Contribution ID: **7** Type: **Presentation (15' + 5' for questions)**

AI Model Inference Pipelines in AI4EOSC and iMagine with OSCAR and AI4Compose

Tuesday, 29 October 2024 12:00 (20 minutes)

AI models require extensive computing power to perform scalable inferences on distributed computing platforms to cope with increased workloads. This contribution summarises the work done in the AI4EOSC and iMagine projects to support AI model inference execution with OSCAR and AI4Compose. AI4EOSC delivers and enhanced set of services to create and manage the lifecycle of AI models (develop, train, share, serve) targeting use cases in automated thermography, agrometeorology and integrated plant protection. In turn, iMagine provides imaging data and services for aquatic science, and leverages the platform created in AI4EOSC.

On the one hand, OSCAR provides the serverless computing platform to run AI model inference on elastic Kubernetes clusters deployed on Cloud infrastructures. Several execution modes are supported to tackle different use cases (synchronous executions to achieve fast AI model inference with pre-provisioned infrastructure, scalable asynchronous executions, to execute multiple inference jobs on auto-scaled Kubernetes clusters and exposed services, to leverage pre-loaded in-memory AI models). Two production OSCAR clusters have been deployed in distributed Cloud sites (INCD and Walton) to support the different use cases.

On the other hand, AI4Compose allows users to visually design AI inference pipelines on Node-RED or Elyra. UPV operates a FlowFuse instance for users to deploy their own Node-RED instances on which they can visually craft the AI inference pipelines. For that, custom nodes have been created to facilitate the execution of the model inference steps in distributed OSCAR clusters for enhanced scalability. The AI models are available in the AI4EOSC Dashboard. These nodes have been contributed to the Node-RED library for enhanced outreach. AI4Compose also supports Elyra to create these AI inference pipelines from a Jupyter Notebook environment. The support has been integrated in EGI Notebooks a popular managed Jupyter Notebook service, thus facilitating adoption of these techniques beyond the project realm.

Together, they provide the ability to craft custom AI inference pipelines using custom nodes from a visual canvas which have been created to facilitate the usage of pre-trained models with the AI4OS platform, the distributed AI platform powering both AI4EOSC and iMagine.

Grant PID2020-113126RB-I00 funded by MICIU/AEI/10.13039/501100011033.

This work was supported by the project AI4EOSC "Artificial Intelligence for the European Open Science Cloud" that has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant 101058593.

This work was supported by the project iMagine "AI-based image data analysis tools for aquatic research"that has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant 101058625.

Primary authors: RODRÍGUEZ, Vicente (Universitat Politècnica de València); CALATRAVA ARROYO, Amanda (Universitat Politècnica de València); AGUIRRE, Diego A. (Universitat Politècnica de València); ALARCÓN, Caterina (Universitat Politècnica de València); LANGARITA, Sergio (Universitat Politècnica de València); MOLTÓ, Germán (Universitat Politècnica de València)

Presenter: MOLTÓ, Germán (Universitat Politècnica de València)

Session Classification: IBERGRID

Track Classification: Developments oriented to foster the Compute Continuum