Carnegie Mellon University
Software Engineering Institute

WHAT IS YOUR METRIC TELLING YOU?
EVALUATING CLASSIFIER CALIBRATION UNDER CONTEXT-SPECIFIC DEFINITIONS OF RELIABILITY

John Kirchenbauer, Jacob Oaks, Eric Heim

ICLR 2022
Workshop on ML Evaluation Standards

Contemporary work on calibration has been towards the goal of learning classifiers such that their largest output (confidence in the "predicted class") is calibrated.

However, this narrow interpretation of classifier outputs does not adequately capture the variety of practical use cases in which classifiers can aid in decision making.

Expressive metrics must be developed that accurately measure calibration error for the specific context in which a classifier will be deployed.

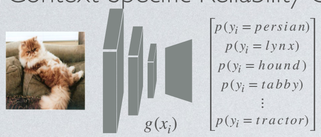
Generalized Expected Calibration Error (GECE) Lenses

SUMMARY FINDINGS

1) Definitions of ECE that focus solely on the predicted class fail to accurately measure calibration error under a selection of practically useful definitions of reliability

2) Many common calibration techniques fail to improve calibration performance uniformly across ECE metrics derived from these diverse definitions of reliability.

Lenses: Context-Specific Reliability Conditions



"Full ECE"

$$\begin{bmatrix} \vdots \\ g(X)_j \\ \vdots \end{bmatrix} \leftarrow \Delta \rightarrow \begin{bmatrix} \vdots \\ \mathbb{P}\left(Y = c_j \,|\, g(X)_j\right) \\ \vdots \end{bmatrix}$$

"Traditional/Top-1 ECE"

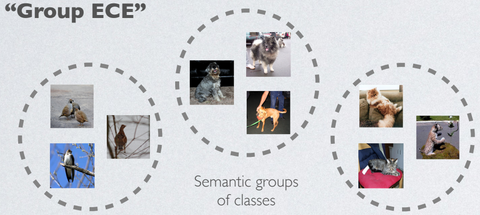
$$\left[g(X)_{max}\right] \leftarrow \Delta \rightarrow \left[\mathbb{P}\left(Y = c_{max} \,|\, g(X)_{max}\right)\right]$$

"Top-k ECE"

sort then select k most confident outputs

$$\begin{bmatrix} g(X)_1 \\ g(X)_2 \\ g(X)_3 \\ \vdots \end{bmatrix} \leftarrow \Delta \rightarrow \begin{bmatrix} \mathbb{P}\left(Y = c_1 \,|\, g(X)_1\right) \\ \mathbb{P}\left(Y = c_2 \,|\, g(X)_2\right) \\ \mathbb{P}\left(Y = c_3 \,|\, g(X)_3\right) \\ \vdots \end{bmatrix}$$

"Group ECE"



Semantic groups of classes

$$\begin{bmatrix} g(x_i)_{bird} \\ g(x_i)_{dog} \\ g(x_i)_{cat} \end{bmatrix} = \begin{bmatrix} p(y_i = finch) + p(y_i = grouse) + p(y_i = quail) \\ p(y_i = hound) + p(y_i = schnauzer) + p(y_i = keeshond) \\ p(y_i = persian) + p(y_i = lynx) + p(y_i = tabby) \end{bmatrix}$$

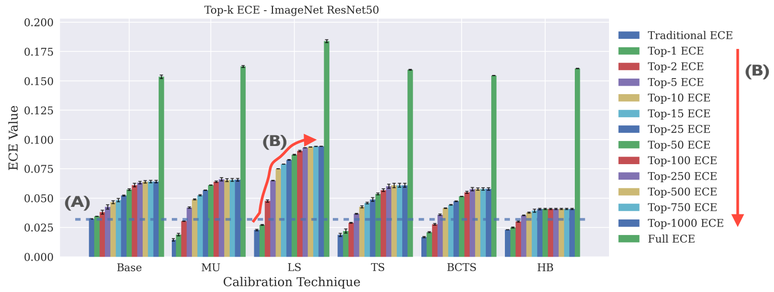
$$\begin{bmatrix} g(X)_{bird} \\ g(X)_{dog} \\ g(X)_{cat} \end{bmatrix} \leftarrow \Delta \rightarrow \begin{bmatrix} \mathbb{P}\left(Y \in bird \,|\, g(X)_{bird}\right) \\ \mathbb{P}\left(Y \in dog \,|\, g(X)_{dog}\right) \\ \mathbb{P}\left(Y \in cat \,|\, g(X)_{cat}\right) \end{bmatrix}$$

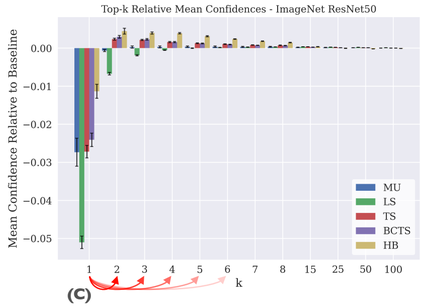
"Inter-Interval ECE"

