Contribution ID: **246** Type: **Poster**

Optimizing Model Inference with FlashAttention, Linformer, and INT8 Quantization for Real-Time Jet Classification

In this work, we present a set of optimizations to the Particle Transformer (ParT), a state-of-the-art model for jet classification, aimed at reducing inference time and memory usage while preserving accuracy. To address the compute and memory bottlenecks of traditional attention mechanisms, we incorporate FlashAttention and memory-efficient attention, enabling exact attention computation through a fused kernel that minimizes read/write overhead and improves GPU utilization. We also integrate Linformer, which reduces attention complexity from $O(n^2)$ to $O(n)$ by projecting sequences into a lower-dimensional space, making ParT more scalable for longer inputs. Additionally, we apply INT8 dynamic quantization to compress matrix multiplications from fp32 to int8, reducing latency and GPU memory usage with minimal impact on performance and no retraining required. By systematically evaluating combinations—FlashAttention + Linformer + INT8 quantization, and the full stack—we demonstrate that our approach yields significant speedups and memory savings while maintaining model accuracy. This synergy enables efficient deployment of transformer models like ParT in real-time, fast paced environments such as those encountered in HL-LHC triggers.

Significance

This presentation goes beyond a status update by showcasing novel integration of FlashAttention, Linformer, and INT8 quantization in the Particle Transformer (ParT) for jet classification. It highlights the synergistic impact of these optimizations on reducing inference time and memory usage without sacrificing accuracy. By systematically evaluating their combined effects, the work provides practical insights for real-time deployment in HL-LHC triggers, marking a significant step toward production-ready transformer models in high-energy physics.

References

Interpreting and Accelerating Transformers for Jet Tagging (Talk at FastML)

Experiment context, if any

CMS

Authors: WANG, Aaron (University of Illinois Chicago (US)); GANDRAKOTA, Abhijith (Fermi National Accelerator Lab. (US)); KHODA, Elham (University of Washington (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); NGADIUBA, Jennifer (FNAL); SAHU, Vivekanand Gyanchand (University of California San Diego); ZHAO, Zihan (Univ. of California San Diego (US))

Presenter: WANG, Aaron (University of Illinois Chicago (US))

Session Classification: Poster session with coffee break

Track Classification: Track 1: Computing Technology for Physics Research