香港浸會大學理學院
HKBU Faculty of Science

DEPARTMENT OF COMPUTER SCIENCE

PhD Degree Oral Presentation

PhD Candidate:	Mr. ZHU Jianing
Date	14 May 2025 (Wednesday)
Time:	10:00 am – 12:00 nn (35 mins presentation and 15 mins Q & A)
Venue:	ZOOM (Meeting ID: 970 4876 9288) (The password and direct link will only be provided to registrants)
Registration:	https://tinyurl.com/bucs-tdreg (Deadline: 12:00 nn, 13 May 2025)

Towards Trustworthy Machine Learning for Out-of-distribution Data

Abstract

The impressive success of machine learning, especially with the advancement of deep neural networks, is generally built on a core assumption in machine learning, i.e., the training and test data are drawn from the same underlying distribution. However, it may not be valid in practice given the complex environments encountered during the inference stage, raising concerns about model trustworthiness. Generally, considering a data pair (x, y), distribution differences can occur on either the x (input) or y (label) side, resulting in out-of-distribution data that differ from the training ones. On the x side, deep neural networks are vulnerable to adversarial examples with imperceptible perturbations that mislead the prediction. On the y side, models can exhibit overconfidence in the input data from a different label space (termed OOD data) in open-world deployment. These issues highlight the need to enhance robustness against adversarial perturbations for better generalization on semantic-invariant changes; and improve reliability in OOD detection for eliciting awareness of unknown data. This thesis focuses on addressing these challenges in building trustworthy machine learning models with out-of-distribution data.

First, we study the reliable adversarial distillation from a well-trained robust model. Building upon the conventional adversarial training method for robust learning, we employ robust model output as soft supervision to alleviate its learning difficulty. By revealing the unreliable issue of teacher supervision, we propose Introspective Adversarial Distillation (IAD), which encourages the student model to partially trust robust supervision from the teacher model and partially trust the model self-introspection for distilling a more adversarially robust model.

Second, we study obtaining an aggregated robust model in distributed learning. Motivated by the data-hungry property of adversarial training on pursuing better adversarial robustness, we consider a decentralized training paradigm like federated learning for involving more data without raising extra concern about privacy and expect to aggregate an adversarially robust model. Tackling the exacerbated heterogeneity issue, we propose Slacked Federated Adversarial Training (SFAT) to address the algorithm in-compatibility and enable distributed adversarial training.

Third, we study unleashing the OOD detection capacity of a well-trained model. Based on the scoring functions for identifying OOD data, we challenge that the model trained on the main task (rather than directly differentiating OOD data from in-distribution data) may not be optimal for OOD detection. Under an in-depth exploration and understanding of the effects of model knowledge from a data perspective, we propose Unleashing Mask (UM) and its variant UMAP to regularize a model to forget some atypical samples that hinder the OOD distinguishability.

Finally, we study outlier exposure to actively enable model to identify OOD data with auxiliary outliers. Given the collection cost and difficulty of representing all unseen OOD samples, we propose Diversified Outlier Exposure (DivOE) to generate diversified outliers based on existing data, which conducts informative extrapolation that enhances model reliability even with limited auxiliary ones. In addition to demonstrating our method effectiveness with comprehensive experimental results, we also conclude general principles to trustworthy algorithm design and discuss the potential future research directions

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