QMUL - China Scholarship Council PhD Project

Primary supervisor: Ahmed M. A. Sayed (ahmed.sayed@qmul.ac.uk)

Group: SAYED Systems Group (https://sayed-sys-lab.github.io)

Deadline for application to QMUL: 29 Jan 2025

Title: Towards Resource Efficient Training and Fine-Tuning of Generative AI and LLMs

Abstract:
In the rapidly evolving landscape of Artificial Intelligence (AI), developing sophisticated Generative AI and Large Language Models (LLMs) has become pivotal for various applications, ranging from natural language processing to creative content generation. However, training these models is computationally intensive, often requiring substantial time and resources and limiting its scalability. This project will study and propose system and algorithmic optimizations to accelerate the training process for Generative AI and LLMs, addressing the challenges posed by the complexity of these models. This research focuses on exploring and implementing advanced parallel computing techniques, leveraging the power of distributed systems and specialized hardware accelerators. By optimizing algorithms, employing parallelization strategies, and harnessing the capabilities of GPUs, TPUs, or emerging AI-specific hardware, this project aims to reduce the training/fine-tuning time of Generative AI and LLMs significantly. Furthermore, the study delves into transfer learning and explores techniques to enhance model convergence and accuracy. By leveraging pre-trained models and developing novel knowledge transfer learning methodologies, the research intends to minimize the data and computational resources required for training, democratizing access to cutting-edge AI technologies. This project will also work on designing efficient architectures for Generative AI models and study network pruning or sparsity techniques to create lightweight yet effective models.