Contribution ID: **53** Type: **Oral Presentation**

Machine Learning as a Service for High Energy Physics on heterogeneous computing resources

Thursday, 25 March 2021 13:00 (30 minutes)

Machine Learning (ML) techniques in the High-Energy Physics (HEP) domain are ubiquitous and will play a significant role also in the upcoming High-Luminosity LHC (HL-LHC) upgrade foreseen at CERN: a huge amount of data will be produced by LHC and collected by the experiments, facing challenges at the exascale. Despite ML models are successfully applied in many use-cases (online and offline reconstruction, particle identification, detector simulation, Monte Carlo generation, just to name a few) there is a constant seek for scalable, performant, and production-quality operations of ML-enabled workflows. In addition, the scenario is complicated by the gap among HEP physicists and ML experts, caused by the specificity of some parts of the HEP typical workflows and solutions, and by the difficulty to formulate HEP problems in a way that match the skills of the Computer Science (CS) and ML community and hence its potential ability to step in and help. Among other factors, one of the technical obstacles resides in the difference of data-formats used by ML-practitioners and physicists, where the former use mostly flat-format data representations while the latter use to store data in tree-based objects via the ROOT data format. Another obstacle to further development of ML techniques in HEP resides in the difficulty to secure the adequate computing resources for training and inference of ML models, in a scalable and transparent way in terms of CPU vs GPU vs TPU vs other resources, as well as local vs cloud resources. This yields a technical barrier that prevents a relatively large portion of HEP physicists from fully accessing the potential of ML-enabled systems for scientific research.

In order to close this gap, a Machine Learning as a Service for HEP (MLaaS4HEP) solution is presented as a product of R&D activities within the CMS experiment. It offers a service that is capable to directly read ROOT-based data (both local and remote), apply the ML solution developed by the user, and ultimately serve predictions by pre-trained ML models "as a service" accessible via HTTP protocol.
This solution can be used by physicists or experts outside of HEP domain and it provides access to local or remote data storage without requiring any modification or integration with the experiment's specific framework. Moreover, MLaaS4HEP is built with a modular design allowing independent resource allocation that opens up a possibility to train ML models on PB-size datasets remotely accessible from the WLCG sites without physically downloading data into local storage.

To prove the feasibility and utility of the MLaaS4HEP service with large datasets and thus be ready for the next future when an increase of data produced is expected, an exploration of different hardware resources is required. In particular, this work aims to provide the MLaaS4HEP service transparent access to heterogeneous resources, which opens up the usage of more powerful resources without requiring any effort from the user side during the access and use phase. We show the comparison of performance using different kinds of resources, both local and remote, for typical physics use-cases, e.g. in signal vs background discrimination problems.

Primary author: Dr GIOMMI, Luca (INFN and University of Bologna)

Co-authors: Prof. BONACORSI, Daniele (University of Bologna); Dr KUZNETSOV, Valentin (Cornell University, Ithaca, USA)

Presenter: Dr GIOMMI, Luca (INFN and University of Bologna)

Session Classification: Infrastructure Clouds and Virtualisation Session

Track Classification: Infrastructure Clouds and Virtualisation