Contribution ID: **111** Type: **Poster + Flashtalk**

Inference optimization with Memory Management and GPU Acceleration in TMVA SOFIE

Within ROOT/TMVA, we have developed SOFIE - System for Optimized Fast Inference code Emit - an engine designed to convert externally trained deep learning models—such as those in ONNX, Keras, or PyTorch formats—into optimized C++ code for fast inference. The generated code features minimal dependencies, ensuring seamless integration into the data processing and analysis workflows of high-energy physics experiments.
SOFIE now supports a comprehensive range of machine learning operators as defined by the ONNX standard, and also supports the translation and inference of Graph Neural Networks trained in DeepMind's Graph Nets. Recent advancements in SOFIE include memory optimizations that enable efficient reuse of in- termediate tensor data during inference, significantly reducing memory overhead. Additionally, SOFIE now incorporates enhanced GPU acceleration, supporting stacks such as SYCL, which have abstractions over platforms like CUDA and ROCm. These improvements result in a runtime- efficient and user-friendly machine learning inference engine, competitive with other state-of-the- art solutions.
This work highlights the latest developments in SOFIE, focusing on its memory optimization ca- pabilities and GPU acceleration enhancements, which collectively deliver efficient inference per- formance for HEP applications.

AI keywords

Fast ML Inference; ML Software;‘ Next Generatiuon Trigger Project; GPU

Primary author: SENGUPTA, Sanjiban (CERN)

Co-author: MONETA, Lorenzo (CERN)

Presenter: MONETA, Lorenzo (CERN)

Track Classification: Real-Time Data Processing