Contribution ID: **244** Type: **Poster**

Efficient Transformer Architectures for Jet Tagging

Particle Transformer has emerged as a leading model for jet tagging, but its quadratic scaling with sequence length presents significant computational challenges, especially for longer sequences. This inefficiency is critical in applications such as the LHC trigger systems where rapid inference is essential. To overcome these limitations, we evaluated several Transformer variants and identified the Linformer as a very promising alternative. Our tests on both small and large models using the JetClass and HLS4ML datasets show that the Linformer dramatically reduces inference time and computational demands measured in FLOPs while nearly matching the performance of the Particle Transformer. We also examined the impact of the input sequence order by testing various strategies, including those based on physics motivated projection matrices, to further improve performance. Finally, we employed interpretability methods such as analyzing the attention matrices and examining the embeddings to gain deeper insights into the model operation.

Significance

Our work presents novel results by demonstrating that the Linformer significantly reduces inference time and computational demands compared to the Particle Transformer. These advances enable real-time jet tagging under the stringent latency constraints of both Level-1 Trigger (L1T) and High-Level Trigger (HLT) systems, marking crucial progress for deploying machine learning in high-energy physics experiments.

References

https://indico.cern.ch/event/1387540/contributions/6153602/attachments/2947435/5180167/Interpreting%20and%20Accelerating%20Trans

Wang, Aaron, et al. “Interpreting Transformers for Jet Tagging.” arXiv preprint arXiv:2412.03673 (2024). https://arxiv.org/abs/2412.03673

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))

Presenters: WANG, Aaron (University of Illinois Chicago (US)); SAHU, Vivekanand Gyanchand (University of California San Diego)

Session Classification: Poster session with coffee break

Track Classification: Track 2: Data Analysis - Algorithms and Tools