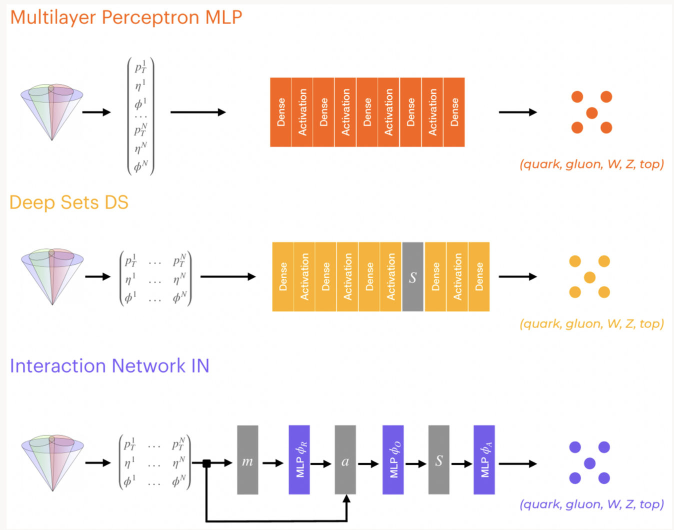
Sets are All you Need: Ultrafast Jet Classification at HL-LHC

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. We compare Deep Sets and Interaction Network models, which are permutation-invariant, with MLP, which is not not permutation-invariant. Through quantization-aware training (QAT) and efficient FPGA implementations, we show that nanosecond inference using complex architectures like Deep Sets and Interaction Networks are feasible at low resource-cost.

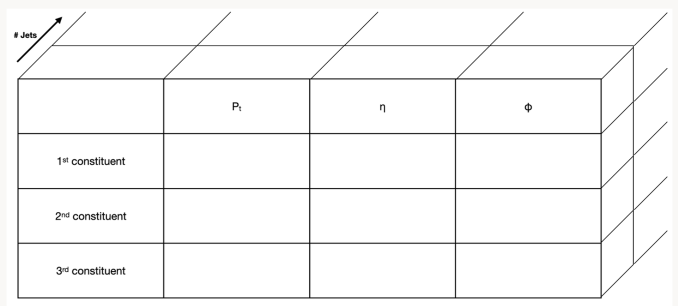
Neural Network Architechtures



The Models are implemented using Keras and Tensorflow libraries. For DS model the input jet data is structured as an unconnected graph, while for the IN it's structured as a fully connected graph. Each graph 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 scenarios of jets with $N = 8, 16, 32$ constituents. We define the constituents features as P_t ,η and ϕ relative to the jet axis.



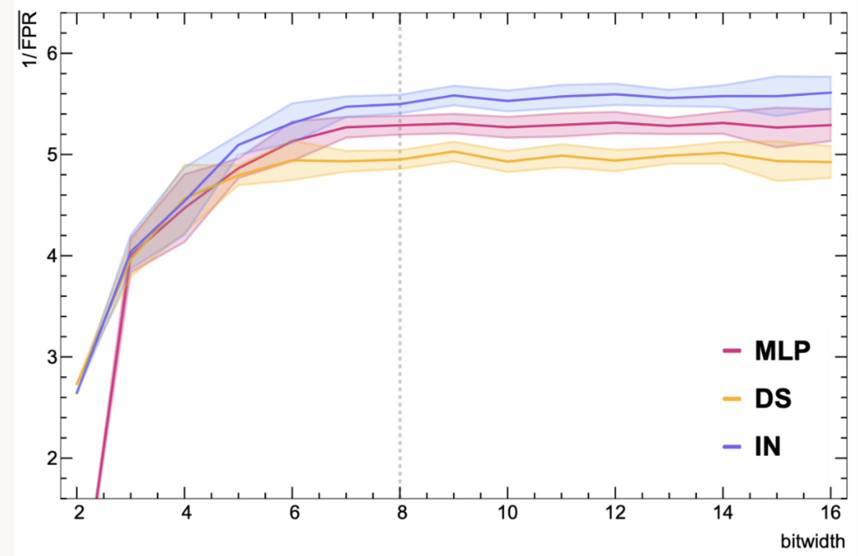
Full Precision Models Performance

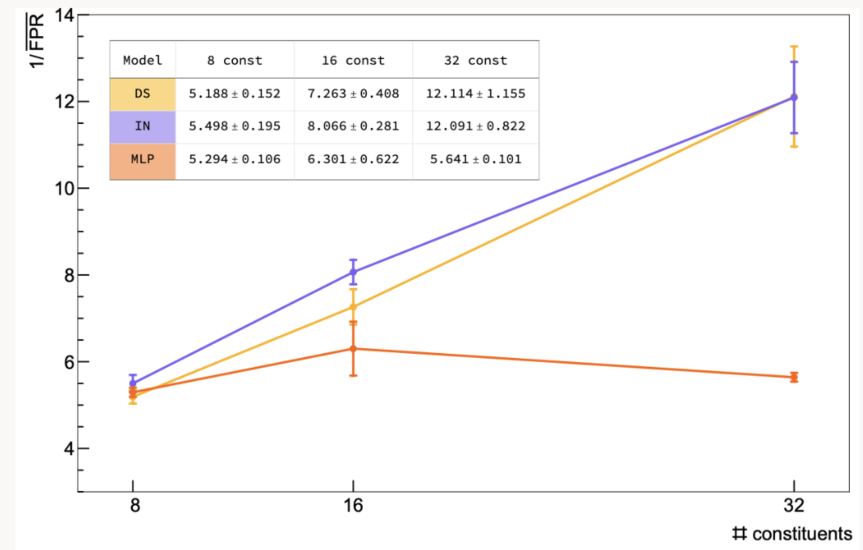
Each model is first trained at floating-point precision, establishing the baseline performance that the quantised model on FPGA should match. The table bellow shows the model performance for 8, 16, and 32 jet constituents. The uncertainties on the AUCs are all ~ 0.001 and thus not included for legibility.

