Contribution ID: **108** Type: **Lightning 5 min talk + poster**

# Interpreting and Accelerating Transformers for Jet Tagging

*Tuesday 15 October 2024 15:25 (5 minutes)*

Attention-based transformers are ubiquitous in machine learning applications from natural language processing to computer vision. In high energy physics, one central application is to classify collimated particle showers in colliders based on the particle of origin, known as jet tagging. In this work, we study the interpretatbility and prospects for acceleration of Particle Transformer (ParT), a state-of-the-art model, leverages particle-level attention to improve jet-tagging performance. We analyzing ParT's attention maps and particle-pair correlations in the $\eta$-$\phi$ plane, revealing intriguing features, such as a binary attention pattern that identifies critical substructure in jets. These insights enhance our understanding of the model's internal workings and learning process and hint at ways to improve its efficiency. Along these lines, we also explore low-rank attention, attention alternatives, and dynamic quantization to accelerate transformers for jet tagging. With quantization, we achieve a 50% reduction in model size and a 10% increase in inference speed without compromising accuracy. These combined efforts enhance both the performance and the interpretability of transformers in high-energy physics, opening avenues for more efficient and physics-driven model designs.

## Focus areas

HEP

**Authors:** WANG, Aaron (University of Illinois at 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)

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

**Session Classification:** Lighting talks