Contribution ID: **150** Type: **Poster**

Machine Learning Pipelines for HEP Using Big Data Tools Applied to Improving Event Filtering

Tuesday, November 5, 2019 4:15 PM (15 minutes)

This work addresses key technological challenges in the preparation of data pipelines for machine learning and deep learning at scale of interest for HEP. A novel prototype to improve the event filtering system at LHC experiments, based on a classifier trained using deep neural networks has recently been proposed by T. Nguyen et al. https://arxiv.org/abs/1807.00083. This presentation covers how we implemented the data pipeline to train the neural network classifier using solutions from the Apache Spark and Big Data ecosystem, integrated with tools, software, and platforms common in the HEP environment. Data preparation and feature engineering make use of PySpark, Spark SQL and Python code run via Jupyter notebooks. We will discuss key integrations and libraries that make Apache Spark able to ingest data stored using ROOT and its integration EOS/XRootD protocol. The presentation will cover the neural network models used, defined using the Keras API, and how the models have been trained in a distributed fashion on Spark clusters using BigDL and Analytics Zoo. We will discuss the implementation, the results of the distributed training, and overall the lessons learned on using Big Data tools to implement an end-to-end ML pipeline.

Consider for promotion

No

Primary authors: ZANETTI, Marco (Universita e INFN, Padova (IT)); MIGLIORINI, Matteo (Universita e INFN, Padova (IT)); CANALI, Luca (CERN)

Presenter: ZANETTI, Marco (Universita e INFN, Padova (IT))

Session Classification: Posters

Track Classification: Track 6 – Physics Analysis