Department of Electrical Engineering
香港城市大學
City University of Hong Kong

CityU Architecture Lab for Arithmetic and Security (CALAS) Seminar Series

Backdoors in Deep Learning: The Good, the Bad and the Ugly

Prof. Yingjie LAO, Associate Professor
Tufts University, Boston, USA

Abstract:

Deep learning is revolutionizing almost all AI domains and has become the core of many modern AI systems. Despite their superior performance compared to classical methods, deep learning also faces new security problems, such as adversarial and backdoor attacks, that are hard to discover and resolve due to their black-box-like property. Backdoor attacks are possible because of insecure model pretraining and outsourcing practices. Malicious third parties can add backdoors into their models or poison their released data before delivering it to the victims to gain illegal benefits. This threat seriously damages the safety and trustworthiness of AI development. While most works consider backdoors “evil”, some research explores leveraging them for positive purposes. A notable approach involves using backdoors as watermarks to detect illegal uses of commercialized data/models. Watermarks can also be used for detecting AI-generated data, particularly with the rise of large generative models like LLMs. In this presentation, I will share insights from our recent research, exploring both the “good” and “bad” aspects of backdoors in deep learning.

Biography:

Yingjie LAO is currently an associate professor in the Department of Electrical and Computer Engineering at Tufts University. He received his Ph.D. degree from the Department of Electrical and Computer Engineering at University of Minnesota, Twin Cities, in 2015. His research has been recognized with an NSF CAREER Award, an IEEE TVLSI Prize Paper Award, and an ISLPED Best Paper Award. His research interests include trusted AI, hardware security, electronic design automation, VLSI architectures for machine learning and emerging cryptographic systems, and AI for healthcare and biomedical applications.

24 May 2024 (Fri); 9:00am - 11:00am HKT; P1402; https://cityu.zoom.us/j/96742093029