Contribution ID: **37** Type: **Lightning Talk**

Efficient sparse matrix multiplication in hls4ml

Pruning enhances neural network hardware efficiency by zeroing out weight magnitude. In order to take full advantage of pruning, efficient implementations of sparse matrix multiplication are required. The current hls4ml implementations of sparse matrix multiplication rely on either the built in high-level synthesis zero suppression operations or a coordinate list representation, which faces scalability issues with model size and reuse factor. These implementations, particularly the coordinate list representation, are limited by their need to have large amounts of fanouts within an FPGA or ASIC to ensure a fully flexible implementation. We introduce a new implementation that preserves coordinate information but avoids the large dedicated logic needed for fanouts through the use of a crossbar. We present results for FPGA implementations scanning the model sparsity and initiation intervals for multiple benchmark models in MLPerf Inference Benchmark for anomaly detection and image classification.

Author: HOANG, Duc Minh (MIT)

Co-author: HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

Presenter: HOANG, Duc Minh (MIT)

Session Classification: Contributed Talks

Track Classification: Contributed Talks