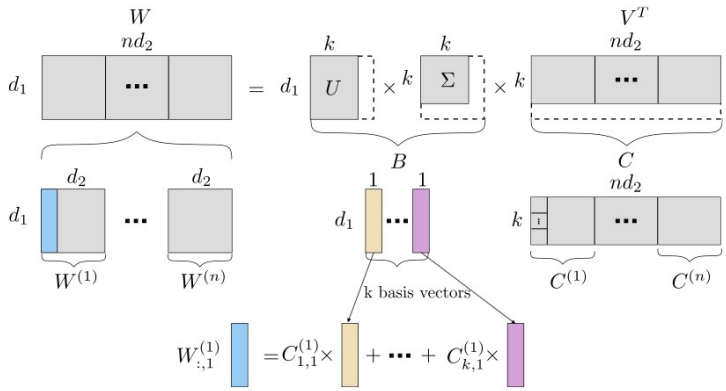
Dynamic Rank Allocation for Efficient LLM Inference on GPUs

Large Language Models (LLMs) have achieved remarkable breakthroughs. However, the huge number of parameters in LLMs require significant amount of memory storage in inference, which prevents their practical deployment in many applications. To reduce memory storage of LLMs, singular value decomposition (SVD) provides a promising solution to approximate weight matrices for compressing LLMs. Parameter sharing across different layers with SVD is also explored to achieve more effective compression for LLMs, as shown in the following figure.

In basis sharing, weight matrices in different layers are decomposed and represented as a linear combination of a set of shared basis vectors and unique coefficients. The types of weight matrices and the layer selection for basis sharing are examined when compressing LLMs to maintain the performance. Comprehensive experiments demonstrate that Basis Sharing outperforms state-of-the-art SVD-based compression approaches and parameter sharing techniques, especially under large compression ratios.

In this master's thesis, the objective is to investigate dynamic rank allocation, where the decomposition rank is adapted based on the difficulty of input prompts. For instance, more complex inputs would be assigned a higher rank to preserve model performance, while simpler inputs would use a lower rank to achieve faster inference.



References: https://arxiv.org/pdf/2410.03765

If you are interested in this topic for master thesis, please contact:

Prof. Dr.-Ing. Li Zhang (grace.zhang@tu-darmstadt.de) and Jingcun Wang (jingcun.wang@tu-darmstadt.de) with your CV and transcripts.