EPS-HEP 2025

Contribution ID: **632** Type: **Parallel**

Low-latency Jet Tagging for HL-LHC Using Transformer Architectures

Transformers are the state-of-the-art model architectures and widely used in application areas of machine learning. However the performance of such architectures is less well explored in the ultra-low latency domains where deployment on FPGAs or ASICs is required. Such domains include the trigger and data acquisition systems of the LHC experiments.

We present a transformer-based algorithm for jet tagging built with the HGQ2 framework, which is able to produce a model with heterogeneous bitwidths for fast inference on FPGAs, as required in the trigger systems at the LHC experiments. The bitwidths are acquired during training by minimizing the total bit operations as an additional parameter. By allowing a bitwidth of zero, the model is pruned in-situ during training. Using this quantization-aware approach, our algorithm achieves state-of-the-art performance while also retaining permutation invariance which is a key property for particle physics applications

Due to the strength of transformers in representation learning, our work serves also as a stepping stone for the development of a larger foundation model for trigger applications.

Secondary track

Authors: GANDRAKOTA, Abhijith (Fermilab); TAPPER, Alexander (Imperial College London); COX, Arianna (Imperial College London); MAIER, Benedikt (Imperial College London); SUN, Chang (Caltech); WOJCICKI, Filip (Imperial College London); NGADIUBA, Jennifer (Fermilab); LAATU, Lauri (Imperial College London); QUE, Zhiqiang (Imperial College London)

Presenter: LAATU, Lauri (Imperial College London)

Session Classification: Joint T12+T16

Track Classification: T16 - AI for HEP (special topic 2025)