Contribution ID: **125** Type: **Poster**

luiGNN - A Graph Neural Network Surrogate Model for hls4ml

Recent advancements in use of machine learning (ML) techniques on field-programmable gate arrays (FPGAs) have allowed for the implementation of embedded neural networks with extremely low latency. This is invaluable for particle detectors at the Large Hadron Collider, where latency and used area are strictly bounded. The hls4ml framework is a procedure that converts trained ML model software to a synthesis result that can be used on an FPGA. However, running the pipeline is a time-consuming procedure, and there is a strong risk of failure. In particular, it may not be possible to successfully convert a model into a synthesis result, or the resource consumption of the model may exceed the resources of the target FPGA. The task is to estimate the chance of success and resource consumption of a given model when passed through the hls4ml pipeline, without needing to run the pipeline. We introduce Latency/Utilization Inference GNN (luiGNN), a surrogate model which uses a graph neural network to represent input architectures in the form of a directed graph. This graph representation allows for a diverse set of model architectures to all be effectively handled by a surrogate model. The GNN allows for effective representation of heterogeneous data and gives a natural avenue for representing certain intricacies of input models, such as skip connections. By representing these models in graph form, the GNN is able to train on the structure of the graph itself, rather than on the specific order of the layers in the input data. In principle, this could allow for luiGNN to perform inferences on model architectures that had never been seen in training data. Tests on the wa-hls4ml benchmark dataset currently point to the GNN being able to match or outperform the predictive power of a standard multilayer perceptron surrogate model. This lends confidence to luiGNN's future potential in the case of strongly heterogeneous data and in inference on previously unseen model architecture.

Focus areas

Authors: HAWKS, Ben (Fermi National Accelerator Lab); PLOTNIKOV, Dennis (Johns Hopkins University (US)); Dr TAME-NARVAEZ, Karla (Fermilab National Accelerator Laboratory); TRAN, Nhan (Fermi National Accelerator Lab. (US))

Presenter: PLOTNIKOV, Dennis (Johns Hopkins University (US))