Accelerating Large Mixture-of-Experts Models via Pipelining and Scheduling

Prof. Xiaowen Chu

(IEEE Fellow, Head of the Data Science and Analytics Thrust)

The Hong Kong University of Science and Technology (Guangzhou)

Abstract: In recent years, large-scale deep neural network models can be easily scaled to trillions of parameters with sparsely activated mixture-of-experts (MoE), which significantly improves the model quality while only requiring a sub-linear increase in computational costs. However, the dynamic activation of MoE experts introduces extensive communications, limiting the scaling efficiency of distributed systems. In this talk, we will present some of our recent work on improving the training efficiency of MoE-based LLM models, which leverages two simple ideas: pipelining and scheduling.