PMPH2025-Flash Attention

By
Wanjing Hu / Computer Science, KU
fng685@alumni.ku.dk

October 24, 2025

1 Problem Statement

Flash Attention proposed by Dao-AILab is an IO-aware exact attention algorithm that uses tiling to reduce the number of memory reads/writes between GPU high bandwidth memory (HBM) and GPU on-chip SRAM(Dao et al., 2022). In this report, we implement Flash Attention 1(FA1) forward calculation with different Domain Specific Languages(DSLs).

For **evaluation metrics**, we focus on Latency speed up across different DSLs, compute efficiency, memory efficiency and throughput.

Latency speedup is assessed using kernel runtime of the execution time (T) of the self-attention forward. Here we choose Baseline as TODO:

$$\text{Speedup} = \frac{T_{\text{Base}}}{T_{\text{FlashAttention}}}.$$

For compute efficiency we use achieved Tera Floating-Point Operations Per Second(TFLOPs):

$$\text{Achieved TFLOPs} = \frac{\text{Total Floating-Point operations in FA1}}{T_{\text{Execution Time}}}$$

.

We compare the achieved TFLOPs with the theoretical hardware TFLOPs to see if we are using the GPU compute resource efficiently.

For memory efficiency we access with achieved bandwidth. We use Nsight Compute to see the total runtime bandwidth.

$$\text{Achieved Bandwidth} = \frac{\text{Total runtime bandwidth in FA1}}{T_{\text{Execution Time}}}$$

.

For throughput we use maximum input batch size with fixed other parameters. This is a more request-oriented index and more end-to-end in Transformer-like model serving.

2 DSL XX by Your Name TODO

3 DSL XX by Your Name TODO

4 DSL XX by Your Name TODO

5 DSL XX by Your Name TODO

References

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.