Contribution ID: **28** Type: **not specified**

Large CNN for HLS4ML and Deepcalo

Wednesday, 5 October 2022 14:05 (5 minutes)

Convolutional neural networks (CNN) have been widely applied in a tremendous of applications that involve image processing, including particle physics. Deepcalo is a package designed for developing CNNs using ATLAS data at CERN, targeting tasks like energy regression of electrons and photons. Although it has been shown that CNNs used in Deepcalo can handle the task smoothly, the extensive computation resources and high-power consumption lead it hard to perform real-time inference during the experiment. As a result, it is limited in software simulation usage.

To accelerate the inference time and lower the power consumption, we implement those CNNs on FPGAs (Field Programmable Gate Arrays) with HLS4ML. HLS4ML is an automated tool for deploying machine-learning models on FPGAs, targeting ultra-low latency using fully-on-chip architecture. Based on HLS C++ codes by Dr. Dylan Rankin, we extend the HLS4ML library for supporting an automatic large CNNs conversion. In this work, we introduce a deeply-optimized workflow for implementing large CNNs on FPGAs. Implemented on an AlveoU50 FPGA running at 200 MHz, the accelerator infers with 0.039 of IQR75 loss in 0.6 ms.

Primary authors: SCHUY, Alexander Joseph (University of Washington (US)); LAI, Bo-Cheng; CHEN, Chi-Jui; RANKIN, Dylan Sheldon (Massachusetts Inst. of Technology (US)); YANG, Lin-Chi; HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); HAUCK, Scott; HSU, Shih-Chieh (University of Washington Seattle (US)); HUANG, Yan-Lun; YIN, Ziang (University of Washington (US))

Presenters: CHEN, ChiJui; YANG, Lin-Chi; HUANG, Yan-Lun

Session Classification: Contributed Talks