THE UNIVERSITY OF SYDNEY

A Benchmark Study on Calibration

YounanZhu 530162636

COMP5329 Deep Learning

Research Problem

Deep neural networks have become increasingly popular in a variety of machine learning tasks. However, as these models become more complex, issues with miscalibration tend to rise, even as prediction accuracy improves. Numerous studieshave focused on enhancing calibration performance through data preprocessing, loss functions, and training frameworks, but the investigation into calibration properties has been somewhat neglected. In our study, we utilize Neural Architecture Search (NAS), which provides a comprehensive model architecture space, to thoroughly investigate calibration properties. We specifically construct a model calibration dataset which assesses 90 bin-based and 12 other calibration measurements across all networks contained in the extensively used NATS-Bench search space, consisting of 117,702 unique neural networks.
From this in-depth analysis, we aim to address several longstanding questions within the field:
(i) Can model calibration be generalized across different tasks?
(ii) How reliable are calibration metrics?
(iii) What is the impact of bin size on calibration measurement?
(iv) Can robustness be used as a calibration measurement?
(v) Which architectural designs are beneficial for calibration?
(vi) Does a post-hoc calibration method affect all models uniformly?
(vii) How does calibration interact with accuracy?
Our study also fills an existing gap in the exploration of calibration in NAS. By providing this dataset, we offer a valuable resource to further research on NAS calibration. To our knowledge, this research represents the first large-scale investigation of calibration properties, and the inaugural study of the calibration problem within NAS.

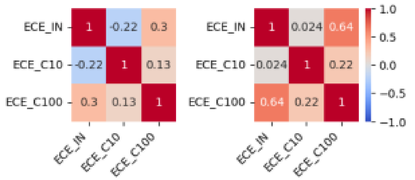
Methodology

In our work, we introduce a calibration dataset based on NATS-Bench. Specifically, we evaluate all 117,702 unique architectures concerning topology and model size, and benchmark them on multiple calibration metrics of different types.

Experiments and Discussion

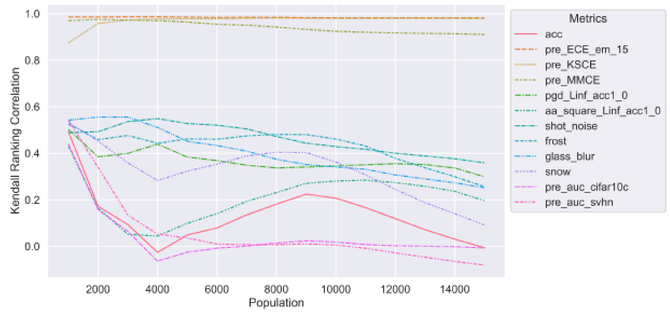
Can model calibration be generalized across different tasks?

The calibration property of a certai architecture can not generalize we to different tasks



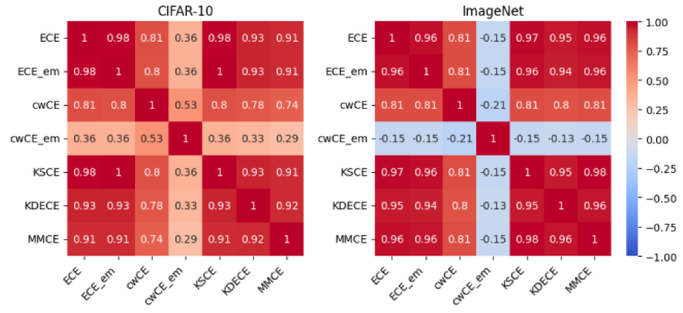
Can robustness be used as a calibration measurement?

Calibration performance can be measured not only by the robustness accuracy on the corruption dataset, but also by other robustness metrics only among models with high prediction accuracy.



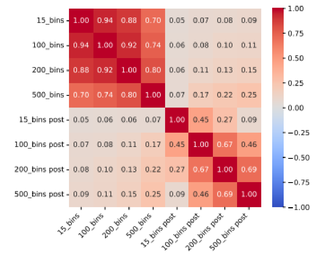
How reliable are calibration metrics

A consistent trend in the ranking of most calibration performance regardless of metric type.
cwCEem may not be a reliable metric for calibration measurement.



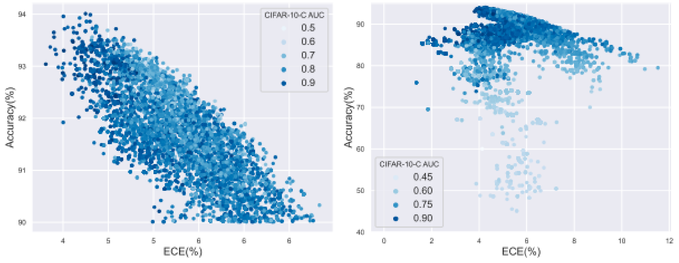
Does a post-hoc calibration method affect all models uniformly?

Well-calibrated models do not necessarily exhibit better calibration performance after post-hoc calibration techniques



How does calibration interact with accuracy

The trade-off between accuracy and calibration exists only among architectures with prediction performance.



What is the impact of bin size on calibration measurement?

Bin size has a more substantial impact on post-ECE
For a holistic comparison, it is recommended to assess post-hoc calibration performance across a range of bin sizes.