Contribution ID: **411** Type: **Poster**

Ultra-low latency recurrent neural network inference on FPGAs for physics applications with hls4ml

Thursday 27 October 2022 16:10 (30 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

Experiment context, if any

References

https://arxiv.org/abs/2207.00559

Significance

RNNs have shown substantial success for many tasks in particle physics. They are particularly well-suited to those problems involving sequences of particle or detector signals, outperforming densely connected deep neural networks (DNNs) and convolution neural networks (CNNs) on certain jet classification tasks. In spite of this success, RNNs have not seen widespread adoption in ultra-low latency environments in physics when compared to DNNs and CNNs. This difference is owed in part to tools such as hls4ml that simplify the adaptation of the latter models from Keras to HLS. The support for GRUs and LSTMs in hls4ml that we present in this work represents the removal of a major barrier to the use of RNNs in ultra-low latency environments. This has ramifications not only for high energy physics but also other research areas where RNNs have become popular. While we have focused on the usage of hls4ml with FPGAs, it is important to note that hls4ml can also be used to create ASIC designs, and thus this work also allows for the possibility of RNN usage on ASICs as well. The recurrent or repeating nature of many modern algorithms, such as RNNs, transformers and graph neural networks, make them very difficult to be run, particularly at low latency, on FPGAs. In this work, we present the successful deployment of RNNs in models with number of trainable parameters ranging from O(1 k) to O(100 k) achieving latencies of O(1 s) to O(100s). This represents an important step in enabling support in hls4ml for more complex architectures with recursive computations.

Primary authors: WANG, Aaron; VERNIERI, Caterina (SLAC National Accelerator Laboratory (US)); 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)); RAO, Richa (University of Washington); HAUCK, Scott; HSU, Shih-Chieh (University of Washington Seattle (US)); SUMMERS, Sioni Paris (CERN); LONCAR, Vladimir (CERN)

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

Session Classification: Poster session with coffee break