Nanosecond Jet Classification at LHC

Introduction

The Large Hadron Collider(LHC) at CERN will go thorugh an upgrade (HL-LHC) to increase the rate of proton collisions, allowing experiments to collect one order of magnitude more data. This will demand a more efficient real-time event filter and this study shows how to perform jet classification on field-programmable gate arrays (FPGA) within $O(100)$ ns. Through quantization-aware training (QAT) and efficient FPGA implementations, we show that nanosecond inference using complex architectures like graph neural networks are feasible at low resource-cost.

Neural Network Architechtures

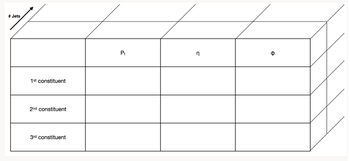
The Models are implemented using Keras and Tensorflow libraries. We compare the following architectures:

- A deep multilayer perceptron (MLP)
- A Graph Convolutional Network(GCN)
- An Interaction Network (IN)

For graph models the jet is represented as a fully connected graph, where each node is associated to a jet constituent and its features.

Dataset

In this study we analyze the public HLS jet dataset [1], consisting of jets from five different origins: quark (q), gluon (g), W boson, Z boson, and top (t) jets, with up to $N = 50$ jet particle constituents. We use a scenario of jets with $N = 8$ constituents, the expected average number at the L1T. We define the constituents features as P_t ,η and ϕ relative to the jet axis.



Full Precision Models Performance

Each model is first trained at floating-point precision, establishing the baseline performance that the quantized model on FPGA should match. The table bellow shows the models size and performance parameters.

Model	Param.	Oper.	Accuracy					AUC					FPR @30% TPR				
			q	g	W	Z	t	q	g	W	Z	t	q	g	W	Z	t
MLP	4,925	9840	0.82	0.86	0.84	0.85	0.90	0.82	0.86	0.87	0.86	0.91	0.043	0.022	0.032	0.023	0.011
GCN	5,045	11847	0.83	0.86	0.86	0.87	0.90	0.84	0.87	0.90	0.88	0.92	0.042	0.019	0.021	0.010	0.009
IN	4,289	92544	0.83	0.86	0.86	0.87	0.90	0.84	0.87	0.89	0.88	0.92	0.042	0.019	0.022	0.012	0.006

References:

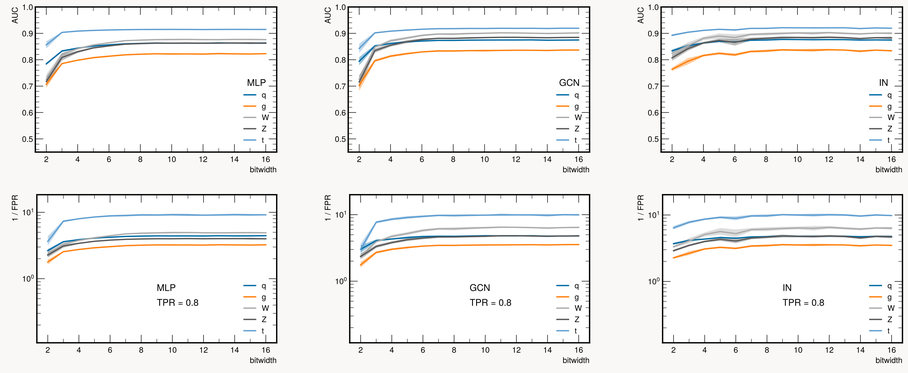
1) *HLS4ML LHC Jet Dataset* , https://doi.org/10.5281/zenodo.3601443 , J.M.Duarte and others, 2020

2) *QKeras* - https://github.com/google/qkeras , C.Coelho, 2019

3) *HLS4ML* - https://fastmachinelearning.org/hls4ml , 2018

Quantized Models Performance

Quantization aware training (QAT) of each model is implemented using the QKeras library [2]. The bit precision is scanned between 2 and 16 with a 1-bit step, when quantizing a model. The QAT performance obtained for each model as a function of the bit width is shown bellow.



Results

The models described above are translated into firmware using HLS4ML library [3], then synthesized with Vivado HLS 2020.1, targeting a Xilinx Virtex UltraScale+ VU9P (xcvu9p-flgb2104-2-e) FPGA with a clock frequency of 200 MHz.
A summary of the average accuracy ratio, resource consumption, and latency for the models is shown in table bellow. We find the FPGA resources are all bellow 16% of the total available, while latencies are less than about 250 ns for the MLP and GCN models and 505 ns for the IN. Moreover the initiation interval (II), which is the number of clock cycles between a new input, for the MLP and GCN are well within the required 150 ns, but IN have larger II making it too large for the L1T.

FPGA: Xilinx Virtex UltraScale+ VU9P

Model	Bits	Acc.Ratio	Latency[ns]	Latency[cc]	II [cc]	DSP	LUT	FF	BRAM
MLP	8	0.97	135	27	1	194 (3.5%)	85903 (12.9%)	36772 (2.8%)	2.0 (0.1%)
GCN	8	0.97	250	50	16	63 (1.1%)	67001 (10.1%)	32360 (2.4%)	2.0 (0.1%)
IN	8	0.99	505	101	67	213 (3.9%)	104885 (15.8%)	53291 (4.0%)	2.0 (0.1%)

Conclusions

We show the feasibility of a particle-based jet tagging algorithm for L1T and how to deploy graph neural classifiers on a FPGA using the HLS4ML library. Using QAT, we can limit resource utilization while retaining accuracy and inference time within a $O(100)$ ns latency, well within the typical time of a collision-event processing. The MLP and GCN architectures are suitable for the L1T , while IN consumes too much of the resources and have larger latencies.
An algorithm of this kind could improve the quality of the trigger decision, reducing false positive at little true-positive cost, increasing the scientific reach of the experiments.

Andre Sznajder in collaboratio with:

andre.sznajder@cern.ch

UERJ (Brazil)

Thea Aarrestad, Vladimir Loncar, Maurizio Pierini, Arpita Seksaria, and Sioni Summers
European Organization for Nuclear Research (CERN) (Geneva, Switzerland)

Jennifer Ngadiuba
Fermi National Accelerator Laboratory (FNAL) (Batavia, IL, USA)

Javier Duarte, Sukanya Krishna
University of California San Diego (La Jolla, CA, USA)

Johannes Haller, Gregor Kasieczka, Ihor Komarov, Finn Labe, Artur Lobanov, Philipp Rincke, Matthias Schroeder
Universität Hamburg (Hamburg, Germany)