香港浸會大學理學院
HKBU Faculty of Science

DEPARTMENT OF COMPUTER SCIENCE

PhD Degree Oral Presentation

PhD Candidate:	Mr. LEI Zijian
Date	15 January 2025 (Wednesday)
Time:	9:00 am – 10:30 am (35 mins presentation and 15 mins Q & A)
Venue:	ZOOM (Meeting ID: 973 1049 9830) (The password and direct link will only be provided to registrants)
Registration:	https://bit.ly/bucs-reg (Deadline: 12:00 nn, 14 January 2025)

Unleashing the Power of Random Projection for Efficient Machine Learning

Abstract

The computational cost of AI algorithms poses challenges for their real-world applications. Random projection (RP) is a promising dimensionality reduction technique that can reduce the computational requirements of training AI models. In thesis research, we investigate three pivotal methodologies for: i) data-driven random projection for dimensionality reduction, ii) efficient on-device training, and iii) fine-tuning of large language models, to achieve scalable and resource-efficient machine learning (ML) systems.

RP produces new low-dimensional embedding without considering the intrinsic property of the datasets. We propose importance sampling methods based on their label information to produce low-dimensional embedding. For highly sparse input data, we also discover a connection between count-sketch and k-means, and apply k-means with gradient descent and ϵ-$\mathcal{L}_1$ ball projection to obtain a low-dimensional sketched matrix for enhancing the efficiency of subsequent ML tasks.

In addition to compact embedding, deploying machine learning algorithms on memory-limited devices poses another challenge We propose a novel memory and computation-efficient kernel SVM model by using binary embedding and ternary model coefficients. We derive a simple and yet effective coordinate descent algorithm for the learning that can support different types of loss function and regularizer.

With the proliferation of large language models, we further explore RP to achieve parameter-efficient fine-tuning. We propose Randomized Walsh-Hadamard Transform to achieve significant reduction in the size of trainable parameters compared to the state-of-the-art methods including low-rank adaptation (LoRA) and its variants. It also allows a PAC-Bayes regularizer to be efficiently incorporated to improve generalization

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