ICTS Seminar

Title : Efficient Large Language Model Inference with HiRE and Tandem Transformers

Speaker : Praneeth Netrapalli (Google Research India, Bengaluru)

Date : Friday, 08 November 2024

Time : 11:00 AM (IST)

Abstract : We will first give an overview of the Transformer architecture and Large Language Models (LLMs), and explain the key bottlenecks in LLM inference. After giving a birds eye view of the various approaches that have been proposed in the literature for speeding up LLM inference, we explain in detail two of the approaches that we have developed recently. The first approach focuses on exploiting the inherent sparsity in different layers of LLMs. More concretely, despite significant sparsity within these layers, efficient exploitation is hindered by a lack of accelerator support for unstructured sparsity and the computational cost of identifying important elements. We introduce HiRE, a novel technique that utilizes dimensionality reduction and quantization to predict the significant elements with high recall, followed by focused computation and an efficient approximate top-k operator. Applied to softmax and a group-sparse FFN layer, HiRE significantly reduces computational cost while preserving accuracy, leading to improved end-to-end inference latency. Second, we tackle the inherent sequential generation bottleneck of LLMs with tandem transformers. This architecture combines a small autoregressive model with a large block-mode model, where the small model leverages the large model's representations for improved accuracy. This results in enhanced prediction, faster inference, and the option of a verification step to ensure quality. Our approach demonstrates superior performance compared to standalone models and addresses the limitations of existing parallel decoding techniques.

Based on joint works with Yashas Samaga B L, Varun Yerram, Aishwarya P S, Pranav Nair, Srinadh Bhojanapalli, Chong You, Toby Boyd, Sanjiv Kumar, Spandana Raj Babbula and Prateek Jain. The talk will assume only basic familiarity with machine learning (ML) in general, and not assume any prior familiarity with transformer architecture or LLMs.

Venue : Emmy Noether Seminar Room
Zoom Link: https://icts-res-in.zoom.us/j/91981337040?pwd=DemWDaAjVdyUumb7TyKrKKQptVIEki.1
Meeting ID: 919 8133 7040
Passcode: 202030