



Efficient Inference and Fine-tuning of LLM on the Edge

by
Prof. Li Li

University of Macau, Macau

Date : 12 Aug 2026 (Wednesday)

Time : 11:00am – 12:00nn

Venue : Rm833, 8/F, Ho Sin-Hang Engineering Building, CUHK

Abstract

This talk focuses on efficient inference and fine-tuning of large language models on edge devices, addressing key challenges in computation, memory, energy, and real-time performance on mobile platforms. The talk begins by pointing out that deploying separate dense models for each application does not scale on mobile devices. To solve this, a unified serving framework is proposed, incorporating QoS-aware control, application-aware reuse, and GPU runtime optimizations to reduce latency while maintaining output quality. The talk then explores memory bottlenecks in on-device training and fine-tuning, introducing hybrid compression strategies and novel tensor layouts that achieve significant speedups. For more complex multimedia and robotic scenarios, the talk presents an NPU-based acceleration scheme for video understanding that fills idle execution resources, as well as a cache-aware approach for robotic control that boosts inference speed without sacrificing task success. Finally, the talk looks ahead to thermal-aware scheduling challenges for next-generation wearables like smart glasses. Overall, this talk systematically presents the complete technical landscape of edge intelligence, from efficient inference and on-device training to physical-world interaction, bridging theory and practice.

Biography

Dr. Li Li is an Assistant Professor at the University of Macau, affiliated with the State Key Laboratory of Internet of Things for Smart City and the Department of Computer and Information Science. He completed his Ph.D. (2018) and M.S. (2014) at The Ohio State University, following a B.S. from Tianjin University in 2011. His research spans efficient fine-tuning and inference of large language models on edge devices, on-device learning, mobile/cloud computing, autonomous driving, and distributed learning. He has published extensively in top-tier venues including MICRO, ATC, RTSS, MobiSys, SENSYS, NIPS (Oral), ICML (Spotlight), ICLR, ACL, KDD, SoCC, and CVPR, with a strong emphasis on resource-efficient LLM adaptation, on-device model optimization, memory-efficient fine-tuning, and real-time collaborative systems. Dr. Li has also gained research experience at institutions such as Shenzhen Institute of Technology, Chinese Academy of Science, and Microsoft Research.

**** ALL ARE WELCOME ****