Contribution ID: **101**

Type: **Poster**

Ahead-of-time (AOT) compilation of Tensorflow models for deployment

Wednesday 13 March 2024 16:15 (30 minutes)

In a wide range of high-energy particle physics applications, machine learning methods have proven as powerful tools to enhance various aspects of physics data analysis. In the past years, various ML models were also integrated in central workflows of the CMS experiment, leading to great improvements in reconstruction and object identification efficiencies. However, the continuation of successful deployments might be limited in the future due to memory and processing time constraints of more advanced models evaluated on central infrastructure.

A novel inference approach for models trained with TensorFlow, based on Ahead-of-time (AOT) compilation is presented. This approach offers a substantial reduction in memory footprint while preserving or even improving computational performance. This talk outlines strategies and limitations of this novel approach, and presents integration workflow for deploying AOT models in production.

Significance

The continuation of successful ML model deployments might be limited in the future due to memory and processing time constraints, and this contribution presents a novel approach for inference on central infrastructure that can drastically reduce resource consumption.

References

Experiment context, if any

CMS

Primary authors: WIEDERSPAN, Bogdan (Hamburg University (DE)); RIEGER, Marcel (Hamburg University (DE))

Presenter: WIEDERSPAN, Bogdan (Hamburg University (DE))

Session Classification: Poster session with coffee break

Track Classification: Track 1: Computing Technology for Physics Research