



Contribution ID: 210

Type: **Flashtalk with Poster**

Fast Inference of Deep Learning Models with SOFIE

Tuesday, April 30, 2024 2:53 PM (3 minutes)

Machine learning, especially Deep Learning, has become a valuable tool for researchers in High Energy Physics (HEP) to process and analyse their data. Popular Python-based machine learning libraries, such as Keras and PyTorch, offer good solutions for training deep learning models also in CPU or GPU environments. However, they do not always provide a good solution for inference. They may only support their own models, often provide only a Python API, or be constrained by heavy dependencies.

To solve this problem, we have developed a tool called SOFIE, within the ROOT/TMVA project. SOFIE takes externally trained deep learning models in ONNX format or Keras and PyTorch native formats and generates C++ code that can be easily included and invoked for fast inference of the model. The code has a minimal dependency and can be easily integrated into the data processing and analysis workflows of the HEP experiments.

We will present the latest developments in SOFIE, which include the support for parsing Graph Neural Networks trained with the Python Graph Net library, as well as the support for ONNX operators needed for transformer models.

We will also show how SOFIE can be used to generate code for accelerators, such as GPU using SYCL, in order to achieve optimal performance in large model evaluations. We will present benchmarks and comparisons with other existing inference tools, such as ONNXRuntime, using deep learning models used by the LHC experiments.

Primary author: MONETA, Lorenzo (CERN)

Co-authors: PANAGOI, Ioanna; SENGUPTA, Sanjiban

Presenter: MONETA, Lorenzo (CERN)

Session Classification: 2.4 Hardware acceleration & FPGAs

Track Classification: Session A