Contribution ID: **69** Type: **Parallel talk**

Tracking Transformer synthesis for low-latency FPGA deployment

The Transformer Machine Learning architecture has been gaining considerable momentum in recent years. Computational High-Energy Physics tasks such as jet tagging and particle track reconstruction (tracking), have either achieved proper solutions, or reached considerable milestones using Transformers. On the other hand, the use of specialised hardware accelerators, especially FPGAs, is an effective method to achieve online, or pseudo-online latencies.

The development and integration of Transformer-based machine learning on FPGAs is still ongoing, and the support from current tools is very limited. Additionally, FPGA resources present a significant constraint. Considering the model size alone, while smaller models can be deployed directly, larger models are to be partitioned in a meaningful and ideally automated way. We aim to develop methodologies and tools for monolithic, or partitioned Transformer synthesis, specifically targeting inference. Our primary use-case involves machine learning models for tracking, derived from the TrackFormers project. We strive for lower latencies compared to GPU deployments.

AI keywords

FPGA deployment; Transformer synthesis; Inference latency

Primary authors: YOUSEFZADEH, Amir (University of Twente); BLANKESTIJN, Arjan (University of Twente); ODYURT, Uraz (University of Twente)

Presenter: ODYURT, Uraz (University of Twente)

Track Classification: Real-Time Data Processing