Contribution ID: **119** Type: **Poster Presentation**

Integration and deployment of Model Serving Framework to serve machine learning models at production scale in easy way.

Wednesday, 1 June 2022 17:13 (3 minutes)

With the help of machine learning systems, we can examine data, learn from that data and make decisions. Now machine learning projects have become more relevant to various use case, but too many models are difficult to manage. For this reason several MLOps tools were born. These tools are the main platforms, hosting the full machine learning process lifecycle, starting with data management and ending with model versioning and deployment.
NEANIAS (Novel EOSC Services for Emerging Atmosphere, Underwater & Space Challenges) is an ambitious project that comprehensively addresses the challenges set out in the 'Roadmap for EOSC'foreseen actions.
NEANIAS drives the co-design and delivery of innovative thematic services, derived from state-of-the-art research assets and practices in three major sectors: Underwater research, Atmospheric research and Space research.
The machine learning core services identified in the NEANIAS Project allow the scientists to define machine learning models and manage their lifecycle.
The Machine learning core services is composed by a JupyterHub instance (C3.1) that enable different profile server:
(C3.2) Serve machine learning model in production enviroment.
(C3.3) Do deep learning training framework using a Horovod cluster.
(C3.4) Perform distributed calculation using Apache Spark.
All these services combined and integrated are used for astrophysics use cases. In particular, two environments including two deep learning models tailored to object detection (based on maskRCNN) and semantic segmentation (based on Tiramisu) and trained on radio dataset from the Square Kilometre Array (SKA) precursors were integrated in C3.1 to classify radio astronomical sources, galaxies and image artefacts.
The "C3.2 - Model Serving"is a set of applications, part of Neanias ML core services, that is employed to simplify machine-learning model deployment and to run high-performance model serving at scale. This makes it easy to turn ML models into prediction services that are easily deployable on container platform. BentoML is the software identified to satisfy these goals. In the project described in this paper, It has been improved in the frame of the NEANIAS project to increase the interoperability of the applications.

"C3.2 - Model Serving"works on top of Kubernetes platform, provide REST API documented with Swagger and use GRPC proxy to guarantee authentication and authorization at the model saved and deployed.

Main Topic

Software tools and services for machine learning

Secondary Topic

Participation mode

In person

Primary authors: CARONTE, Francesco; Mr MESSINEO, Rosario (Altecspace)

Co-author: SCIACCA, Eva (Istituto Nazionale di Astrofisica (INAF))

Presenter: CARONTE, Francesco

Session Classification: Poster Session Day 3.2