Contribution ID: **611** Type: **Poster**

Optimized Fast Machine Learning Inference using TMVA SOFIE

While the development of machine learning models for analyzing physical processes—such as simulations, reconstruction, and triggers—has progressed rapidly, efficient inference remains a major challenge. Despite the availability of popular frameworks like TensorFlow and PyTorch for model development, training, and evaluation, experiments at CERN face difficulties during inference due to issues such as limited flexibility, integration complexity, heavy dependencies, and high latency during single-event evaluation.

Addressing these challenges, the ML4EP team at CERN has developed **SOFIE** (System for Optimized Fast Inference code Emit)—a tool that translates externally trained deep learning models in ONNX format, or those developed in Keras or PyTorch, into an intermediate representation, which is then used to generate highly optimized C++ code for fast inference. The generated code has only BLAS as an external dependency, making it easy to integrate into the data processing workflows of high-energy physics experiments.

SOFIE's IR can be stored in .root format, providing flexibility for storing and transporting large models as compressed files. SOFIE integrates with the ROOT ecosystem, offering a Python interface and support for multi-threaded evaluation via RDF slots. However, it does not depend explicitly on other ROOT libraries, enhancing its portability and ease of use.

SOFIE now supports a wide range of ML operators as defined by the ONNX standard, along with user-defined functions. It also enables inference for in-memory graph neural network models trained using DeepMind's Graph Nets.

We present the recent developments in SOFIE, including space optimizations through a custom memory allocator, operator fusion, kernel-level optimizations, and improvements in processing time—that significantly reduces inference latency.

Secondary track

T12 - Data Handling and Computing

Authors: SENGUPTA, Sanjiban (CERN, The University of Manchester); Dr MONETA, Lorenzo (CERN)

Presenter: SENGUPTA, Sanjiban (CERN, The University of Manchester)

Session Classification: T16

Track Classification: T16 - AI for HEP (special topic 2025)