Name: Aditya Kendre

Campus Affiliation: Penn State Harrisburg

Major: Computer Science

Anticipated Graduation Date: May 2024

Mentors: Hien Nguyen (Penn State Harrisburg), Suman Saha (Penn State University Park)

Partner: None

Project Title: Next-Generation Language Model for On-Device Natural Language Processing

Large Language Models (LLMs) have brought significant advancements in Natural Language Processing (NLP), enabling breakthroughs in tasks like translation and text generation. However, their immense size and computational requirements pose challenges for deployment in resource-constrained environments. To address this, knowledge distillation has emerged as a technique to transfer knowledge from larger models to smaller ones. In this paper, we demonstrate the effectiveness of fine-tuning by training the Pythia-1B model on an instruction database. This allows the smaller model to benefit from the advanced language capabilities and reasoning abilities. Furthermore, we emphasize the deployability of the proposed model by choosing a well-calibrated smaller model, ensuring optimal computation-to-performance ratio, particularly for resource-constrained devices like smartphones. By deploying the LLM closer to the consumer, we observe decreased latency, enhanced privacy, and the ability to function effectively in offline or low-connectivity scenarios. This work contributes to making LLM technology more accessible and practical, with implications for improving user experience and privacy in real-world applications.