Contribution ID: **174** Type: **Poster**

Zero-overhead ML training with ROOT in an ATLAS Open Data analysis

The ROOT software framework is widely used in HEP for storage, processing, analysis and visualization of large datasets. With the large increase in usage of ML for experiment workflows, especially lately in the last steps of the analysis pipeline, the matter of exposing ROOT data ergonomically to ML models becomes ever more pressing. In this contribution we discuss the experimental component of ROOT that exposes ROOT datasets in batches ready for the training phase. A new shuffling strategy for creating the batches to prevent biased training is discussed, taking as examples real-life use cases relative to ATLAS Open Data.
An end-to-end ML physics analysis is carried out to show how training a model with common ML tools can be done directly from ROOT datasets to avoid intermediate data conversions, streamline workflows and used in the case where the training data does not fit in memory. Datasets from ATLAS Open Data are used as input to analyses searching for the Higgs boson or new BSM particles such as supersymmetric particles. The datasets are stored in the new on-disk ROOT format called RNTuple.

Significance

This presentation covers a new shuffling strategy in the experimental component of ROOT that exposes ROOT datasets in batches ready for the training phase to prevent biased training when used with common ML tools. This enabled ML training to be done directly from ROOT datasets avoiding the need for intermediate data conversions, streamlining workflows and used in the case where the training data does not fit in memory. In this contribution an end-to-end physics analysis is carried out to show how it can be used when training a model with common ML tools with ATLAS Open Data as input datasets stored in the new on-disk ROOT format called RNTuple.

References

CHEP 2024: Zero-overhead training of machine learning models with ROOT data
https://indico.cern.ch/event/1338689/contributions/6015940/

Experiment context, if any

ATLAS

Authors: FOLL, Martin (University of Oslo (NO)); Dr PADULANO, Vincenzo Eduardo (CERN); PIPARO, Danilo (CERN); Prof. OULD-SAADA, Farid (University of Oslo (NO)); Dr GRAMSTAD, Eirik (University of Oslo (NO)); CATMORE, James (University of Oslo (NO))

Presenter: FOLL, Martin (University of Oslo (NO))

Session Classification: Poster session with coffee break

Track Classification: Track 1: Computing Technology for Physics Research