Department of Data Science

香港城市大學
City University of Hong Kong

DS SEMINAR

LightSeq: High-Performance Training and Inference for Transformer Models

Date: 9 April 2025 (Wednesday)

Time: 8:00pm - 9:30pm

Seminar Link: https://cityu.zoom.us/j/82927639678

ABSTRACT

Transformer-based models, with self-attention mechanisms at their core, have been widely adopted in natural language processing and computer vision. However, their massive computational demands pose challenges for industrial applications, including high costs, lengthy offline training times, and high online inference latency.

To address these challenges, LightSeq integrates multiple optimization techniques—such as operator fusion, parallel decoding, asynchronous computation, and memory optimization—boosting Transformer model inference speeds by 4-10x and training speeds by 1.5-3x, while maintaining compatibility with any deep learning framework.

Currently, LightSeq has been deployed across ByteDance's internal products, including Volcano Translation, Search, Ads, Recommendation, Education, and E-commerce, and has gained significant traction in the open-source community.

Mr. XIONG Ying

GUEST SPEAKER'S PROFILE

Mr. XIONG Ying is a High-Performance Computing Algorithm Engineer at Microsoft Asia Engineering. He holds bachelor's and master's degrees from Beijing University of Posts and Telecommunications. Previously at ByteDance AI Lab, he focused on training and inference acceleration, contributing to the LightSeq open-source project and the SEED foundation model.

Enquiries: ds.go@cityu.edu.hk

All are welcome