Contribution ID: **40** Type: **not specified**

# Quantized Neural Networks on FPGAs using HAWQ-V3 and hl4ml

*Wednesday, 5 October 2022 14:45 (15 minutes)*

Neural networks have been shown to be helpful in identifying events of interest in particle physics. However, to be used for live trigger decisions, they must meet demandingly low latencies and resource utilization for deployment on Field Programmable Gate Arrays (FPGAs). HAWQ-V3, a Hessian-based quantization-aware training framework, and hls4ml, an FPGA firmware implementation package, address these issues. HAWQ-V3 is a training framework enabling ultra-low and mixed-precision quantization. It introduced an approach to determining the relative quantization precision of each layer based on the layer's Hessian spectrum. More recently, it implements a computational graph with only integer addition, multiplication, and bit-shifting. We present a neural network classifier implemented with HAWQ-V3 for high-pT jets from simulations of LHC proton-proton collisions. We then introduce an extension for HAWQ-V3 to translate our classifier into the Quantized ONNX (QONNX) intermediate representation format, an extension of the Open Neural Network Exchange (ONNX) format, supporting arbitrary-precision and low-precision neural networks. We demonstrate how the conversion of HAWQ-V3 models leverages the PyTorch Just-in-Time compiler to trace and translate models to QONNX operators. We then proceed to hls4ml to create firmware implementation of our quantized neural network and review its estimated latency and resource utilization for an FPGA.

**Primary author:** CAMPOS, Javier Ignacio (Fermi National Accelerator Lab. (US))

**Presenter:** CAMPOS, Javier Ignacio (Fermi National Accelerator Lab. (US))

**Session Classification:** Contributed Talks