COMPSCI 690AB SYSTEMS FOR DEEP LEARNING

Reading Assignment 3

FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness – Summary

This paper mainly focuses on introducing FlashAttention which is an IO-aware attention algorithm which minimizes reads and writes to the memory, resulting in 7.6x speedup over standard attention. This is mainly come into picture because transformers are used in NLP and image classification but they struggle with longer contexts because of quadratic time and memory complexity of self-attention. Various existing attention methods often overlook memory access over heads. Efficiently handling long sequence is important as it allows the model to understand and perform complex tasks better. The paper proposes a solution, FlashAttention which divides input matrix into multiple blocks and performs attention computations incrementally and minimizes HBM accesses, leading to lower memory footprint and faster execution. Through experimentation results we find that FlashAttention outperforms MLPerf speed record for BERT by 15% and speeds up GPT2 by 3 times compared to HuggingFace and Megatron. Overall, it is the first Transformer model to demonstrate superior runtime and memory performance compared to other standard mechanisms. One major drawback in the paper proposal is there could be a need for a lot of manual effort and engineering expertise in order to implement FlashAttention efficiently. This could be overcome by focusing future studies on developing automated tools which integrates novel attention mechanism into Transformer architectures.

Question: Can this technique be applied to other deep learning architectures?

Srimathi Mahalingam