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

Transformer in HLS4ML

Friday 7 October 2022 13:08 (3 minutes)

The transformer become widely use for the Natural language processing (NLP) task. Besides the NLP tasks, the transformer could be used to analyze any other time series signal processing data, such as images prediction, sensor detection, or glitch detection.
However, most of studies on transformer were implemented on GPU. The implementation of transformer on the FPGA was quite undiscovered. In this poster, we present an implementation of transformer on FPGA within the hls4ml framework.
We demonstrate how the transformer block, especially the multi-head attention layer, would be implemented on FPGA. We show a small transformer model as a benchmark to indicate the latency and resource usage on FPGA. Eventually, we discuss our next steps and expectation of the transformer on FPGA.

Authors: JIANG, Zhixing "Ethan" (University of Washington (US)); KHODA, Elham E (University of Washington (US))

Co-authors: HAUCK, Scott; HSU, Shih-Chieh (University of Washington Seattle (US))

Presenters: JIANG, Zhixing "Ethan" (University of Washington (US)); KHODA, Elham E (University of Washington (US))

Session Classification: Lightening talks