NJIT | INSTITUTE FOR DATA SCIENCE

Data Science Seminar Series

In Collaboration with
The Department of Data Science

Backdoor in AI: Algorithms, Attacks, and Defenses

Hosted by Shuai Zhang

Ruixiang Tang

Rice University

Date: Thursday, February 22, 2024
Time: 2:30 PM - 3:30 PM (Coffee served at 2:15 PM)
Location: GITC Building Room 4402 (4th floor Seminar Room)
Web Link: Zoom Meeting Link

As deep learning models are increasingly integrated into critical domains, their safety emerges as a critical concern. This talk delves into the emerging threat of backdoor attacks. These attacks involve embedding a backdoor function within the victim model, allowing attackers to manipulate the model's behavior using specific triggers. I will begin by identifying key stages in the machine learning lifecycle where backdoors can be injected into deep learning models. Specifically, I will discuss a new attack vector that operates during the post-training stage. Then I will introduce my recent research on defending against backdoor attacks. I have designed a honeypot module to absorb all the backdoor functionality, safeguarding the integrity of the stem network. The talk will also explore the security risks in advanced large language models, with a particular focus on understanding how backdoor and data poisoning attacks pose threats to these generative models. Finally, I will conclude by providing an overview of my research in trustworthy AI and outline future research directions.

Ruixiang (Ryan) Tang is a final-year Ph.D. student at Rice University. His research is primarily concentrated on Trustworthy Artificial Intelligence (AI), with specific emphases on security, privacy, and explainability. He has over 20 research in leading machine learning, data mining, and natural language processing venues such as NeurIPS, ICLR, AAAI, KDD, WWW, TKDD, ACL, EMNLP, and Communications of the ACM. Additionally, He closely collaborates with healthcare institutes, such as Yale, Baylor, and UThealth to facilitate the deployment of reliable large language models in the healthcare sector. He has been acknowledged as AMIA'23 Best Student Paper Award, AMIA'22 Best Student Paper (Shortlist) Award, as well as CIKM'23 Honorable Mention for Best Demo Paper Award.