Contribution ID: **1379** Type: **Parallel Talk**

SOFIE: C++ Code Generation for Fast Deep Learning Inference

Friday, 8 July 2022 17:15 (15 minutes)

Deep neural networks are rapidly gaining popularity in physics research. While python-based deep learning frameworks for training models in GPU environments develop and mature, a good solution that allows easy integration of inference of trained models into conventional C++ and CPU-based scientific computing workflow seems lacking.

We report the latest development in ROOT/TMVA that aims to address this problem. This new framework takes externally trained deep learning models in ONNX format or Keras and PyTorch native formats, and emits C++ code that can be easily included and invoked for fast inference of the model, with minimal dependency on linear algebra libraries only. We provide an overview of this current solution for conducting inference in C++ production environment and discuss the technical details.

More importantly, we present the latest and significant updates of this framework in supporting commonly used deep learning architectures, such as convolutional and recurrent networks.

We show also how the inference code can be integrated in the analysis users code together with tools such as ROOT RDataFrame.
We demonstrate its current capabilities with benchmarks in evaluating popular deep learning models used by LHC experiments against popular deep learning inference tools, such as ONNXRuntime.

In-person participation

Yes

Primary authors: MONETA, Lorenzo (CERN); AN, Sitong (CMU); GUIRAUD, Enrico (EP-SFT, CERN)

Presenter: GUIRAUD, Enrico (EP-SFT, CERN)

Session Classification: Computing and Data handling

Track Classification: Computing and Data handling