For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics

PHD STUDENT SEMINAR

Stabilizing Pre-LN Transformers via Gated Residual Paths

By

Mr. Tianhao CHEN

Abstract

Residual connections and normalization have been shown to improve the training stability of deep neural networks. Modern Large Language Models (LLMs), such as the LLaMA, Qwen, and DeepSeek series, predominantly adopt the Pre-LayerNorm (Pre-LN) architecture, which integrates residual connections with LayerNorm around Self-Attention and Feed-Forward Networks (FFNs).

However, Pre-LN Transformers often exhibit an exponential growth in layer activation variance. Since the gradient of LayerNorm is approximately inversely proportional to its input variance, this leads to vanishing gradients in the Attention and FFN blocks—especially in deeper layers where the variance accumulates. As a result, deeper layers contribute less to learning, which can hinder both the performance and efficiency of LLM pretraining.

To address this issue, several architectural modifications have been proposed. In this seminar, we briefly review existing approaches and introduce a simple modification to the Pre-LN architecture that improves pretraining efficiency.

Date : 6 May 2025 (Tuesday)
Time : 11:00am
Venue : Room 4503 (near Lifts 25/26)

All are Welcome!