A Novel Theory-Driven Approach for Model Compression in Deep Neural Networks

Yu, Alyssa (School: Poolesville High School)

In this study, I present a novel approach for model compression to reduce the size of neural network models. In recent years, with the rise of LLMs such as ChatGPT, deep learning models have grown significantly in complexity, garnering attention for remarkable performance but also raising concerns about the substantial computing resources they require. This has created significant challenges in fine-tuning or re-training models. I approach model compression by integrating rank selection into the low-rank training process and performing independent layer-wise rank selection guided by a theoretical loss error bound. Specifically, I conducted a comprehensive theoretical analysis to quantify how low-rank approximations impact the training losses. Building on these insights, I developed an efficient layer-wise rank search algorithm and incorporated it into low-rank singular value decomposition (SVD) training. The evaluation results on benchmark datasets demonstrate that the proposed approach can achieve high prediction accuracy while delivering significant compression performance. My research has broad impacts in future AI applications in medical diagnostics, where faster detection of diseases is needed in emergency scenarios. Moreover, it will allow for quicker object detection and decision-making in autonomous vehicles, where fast-reaction times are important. Finally, AI applications will become more eco-friendly, requiring less power to make predictions and lower carbon emissions from data centers.

Awards Won: