Contribution ID: **31** Type: **Oral**

Exploration of Resource-efficient Implementations of ML Models Targeting eFPGAs

Wednesday, 8 November 2023 13:50 (15 minutes)

Implementing machine learning (ML) models in hardware has received considerable interest over the last several years from the physics community. The Python packages, hls4ml and conifer, has enabled porting models trained using Python ML libraries to register transfer level (RTL) code. Most of the attention, thus far, has been focused on porting ML models to commercial FPGAs or synthesized blocks on ASICs. With the latter, a (physical) area-optimized implementation of a ML model can be integrated on-chip. The reduction in area generally results in reduced costs for chip fabrication. The usual trade-off with an ASIC implementation is the inability to update model architecture post-synthesis. However, updating of biases/weights has been demonstrated at an additional area cost with techniques such as distributed I2C networks. Regardless, recent developments in open-source embedded FPGA (eFPGA) frameworks now provide an alternate and more flexible pathway for implementing ML models in hardware: customized eFPGA fabrics, which can also be integrated as part of an overall chip design. In general, the decision between an ASIC or eFPGA ML implementation will depend on the target application. We explored the design parameter space for eFPGA implementations of fully connected neural network (fc-NN) and boosted decision tree (BDT) models using the classification task of neutron/gamma identification, with a specific focus on resource efficiency. We used training data from an AmBe sealed source incident on a plastic scintillator read out by SiPMs. We studied relevant input features, the required bit-resolution, sampling rate and trade-offs in hyperparameters for both ML models while tracking resource usage and neutron efficiency at a gamma leakage of 10^-3. The results of the study will be used in the specification of an eFPGA fabric, which will be integrated as part of a 130 nm test chip next year.

Early Career

Yes

Primary author: JOHNSON, Jyothisraj (Lawrence Berkeley National Laboratory (LBNL))

Co-authors: Dr BOXER, Billy (UC Davis); Dr GRACE, Carl (LBNL); Dr PRAKASH, Tarun (LBNL)

Presenter: JOHNSON, Jyothisraj (Lawrence Berkeley National Laboratory (LBNL))

Session Classification: RDC4

Track Classification: RDC Parallel Sessions: RDC4: Readout and ASICs