Contribution ID: **4** Type: **Talk**

A Quartus backend for hls4ml: deploying low-latency Neural Networks on Intel FPGAs

Monday, 30 November 2020 15:38 (6 minutes)

We describe the new Quartus backend of hls4ml, designed to deploy Neural Networks on Intel FPGAs. We list the supported network components and layer architectures (dense, binary/ternary, and convolutional neural networks) and evaluate its performance on a benchmark problem previously considered to develop the Vivado backend of hls4ml. We also introduce the support for recurrent layers and introduce a new asynchronous calling model to increase performance for larger models. In addition to that, we also demonstrate the use of this new model to optimize large-sparse networks.

Primary author: JAVED, Hamza (Pakistan Institute of Engin. and (PK))

Presenter: JAVED, Hamza (Pakistan Institute of Engin. and (PK))