Contribution ID: **118** Type: **Oral Presentation**

Recent Developments in hls4ml

Wednesday 3 August 2022 14:55 (20 minutes)

Neural Network (NN)-based inference deployed in FPGAs or ASICs is a powerful tool for real-time data processing and reduction. FPGAs or ASICs may be needed to meet difficult latency or power efficiency requirements in data acquisition or control systems. Alternately one may need to design ASICs for special conditions, like use in high radiation areas. The software package, hls4ml, was designed to make deploying optimized NNs on FPGAs and ASICs accessible for domain applications. In this talk, we will present the package capabilities, give updates on the current status, and give examples of its application.

The hls4ml's internal structure has been reorganized to allow for more flexibility to expand our backend support. Deploying on Intel FPGAs is now supported in the main branch, as is support for accelerators. We are continuously improving our NN support, such as adding new CNN implementations. We are integrating support for LSTM and GRU networks. We have also collaborated with the AMD/Xilinx FINN group to develop a Quantized ONNX (QONNX) representation to serve as a common way to to represent quantized NNs. Through common tools we provide, both Brevitas and QKeras can produce QONNX, which can then be ingested by both FINN and hls4ml. Usage of hls4ml continues to grow. Demonstrator systems for accelerator control are being built at Fermilab, and it is being studied for use by DUNE to detect supernova neutrino bursts. Together with the FINN team, we have provided solutions for MLPerf Tiny benchmarking results.

Minioral

Yes

IEEE Member

Yes

Are you a student?

No

Primary author: MITREVSKI, Jovan (Fermi National Accelerator Lab. (US))

Presenter: MITREVSKI, Jovan (Fermi National Accelerator Lab. (US))

Session Classification: Deep Learning and Machine Learning

Track Classification: Deep Learning and Machine Learning