUPPRESSION (NMS)

Method: selects a single bounding box with the highest confidence score out of many overlapping bounding boxes

Intersection over Union (IoU): controls how aggressively overlapping boxes are discarded

\[ \text{IoU}(A, B) = \frac{A \cap B}{A \cup B} \]
Miscalibration of object detectors is measured before (white-box) and after (black-box) NMS.

Training and evaluation is done on the COCO2017 validation dataset.

**WHITE-BOX VS. BLACK-BOX**

TP: \( IoU(\square, \square) \geq threshold \)

FP: \( IoU(\square, \square) < threshold \)
WHITE BOX CALIBRATION RESULTS

RetinaNet

Before
Uncalibrated D-ECE: 22.9%

After
Calibrated D-ECE: 0.981%

Faster R-CNN

Before
Uncalibrated D-ECE: 4.20%

After
Calibrated D-ECE: 0.861%
QUALITATIVE RESULTS

NMS potentially degrades initially well-calibrated predictors

After white-box calibration

Confidence Histogram (top) and Reliability Diagram (bottom) of RetinaNet

Position-dependent Heatmap of RetinaNet

NMS
CONCLUSION

- Focal loss decreases average confidence
- White Box Calibration works
- NMS making predictions overconfident
- Detections on border of image tend to be worse calibrated than near the center
- Future work should focus on other box aggregation methods than NMS (e.g. average boxes)
All quantitative calibration results are given in our paper in more detail.

References


On Calibration of Modern Neural Networks.


Multivariate Confidence Calibration for Object Detection.