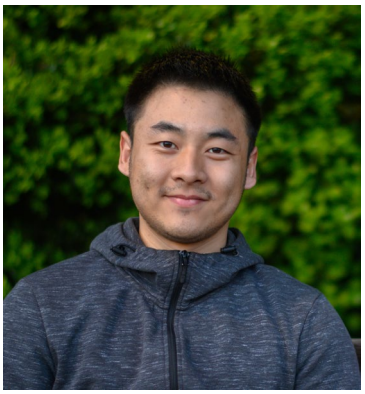
PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering

Faculty Candidate Seminar -- Purdue Computes: AI/ML



Yaodong Yu

PhD Student, EECS Dept.
UC Berkeley

Thursday, March 21, 2024
10:30 A.M. – 11:30 A.M.
MSEE 112

Towards Transparent Representation Learning

Abstract

Machine learning models trained on vast amounts of data have achieved remarkable success across various applications. However, they also pose new challenges and risks for deployment in real-world high-stakes domains. Decisions made by deep learning models are often difficult to interpret, and the underlying mechanisms remain poorly understood. Given that deep learning models operate as black boxes, it is challenging to understand, much less resolve, various types of failures in current machine learning systems.

In this talk, I will describe our work towards building transparent machine learning systems through the lens of representation learning. First, I will present a white-box approach to understanding transformer models. I will show how to derive a family of mathematically interpretable transformer-like deep network architectures by maximizing the information gain of the learned representations. Furthermore, I will demonstrate that the proposed interpretable transformer achieves competitive empirical performance on large-scale real-world datasets, while learning more interpretable and structured representations than black-box transformers. Next, I will present our work on training the first set of vision and vision-language foundation models with rigorous differential privacy guarantees and demonstrate the promise of high-utility differentially private representation learning. To conclude, I will discuss future directions towards transparent and safe AI systems we can understand and trust.

Bio

Yaodong Yu is a PhD student in the EECS department at UC Berkeley, advised by Michael I. Jordan and Yi Ma. His research focuses on foundations and applications of trustworthy machine learning, including interpretable deep neural networks, privacy-preserving foundation models, and uncertainty quantification under complex environments. He is the recipient of CPAL-2024 Rising Star Award, and first place in the NeurIPS-2018 Adversarial Vision Challenge.