ACAT 2021

Contribution ID: **732** Contribution code: **contribution ID 732** Type: **Poster**

SOFIE: C++ Code Generation from ROOT/TMVA for Fast Deep Learning Inference

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 with examples of the generated code.

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, as well as a new capability to store deep learning models in ROOT format.

Furthermore, we demonstrate its current capabilities with benchmarks in evaluating popular deep learning models like resnet against popular deep learning inference tools, such as ONNXRuntime.

Significance

This is a significant and extensive update for SOFIE that vastly expands our supported NN models and functionalities. It is an important step that promises a practical solution for physicists looking to integrate their NN into their C++ workflow.

References

Previous CHEP 2021 proceeding that covers only basic NN and benchmark:
https://www.epj-conferences.org/articles/epjconf/abs/2021/05/epjconf_chep2021_03040/epjconf_chep2021_03040.html

Speaker time zone

Compatible with America

Primary authors: AN, Sitong (CERN, Carnegie Mellon University (US)); MONETA, Lorenzo (CERN); SENGUPTA, Sanjiban (IIIT Bhubaneswar); HAMDAN, Ahmat (ISSEA, Cameroon); SOSSAI, Federico (University of Padova); SAXENA, Aaradhya (Indian Institute of Technology, Roorkee)

Presenter: AN, Sitong (CERN, Carnegie Mellon University (US))

Session Classification: Posters: Walnut

Track Classification: Track 1: Computing Technology for Physics Research