GEORGIA

1785

Department of Statistics
UNIVERSITY OF GEORGIA
Colloquium Series

Toward Trustworthy Machine Learning Under Training-Time Adversaries

When and Where:

10/17/2024

4:00 PM — 5:00PM

Room 204 Caldwell Hall

Zhen Xiang

Abstract:

Machine learning has demonstrated remarkable performance across various applications. However, significant concerns have arisen about its trustworthiness, such as risks posed by training-time adversaries. Facilitated by the need for 'big data', whether publicly available or locally accessible, training-time adversaries can easily manipulate a machine learning model by altering the distribution of its training data. This manipulation can lead to potentially catastrophic consequences in safety-critical applications. In this talk, I will focus on a type of threatening yet stealthy training-time adversarial attack known as a 'backdoor attack'. During inference, a backdoored model will predict a specific adversarial target class whenever a test instance is embedded with a backdoor trigger prescribed by the adversary, while maintaining high performance on benign instances. First, we will explore the role of downstream users or third-party inspectors in detecting whether a given model is backdoored, without access to its training dataset. In particular, we will apply optimization and statistical approaches to address two challenging problems: 1) How can we detect backdoors irrespective of the trigger type? 2) How can we 'certify' a backdoor detector, i.e., provide a theoretical guarantee of its detection performance? Second, we will perform red-teaming on commercial large language models, such as GPT3.5, GPT4, and PaLM2, by proposing a backdoor attack for in-context learning with chain-of-thought prompting. We advocate for the development of new approaches to address evolving practical challenges in ensuring the trustworthiness of models.

About the Speaker:

Zhen Xiang is an assistant professor from School of Computing, University of Georgia. Before that, he was a postdoc at the Secure Learning Lab, University of Illinois Urbana-Champaign. Zhen Xiang earned his PhD from Pennsylvania State University in 2022, with the Dr. Nirmal K. Bose Dissertation Excellence Award. Zhen Xiang's research interests include trustworthy machine learning, AI safety, large language models and agents, and statistical signal processing. He serves as the PC member for conferences such as ICLR, NeurIPS, and ICASSP, the reviewer for journals including TPAMI, TNNLS, and TIFS, and the associate editor for TCSVT. Zhen Xiang is also the leading organizer of the Competition for LLM and Agent Safety at NeurIPS 2024, and the co-organizer of the first Trojan Removal Competition and the second Trojan Detection Challenge.