Contribution ID: **95** Type: **Oral**

Accelerating Machine Learning Inference on GPUs with SYCL using SOFIE

Tuesday, March 12, 2024 11:50 AM (20 minutes)

Recently, machine learning has established itself as a valuable tool for researchers to analyze their data and draw conclusions in vari- ous scientific fields, such as High Energy Physics (HEP). Commonly used machine learning libraries, such as Keras and PyTorch, might provide functionality for inference, but they only support their own models, are constrained by heavy dependencies and often provide only a Python API and not a C++ one. SOFIE [13], which stands for System for Optimized Fast Inference code Emit, a part of the ROOT project developed at CERN, creates standalone C++ inference code from an input model in one of the popular machine learning formats. This code is directly invokable from other C++ projects and has minimal dependencies. We will present the new developments of SOFIE extending the functionality to generate SYCL code for machine learning model inference that can run on various GPU platforms and is only dependent on Intel MKL BLAS and portBLAS libraries, achieving a speedup of up to x258 over plain C++ code for large convolutional models.

Significance

This presentation covers new results coming from new developments that happened last year

References

Experiment context, if any

Work happening within the ROOT project (CERN EP/SFT) and in collaboration with CERN Openlab

Authors: PANAGOU, Ioanna Maria; MONETA, Lorenzo (CERN); SENGUPTA, SANJIBAN; Dr PADULANO, Vincenzo (CERN)

Presenter: Dr PADULANO, Vincenzo (CERN)

Session Classification: Track 1: Computing Technology for Physics Research

Track Classification: Track 1: Computing Technology for Physics Research