Contribution ID: **135** Type: **Lightning 5 min talk + poster**

EnsembleLUT: Scaling up LUT-based Neural Networks with Ensemble Learning

Wednesday 16 October 2024 15:30 (5 minutes)

Applications like high-energy physics and cybersecurity require extremely high throughput and low latency neural network (NN) inference. Lookup-table-based NNs address these constraints by implementing NNs purely as lookup tables (LUTs), achieving inference latency on the order of nanoseconds. Since LUTs are a fundamental FPGA building block, LUT-based NNs map to FPGAs easily. LogicNets (and its successors) form one such class of LUT-based NNs that target FPGAs, mapping neurons directly to LUTs to meet the low latency constraints with minimal resources. However, it is difficult to implement larger, more performant LUT-based NNs like LogicNets because LUT usage increases exponentially with respect to neuron fan-in (i.e., number of synapses $\times$ synapse bitwidth). A large LUT-based NN quickly runs out of LUTs on an FPGA, which is unideal. Our work EnsembleLUT addresses this issue by creating ensembles of smaller LUT-based NNs that scale linearly with respect to the number of models, achieving higher accuracy within the resource constraints of an FPGA. We demonstrate that EnsembleLUT improves the scalability of LUT-based NNs on various scientific machine learning benchmarks such as jet substructure classification and high-granularity endcap calorimeter data compression found at the LHC CMS experiment, reaching higher accuracy with fewer resources than the largest LogicNets.

Focus areas

Primary authors: WENG, Olivia; ANDRONIC, Marta (Imperial College London); ZUBERI, Danial (UC San Diego); CHEN, Jiaqing (Arizona State University); GENIESSE, Caleb (Lawrence Berkeley National Laboratory); CONSTANTINIDES, George (Imperial College London); DUARTE, Javier (UCSD); TRAN, Nhan (Fermi National Accelerator Lab. (US)); FRASER, Nicholas (AMD); KASTNER, Ryan

Presenter: WENG, Olivia

Session Classification: Lighting talks