Contribution ID: **23**

Type: **not specified**

Fast recurrent neural networks on FPGAs with hls4ml

Tuesday, 4 October 2022 14:45 (15 minutes)

Recurrent neural networks have been shown to be effective architectures for many tasks in high energy physics, and thus have been widely adopted. Their use in low-latency environments has, however, been limited as a result of the difficulties of implementing recurrent architectures on field-programmable gate arrays (FPGAs). In this paper we present an implementation of two types of recurrent neural network layers- long short-term memory and gated recurrent unit- within the hls4ml [1] framework. We demonstrate that our implementation is capable of producing effective designs for both small and large models, and can be customized to meet specific design requirements for inference latencies and FPGA resources. We show the performance and synthesized designs for multiple neural networks, many of which are trained specifically for jet identification tasks at the CERN Large Hadron Collider.

[1] J. Duarte et al., "Fast inference of deep neural networks in FPGAs for particle physics", JINST 13 (2018) P07027, arXiv:1804.06913

Primary authors: WANG, Aaron; VERNIERI, Caterina (SLAC National Accelerator Laboratory (US)); Mr PAIKARA, Chaitanya (University of Washington); RANKIN, Dylan Sheldon (Massachusetts Inst. of Technology (US)); KHODA, Elham E (University of Washington (US)); KAGAN, Michael Aaron (SLAC National Accelerator Laboratory (US)); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); TEIXEIRA DE LIMA, Rafael (SLAC National Accelerator Laboratory (US)); Ms RAO, Richa (University of Washington); HAUCK, Scott; HSU, Shih-Chieh (University of Washington Seattle (US)); SUMMERS, Sioni Paris (CERN); LONCAR, Vladimir

Presenter: KHODA, Elham E (University of Washington (US))

Session Classification: Contributed Talks