fast machine learning
for science

Contribution ID: **137**

Type: **Poster**

AutoDeploy-HEP: An Intelligent Toolkit for ML Deployment on Heterogeneous Hardware in High-Energy Physics

Deploying ML models today requires deep expertise in both hardware and software optimization. It often involves laborious trial-and-error to determine the right combination of tools, techniques, and configurations. While industry and academia benefit from a wide array of deployment frameworks and automation tools, the High-Energy Physics (HEP) community still faces major challenges in adopting and adapting these techniques. The fragmentation and volume of available methods make it nearly impossible for HEP engineers to explore and implement optimal solutions for each specific use case, leading to slow, non-scalable, and often suboptimal deployment workflows.

We propose **AutoDeploy-HEP**, an adaptive toolkit that bridges this gap by automating the end-to-end deployment of ML models across heterogeneous hardware platforms such as GPUs, FPGAs, NPUs, and DCUs. Given a model and deployment objectives via configuration, CLI, or API, the system uses a continually evolving **knowledge graph** to recommend efficient deployment pipelines composed of architecture optimization, model compression, and hardware acceleration steps. These pipelines are executed in a feedback-driven manner, adapting dynamically to intermediate results.

Over time, the toolkit evolves from rule-based logic to learning-driven intelligence using knowledge graph embeddings, ML-based recommenders, and inference cost predictors to optimize decisions before deployment. AutoDeploy-HEP is designed to support a wide range of HEP applications, across experiments and hardware setups, empowering researchers to deploy ML models efficiently without needing deep deployment expertise, and accelerating scientific innovation in the field.

Author: MUSTOFA, MUSTOFA ABDULHAFIZ AHMED

Presenter: MUSTOFA, MUSTOFA ABDULHAFIZ AHMED

Session Classification: Posters and coffee