Motivation

By

Debaditya Roy

For the course - **FID3018**

Debaditya Roy

May 2021

Neural networks are one of the primary machine learning methods that are omnipresent in modern applications. The expressive power of neural networks in terms of feature composition and their results make them a formidable tool for modern-day machine learning applications.

As the wise say *"With great power comes great responsibility"*, and this phrase needs to be relooked from the perspective of plugging in neural network architectures in almost all ML-based systems. Modern-day neural networks produce excellent results in the downstream tasks on which they are applied. However, they suffer from reliability and safety because the predictions they produce are not well calibrated.

In my review task I wish to explore the aspect of *neural network calibration* to improve model reliability and safety. The three papers selected are as follows:

1. *On calibration of modern neural networks*[1]
2. *Verified Uncertainty Calibration*[2]

3. *Learning of Single-Shot Confidence Calibration in Deep Neural Networks through Stochastic Inferences*[3]

While confidence calibration of neural network predictions was explored in early 2000, it took a back seat. The first two papers focus on the post-processing method of confidence calibration requiring a separate validation set. In contrast, the last paper focuses on constructing a loss function that calibrates the confidence without requiring any validation data-set.

References

[1] Chuan Guo et al. "On calibration of modern neural networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1321–1330.

[2] Ananya Kumar, Percy Liang, and Tengyu Ma. "Verified uncertainty calibration". In: *arXiv preprint arXiv:1909.10155* (2019).

[3] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. "Learning for single-shot confidence calibration in deep neural networks through stochastic inferences". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9030–9038.