AI Inference Pipeline Composition with AI4Compose and OSCAR

Amanda Calatrava[a], Diego Aguirre[a], Vicente Rodríguez[a], Caterina Alarcón[a], Miguel Caballer[a] and Germán Moltó[a]

[a] Instituto de Instrumentación para Imagen Molecular (I3M), Universitat Politècnica de València, Valencia, Spain.
amcaar@i3m.upv.es, dieagra@i3m.upv.es, vrodben1@i3m.upv.es, calarcon@i3m.upv.es, micafer1@upv.es, gmolto@dsic.upv.es

OSCAR[1] is an open-source serverless framework to support the event-driven serverless computing model for data-processing applications. It can connect to an object storage solution where users upload files to trigger the execution of parallel invocations to a service responsible for processing each file. It also supports other flexible execution approaches such as programmatic synchronous invocations and exposing user-defined APIs for fast AI inference.

Serverless computing is very appropriate for the inference phase of the AI model lifecycle, as it offers several advantages such as automatic scalability and resource optimization, both at the level of costs and energy consumption. This model, in combination with the composition of workflows using visual environments, can significantly benefit AI scientists. With this objective, we have designed, in the context of the AI4EOSC project, AI4Compose[2], a framework responsible for supporting composite AI by allowing the workflow composition of multiple inference requests to different AI models. This solution relies on Node-RED[3] and Elyra[4], two widely adopted open-source tools for graphical pipeline composition, employing a user-friendly drag-and-drop approach. Node-RED, in combination with Flowfuse to support multitenancy, serves as a powerful graphical tool for rapid communication between different services; meanwhile, Elyra provides a visual Notebook Pipeline editor extension for JupyterLab Notebooks to build notebook-based AI pipelines, simplifying the conversion of multiple notebooks into batch jobs or workflows. The integration with OSCAR is made through flow and node implementations offered as reusable components inside both Node-RED and Elyra visual pipeline compositors.

During the session, we want to demonstrate how AI4Compose works, for both Node-RED and Elyra environments, making use of the Flowfuse instance of AI4EOSC and the EGI Notebooks service, empowered by the Elyra extension. We will present how to trigger the inference of AI models available in the AI4EOSC marketplace and compose the workflow graphically, demonstrating that, with AI4Compose, AI scientists can easily design, deploy, and manage workflows using an intuitive visual environment. This reduces the time and effort required for pipeline composition, while the AI model inference can be executed on remote OSCAR clusters running in the EGI Cloud.

---

[1] OSCAR: https://oscar.grycap.net/
[2] AI4Compose source code on Github: https://github.com/ai4os/ai4-compose
[3] Node-RED: https://nodered.org/
[4] Elyra: https://elyra.readthedocs.io/en/latest/index.html

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.