Tata Institute of Fundamental Research

Survey No. 36/P, Gopanpally Village, Serilingampally, Ranga Reddy Dist., Hyderabad – 500 046

Colloquium

Mitigating Simplicity Bias in Deep Learning

Praneeth Netrapalli

Google Research India, Bengaluru

While deep neural networks have achieved large gains in performance on benchmark datasets, their performance often degrades drastically with changes in data distribution encountered during real-world deployment. In this work, through systematic experiments and theoretical analysis, we attempt to understand the key reasons behind such brittleness of neural networks in real-world settings and propose algorithms that can train models that are more robust to distribution shifts. We first hypothesize, and through empirical + theoretical studies demonstrate, that (i) neural network training exhibits "simplicity bias" (SB), where the models learn only the simplest discriminative features and (ii) SB is one of the key reasons behind non-robustness of neural networks. We then delve deeper into the nature of SB and find that while the network's backbone learns both simple and complex features, it is the final classifier layer which does not use these features in the eventual prediction. We posit two reasons for this: 1. Dominance of non-robust features and 2. Replication of simple features, leading to over-dependence of the final layer linear classifier on these and empirically validate these hypotheses on semi-synthetic and real-world datasets. We then propose two methods to deal with both of these phenomena, and show gains of upto 1.5% over the state-of-the-art on DomainBed - a standard and large-scale benchmark for domain generalization.

We will end with some thoughts on what SB and robustness mean in the new world of large language models (LLMs).

Based on several joint works with Anshul Nasery, Sravanti Addepalli, Depen Morwani, Harshay Shah, Jatin Batra, Kaustav Tamuly, Aditi Raghunathan, R. Venkatesh Babu and Prateek Jain.

Monday, Aug 28th 2023

4:00 PM (Tea/Coffee at 03:45 PM)

Auditorium, TIFR-H