Contribution ID: **31**   Type: **poster**

Model Logging for FPGA Deployment

Monday 14 October 2024 17:50 (5 minutes)

Deploying lightweight models on FPGAs requires robust workflows for tracking, saving, and transferring model information, and ensuring that this information adheres to FAIR (Findable, Accessible, Interoperable, and Reproducible) principles. We present a Python package that automates the identification and documentation of key metadata for machine learning models developed in PyTorch or TensorFlow. This tool captures the model architecture, Python environment, and system hardware details, integrating this information with the code, visualizations, and checkpoints. The entire package is then uploaded to DataFed, a federated scientific data management system chosen for its enforcement of FAIR principles, making model data easily discoverable and shareable across collaborations.

Additionally, provenance data is embedded to explicitly track the progression of each model, ensuring traceability from the original code through each checkpoint. This is particularly advantageous for small, fast models deployed on FPGAs, where iteration speed and accountability are critical. By automating these processes, the package ensures that FPGA-based machine learning systems remain efficient, reproducible, and optimized for performance, all while adhering to open science standards for data management and collaboration.

Author: GODDY, Julian

Co-author: Dr AGAR, Joshua (Drexel University)

Presenter: GODDY, Julian

Session Classification: Poster Session / Reception