香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

TMLR Young Scientist SEMINAR

2023 SERIES

Trustworthy Machine Learning and Reasoning Group

Mr. Yi Zeng

Ph.D. candidate,
Computer Engineering,
Virginia Tech.

Date: 09 Sept 2023 (Saturday)

Time: 09:00 – 10:40 (HKT)

Meeting: https://hkbu.zoom.us/j/6603117755

Data-centric Backdoor Attacks and Countermeasures

ABSTRACT

As machine learning models trained on vast data become integral to various applications, they have also become vulnerable to adversarial attacks. Backdoor attacks pose a particularly stealthy danger among these threats by subtly injecting malicious data that causes the model to make incorrect predictions. To shed light on this emerging challenge, we will go through our recent efforts to identify and defend against state-of-the-art backdoor attacks in this presentation. We will first introduce our latest work, NARCISSUS (CCS'23). NARCISSUS is a practical clean-label backdoor attack that can bypass most existing defenses by synthesizing and exploiting representative in-ward pointing noise from the target class. This highlights the need for more robust countermeasures. Accordingly, we will discuss our proposed defense, ASSET (Usenix Security'23). ASSET actively differentiates between potential backdoor samples and clean samples by introducing opposite learning behaviors, enabling reliable and effective detection of backdoor attacks across various learning paradigms. However, like many existing backdoor defenses, ASSET's efficacy also hinges on a clean subset of the same distribution as the training data. To study this limitation, we have developed META-SIFT (Usenix Security'23). META-SIFT precisely sifts out clean data subsets from contaminated datasets, surpassing existing automated methods and human inspection. Together, these studies demonstrate recent pioneering techniques for Data-centric Backdoor Attacks and Countermeasures.

BIOGRAPHY

Yi Zeng, a third-year Ph.D. candidate in Computer Engineering at Virginia Tech, specializes in adversarial machine learning under the guidance of Professor Ruoxi Jia. An alumnus of the University of California, San Diego, with an M.S. degree and Xidian University with a B.S., Yi's research has been showcased at premier learning and security venues, including ICLR, NeurIPS, ICML, TMLR, IJCAI, ICCV, USENIX Security, and ACM CCS. As the chair of the ICLR 2022 IEEE Trojan Removal Competition, he garnered significant media attention. Among his accolades are the USENIX Student Travel Grant (2023), Amazon Fellowship (2022), and the ICA3PP Best Paper Award (2019).

ENQUIRY

Email: bhanml@comp.hkbu.edu.hk