Contribution ID: **48** Type: **Oral Presentation**

An application-agnostic AI platform to accelerate Machine Learning adoption for basic to hard ML/DL scientific use cases. (Remote Presentation)

Tuesday, March 26, 2024 4:30 PM (30 minutes)

Researchers at **INFN** (National Institute for Nuclear Physics) face challenges from basic to hard science use cases (e.g., big-data latest generation experiments) in many areas: HEP (High Energy Physics), Astrophysics, Quantum Computing, Genomics, etc.

Machine Learning (ML) adoption is ubiquitous in these areas, requiring researchers to solve problems related to the specificity of applications (e.g., tailored models and intricate domain knowledge), but also requiring solving general infrastructure-level and ML-workflow related problems.

As the demand for ML solutions continues to rise across the diverse research domains, there exists a critical need for an innovative approach to accelerate ML adoption.

In this regard we propose an **AI platform** designed as an **application-agnostic MLaaS** (Machine Learning as a Service) solution, which provides a paradigm shift by offering a flexible and generalized infrastructure that decouples the ML development process from specific use cases.

The AI platform is implemented as a software layer on top of our cloud serviceplatform, the INFN Cloud, which offers composable, scalable, and open-source solutions on a dedicated geographically distributed infrastructure. The INFN Cloud core mission is to facilitate resource sharing and enhance accessibility for INFN users, encompassing a wide range of resources, including GPUs and storage.

The AI platform leverages INFN Cloud resources and principles, gathering and orchestrating technologies to support end-to-end scalable ML solutions: Kubernetes, Kubeflow, KServe, KNative, Kueue, Horovod, etc., ensuring support for many ML frameworks: TensorFlow, PyTorch, Apache MXNet, XGBoost, etc.

This contribution will describeTogether with the platform's design and principles, we presentas well as some selected use cases from NLP and HEP domains that benefitted from the "aaS" approach.

The platform's agnostic nature extends beyond model compatibility to address the practical challenges associated with deploying ML solutions in real hard science scenarios: streaming services, exabyte-scale storage solutions, high-bandwidth networking, support for native HEP data (e.g., CERN ROOT data format), etc.

Furthermore, the AI platform promotes transfer learning and model reuse, to accelerate the ML development lifecycle. Developers can leverage pre-trained models and share knowledge across different applications, reducing the time and resources required for training new models from scratch. This collaborative aspect not only enhances efficiency but also promotes a collective learning environment within the research community.

In conclusion, the application-agnostic AI platform serves as a unified ecosystem where developers, data scientists, and domain experts can collaborate seamlessly. By providing a standardized framework for ML model development, training, and deployment, the platform eliminates the need for extensive domain expertise in every application area. This democratization of ML empowers a broader audience to leverage the benefits of machine learning, breaking down barriers and fostering innovation across diverse research domains.

Primary authors: GATTARI, Mauro (INFN (National Institute for Nuclear Physics)); GIOMMI, Luca (INFN and University of Bologna); ANTONACCI, Marica (INFN); Mr VINO, Gioacchino (INFN)

Presenters: GATTARI, Mauro (INFN (National Institute for Nuclear Physics)); GIOMMI, Luca (INFN and University of Bologna); ANTONACCI, Marica (INFN); Mr VINO, Gioacchino (INFN)

Session Classification: Artificial Intelligence (AI)

Track Classification: Track 10: Artificial Intelligence (AI)