Binary outputs interpreted using a Likert scale

$$\bar{g}_{[l,h)} = \mathbb{E}_X[g(X) \in [l,h)]$$

$$\begin{bmatrix} \bar{g}_{[0,0.33)} \\ \bar{g}_{[0.33,0.66)} \\ \bar{g}_{[0.66,1.00)} \end{bmatrix} \leftarrow \Delta \rightarrow \begin{bmatrix} \mathbb{E}_{X,Y}[Y \mid g(X) \in [0.0,0.33)] \\ \mathbb{E}_{X,Y}[Y \mid g(X) \in [0.33,0.66)] \\ \mathbb{E}_{X,Y}[Y \mid g(X) \in [0.66,1.0)] \end{bmatrix}$$



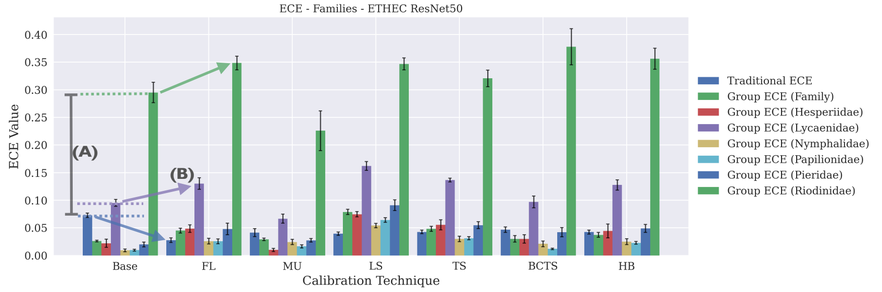


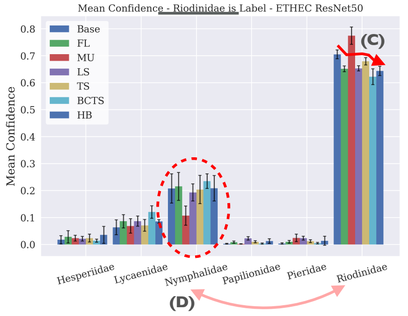
The data consists of the standard ImageNet dataset, with the model trained from scratch on the full label space of 1000 classes, before applying the Top-k lens.

(A) While all techniques do improve calibration by reducing the Traditional and Top-1 ECE as k increases, calibration techniques do not uniformly improve the top-k error metrics.

(B) Technique such as labels smoothing can actually worsen model reliability measured with respect to the top-5, top 10, and higher ECE lenses.

(C) In expectation, each technique reduces the magnitude of the most confident output and increases the magnitude of non-max outputs, suggesting that improvement to Traditional ECE is achieved simply by moving probability mass off of the maximum output.





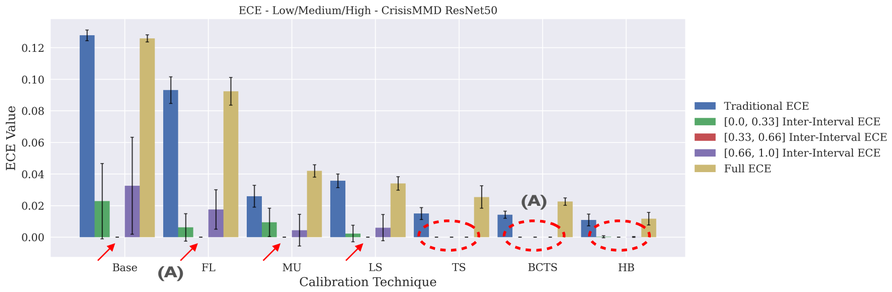
The data consists of butterfly images, with the model (pre-trained on ImageNet) trained on all 200 individual species labels, but then applying the Group lens for the species belonging to each of 6 taxonomical families. Confidence is visualized specifically for validation instances where the true family of the label was Riodinidae.

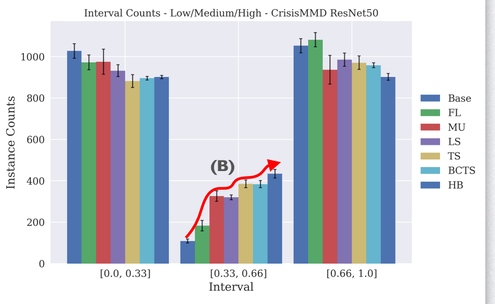
(A) The baseline model calibration is non-uniform across families.

(B) Calibrations technique may reduce Traditional ECE but can worsen calibration with respect to individual Group ECE metrics.

(C) Calibration techniques encourage lower confidence in the correct family, Riodinidae, because they move some probability mass over to incorrect classes.

(D) Model reports relatively high confidence in the Nymphalidae family, supported by domain literature that in fact Nymphalidae and Riodinidae butterflies are difficult to distinguish in the wild.





The data consists of image tweets that were either labeled "informative" or "not informative" for analysis and tracking of humanitarian crises. We train the model (pre-trained on ImageNet) for this binary task and evaluate classifier calibration for this problem under the assumption that classifier confidences will be used to express low ([0.0,0.33]), medium ([0.33,0.66]), and high ([0.66,1.0]) confidence categories to end-users.

(A) All models have zero calibration error for the medium category, but some calibration techniques result in zero calibration error for all intervals., indicating that all interventions produce outputs that better adhere to the specified Likert scale than the baseline

(B) All models output confidences in the medium interval less often than the two extremes, though the interventions produce more medium confidence outputs indicating that the baseline model incurs interval-specific error that the interventions reduce by moving high-confidence outputs to the medium interval.