香港浸會大學理學院
HKBU Faculty of Science

DEPARTMENT OF COMPUTER SCIENCE

MPhil Degree Oral Presentation

MPhil Candidate:	Mr. Buhua LIU
Date	24 February 2025 (Monday)
Time:	4:30 pm – 6:30 pm (35 mins presentation and 15 mins Q & A)
Venue:	Zoom (Meeting ID: 998 6975 3218) (The password and direct link will only be provided to registrants)
Registration:	https://bit.ly/bucs-reg (Deadline: 1:00pm, 23 February 2025)

Advances in Backdoor Attacks and Defenses

Abstract

Deep neural networks (DNNs) are vulnerable to backdoor attacks, where an attacker can embed hidden functionality into a model by manipulating only a small portion of its training data, while leaving its normal operations unaffected. This thesis introduces two major contributions to this filed: (1) an innovative influence-based detection approach to improve backdoor defenses in classic supervised learning (SL) settings, and (2) MetaBackdoor, a cross-paradigm backdoor attack framework that operates effectively across both supervised and contrastive learning (CL) paradigms.

First, we propose a novel defense strategy that enhances backdoor detection by introducing the concept of influence space, where data points are traced through their influence on model parameters rather than conventional model outputs. While many existing backdoor detection methods rely on the latent separability assumption—that clean and poisoned samples can be differentiated in representation space—recent adaptive backdoor attacks can easily bypass this assumption, rendering such methods ineffective. Our influence-based approach overcomes this limitation by achieving significantly stronger separability across statistical measures, with the Silhouette Score increasing by 122% on average. This enhanced separability enables robust backdoor detection and defense, with our influence-based methods outperforming traditional representation-based baselines across eight representative backdoor attacks. Specifically, our approach reduces the average attack success rate by 43.4 points (from 47.2% to 3.8%) across three benchmark datasets.

Second, we introduce MetaBackdoor, a novel backdoor attack framework that unifies attack mechanisms across SL and CL. MetaBackdoor employs a bi-level optimization process to design transferable backdoor triggers that exploit shared feature space properties across both SL and CL. This framework achieves an average increase in attack success rate of over 77.8 points in CL settings while effectively transferring to SL. MetaBackdoor reveals that differences in backdoor behaviors between SL and CL may stem from heuristic trigger designs rather than fundamental distinctions between the paradigms. This insight provides a foundation for developing future cross-paradigm backdoor attacks and defenses.

***** ALL INTERESTED ARE WELCOME *****