COMPSCI 690AB SYSTEMS FOR DEEP LEARNING

Reading Assignment 5

Flash-LLM: Enabling Cost-Effective and Highly-Efficient Large Generative Model Inference with Unstructured Sparsity – Summary

Many generative models have proven to be effective in various tasks, but challenge arises when their size is exponentially increasing in deployment because of memory constraints and high inference latency. This seems to be a problem to fit models into GPU memory and in turn inference latency increases. To address this challenge the paper proposes Flash-LLM which is an efficient GPU library supporting model inference where its goal is to reduce memory footprint, improve efficiency and lower costs. Flash-LLM further introduces a Load-as-Sparse and Compute-as-Dense strategy which leverages sparse memory load. By making use of the capabilities of tensor cores and optimizing the Sparse Matrix-Matrix Multiplication, its possible to achieve inference for such large generative models. This approach reduces global memory footprint, addresses the memory access bottlenecks and helps with the redundant computations of skinny MatMuls. The paper also discusses the implementation and evaluation of Flash-LLM. It is implemented as a high-performance Flash-LLM kernel integrated with Faster Transformer. On evaluation, it is found that Flash-LLM outperforms Sputnik, cuSparse, SparTA, etc. Overall, it achieves significant speedups, proving its superiority in kernel-level performance. One limitation of the proposal could be that it mainly focuses only on supporting model pruning for achieving sparsity and this might require a high fine-tuning cost, in turn limiting its applicability in places where fine-tuning resources are limited. To improve this, there could be more future study to explore methods to reduce the fine-tuning cost or investigate alternative approaches to achieve sparsity without extensive retraining. One question that could be discussed is about how the proposed solution can be more generalized?

Srimathi Mahalingam