Stanford MEDICINE | Department of Biomedical Data Science

BIODS 260B

How do neural networks learn features from data?

Adit Radhakrishnan

George F. Carrier Postdoctoral Fellow
in the School of Engineering and Applied Sciences at Harvard
and an affiliate with the Broad Institute of MIT and Harvard

January 18, 2024

1:30PM-2:50PM

MSOB X303

Abstract:

Understanding how neural networks learn features, or relevant patterns in data, for prediction is necessary for their reliable use in technological and scientific applications. We propose a unifying mechanism that characterizes feature learning in neural network architectures. Namely, we show that features learned by neural networks are captured by a statistical operator known as the average gradient outer product (AGOP). Empirically, we show that the AGOP captures features across a broad class of network architectures including convolutional networks and large language models. Moreover, we use AGOP to enable feature learning in general machine learning models through an algorithm we call Recursive Feature Machine (RFM). We show that RFM automatically identifies sparse subsets of features relevant for prediction and explicitly connects feature learning in neural networks with classical sparse recovery and low rank matrix factorization algorithms. We conclude with a culminating biomedical application where we use RFM to screen for synthetically lethal gene pairs, a potential avenue for selectively targeting cancer cells based on genetic vulnerabilities. Overall, this line of work advances our fundamental understanding of how neural networks extract features from data, leading to the development of novel, interpretable, and effective models for use in scientific applications.