Contribution ID: **3** Type: **Practical Demonstration (20')**

AI4Compose: low-code composition of AI inference pipelines

Tuesday, 26 September 2023 13:20 (20 minutes)

The composition of workflows using visual environments can significantly benefit AI scientists in leveraging the Function-as-a-Service (FaaS) paradigm for the execution of inference pipelines. With this goal, we have designed, in the context of the AI4EOSC project, AI4Compose [https://github.com/AI4EOSC/ai4-compose], an approach to perform low-code composition of AI inference pipelines. It leverages Node-RED [https://nodered.org/] and Elyra [https://elyra.readthedocs.io/en/latest/index.html], two widely used open-source tools for the graphical composition of pipelines, based on a drag-and-drop approach. On the one hand, Node-RED is a flow-based programming tool, originally developed by IBM's Emerging Technology Services team and now a part of the OpenJS Foundation. It is a powerful tool that serves to communicate hardware and services in a fast and easy way. On the other hand, Elyra is a set of AI-focused extensions for JupyterLab Notebooks. It provides a visual Notebook Pipeline editor to build notebook-based AI pipelines, simplifying the conversion of multiple notebooks into batch jobs or workflow.

The FaaS model enables scientists to efficiently manage application components, executed on-demand as functions. To exploit this model, AI4Compose is integrated with the OSCAR serverless framework [https://oscar.grycap.net/], to run the AI models for inference. OSCAR is an open-source platform to support the event-driven serverless computing model for data-processing applications that can run on top of multi-clouds thanks to the Infrastructure Manager (IM) [https://www.grycap.upv.es/im/index.php]. Its functionality flow is mainly based on the monitoring of an object storage solution where users upload files to a bucket and this automatically triggers the execution of parallel invocations to a function responsible for processing each file (asynchronous mode). It also supports synchronous invocations through highly scalable HTTP-based endpoints (based on KNative). The integration with OSCAR is made through flow implementations offered as reusable components inside both Node-RED and Elyra visual pipeline compositors.

With AI4Compose, users will gain agility and resource efficiency as they can leverage the management of the computing platform to OSCAR, which provides a highly scalable infrastructure to support complex computational tasks. Also, AI scientists can easily design, deploy and manage their workflows using an intuitive visual environment, reducing the time and effort required for the maintenance of inference pipelines. Lastly, our platform aims to lower the learning curve for researchers to implement AI and FaaS pipelines.

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. Also, Project PDC2021-120844-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR and Grant PID2020-113126RB-I00 funded by MCIN/AEI/10.13039/501100011033.

Primary authors: Mr RODRÍGUEZ, Vicente (Universitat Politècnica de València); Mr AGUIRRE, Diego (Universitat Politècnica de València); CALATRAVA, Amanda (Universitat Politècnica de València); Dr MOLTÓ, Germán (Universitat Politècnica de València)

Presenter: Mr RODRÍGUEZ, Vicente (Universitat Politècnica de València)

Session Classification: IBERGRID Contributions with Demonstrations

Track Classification: Development of innovative software and services