Contribution ID: **160** Type: **Oral presentation**

End-to-end Codesign of Hessian-aware Quantized Neural Networks for FPGAs

Wednesday 24 April 2024 09:00 (40 minutes)

We develop an end-to-end workflow for the training and implementation of co-designed neural networks (NNs) for efficient field-programmable gate array (FPGA). Our approach leverages Hessian-aware quantization (HAWQ) of NNs, the Quantized Open Neural Network Exchange (QONNX) intermediate representation, and the hls4ml tool flow for transpiling NNs into FPGA firmware. This makes efficient NN implementations in hardware accessible to nonexperts, in a single open-sourced workflow that can be deployed for real-time machine learning applications in a wide range of scientific and industrial settings. We demonstrate the workflow in a particle physics application involving trigger decisions that must operate at the 40 MHz collision rate of the CERN Large Hadron Collider (LHC). Given the high collision rate, all data processing must be implemented on custom ASIC and FPGA hardware within a strict area and latency. Based on these constraints, we implement an optimized mixed-precision NN classifier for high-momentum particle jets in simulated LHC proton-proton collisions. In addition, we use a second particle physics example of lossy data compression with an autoencoder for the High-Granularity Endcap Calorimeter subdetector. We report on these two NNs with our end-to-end codesign workflow.

Minioral

Yes

IEEE Member

No

Are you a student?

No

Authors: CAMPOS, Javier; Dr TRAN, Nhan (Fermilab)

Co-authors: Dr GHOLAMI, Amir (University of California, Berkeley); FLUMERFELT, Eric Lewis (Fermi National Accelerator Lab. (US)); Dr DUARTE, Javier (University of California, San Diego); MITREVSKI, Jovan (Fermi National Accelerator Lab. (US)); Dr MAHONEY, Michael (University of California, Berkeley); DONG, Zhen (University of California, Berkeley)

Presenter: LONCAR, Vladimir (Massachusetts Inst. of Technology (US))

Session Classification: Invited Talk, Oral presentations, Mini-Orals

Track Classification: AI, Machine Learning, Real Time Simulation, Intelligent Signal Processing