Architecture	Constituents	Parameters	FLOPs	Accuracy	AUC g	AUC q	AUC W	AUC Z	AUC t
MLP		26,826	53,162	64.6 ± 0.1%	0.84	0.88	0.90	0.88	0.92
DS	8	3,461	36,805	64.0 ± 0.3%	0.84	0.88	0.90	0.88	0.92
IN		3,347	37,232	64.9 ± 0.2%	0.84	0.88	0.91	0.89	0.92
MLP		20,245	40,485	68.4 ± 0.3%	0.87	0.89	0.91	0.90	0.94
DS	16	3,461	71,109	69.4 ± 0.2%	0.87	0.89	0.93	0.92	0.94
IN		3,347	140,432	70.8 ± 0.2%	0.88	0.90	0.94	0.92	0.94
MLP		24,101	48,197	66.2 ± 0.2%	0.90	0.89	0.89	0.88	0.94
DS	32	3,461	139,717	75.9 ± 0.1%	0.91	0.91	0.96	0.95	0.95
IN		7,400	109,556	75.8 ± 0.3%	0.91	0.91	0.96	0.95	0.95

Quantized Models Performance

Quantization aware training (QAT) of each model is implemented using the QKeras library. Pruning to 50% sparsity is applied to the 32-constituent IN model so it can fit within the FPGA resource constrains and for consistency, the same pruning sparsity is applied to the 32-constituent MLP and DS models. Plots shows the models performance in terms of inverse average FPR (False Positive Rate) as a function of the bitwidth and the number of constituents.





Results

The models described above are translated into firmware using HLS4ML library [3], then synthesized with Vivado HLS, targeting a Xilinx Virtex UltraScale+ VU13P FPGA with a clock frequency of 200 MHz. Number of jet constituents, reuse factor(RF), latency, initialization interval (II) and resource consumption for the models quantized to 8 bits are shown in the table. The cc next to the latency and II represents the number of clock cycles on the FPGA. ‘

Architecture	Constituents	RF	Latency [ns] (cc)	II [ns] (cc)	DSP	LUT	FF	BRAM18
	8	1	105 (21)	5 (1)	262 (2.1%)	155,080 (9.0%)	25,714 (0.7%)	4 (0.1%)
MLP	16	1	100 (20)	5 (1)	226 (1.8%)	146,515 (8.5%)	31,426 (0.9%)	4 (0.1%)
	32.	1	105 (21)	5 (1)	262 (2.1%)	155,080 (7.2%)	25,714 (0.7%)	4 (0.1%)
	8	2	95 (19)	15 (3)	626 (5.1%)	386,294 (22.3%)	121,424 (3.5%)	4 (0.1%)
DS	16	4	115 (23)	15 (3)	555 (4.5%)	747,374 (43.2%)	238,798 (6.9%)	4 (0.1%)
	32.	8	130 (26)	10 (2)	434 (3.5%)	903,284 (52.3%)	358,754 (10.4%)	4 (0.1%)
	8	2	160 (32)	15 (3)	2,191 (17.8%)	472,140 (27.3%)	191,802 (5.5%)	12 (0.2%)
IN	16	4	180 (36)	15 (3)	5,362 (43.6%)	1,387,923 (80.3%)	594,039 (17.2%)	52 (1.9%)
	32.	8	205 (41)	15 (3)	2,120 (17.3%)	1,162,104 (67.3%)	761,061 (22.0%)	132 (2.5%)

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
4) Ultrafast Jet Classification on FPGA at the HL-LHC - arXiv:2402.01876v2 [hep-ex]

Conclusions

Neural network based jet classification algorithms are synthesized on FPGA that mimic the environment within the hardware layers of the real-time data processing systems for a typical HL-LHC experiment. Using jet data with constituent level information, we show how one could synthesize machine learning algorithms pertaining to three different data representations on an FPGA by using the hls4ml library. Deep Sets network strikes a good balance between accuracy, latency, and resource consumption compared with the deployed and tested MLP and IN models. In conclusion, we have identified and shown the necessary ingredients to deploy a jet classifier in the level-1 trigger of the HL-LHC

Andre Sznajder (UERJ)

in collaboration with: P.Odagiu (ETH), Z.Que(ICL), J.Duarte(UCSD), J.Haller(UH), G.Kasieczka(UH), A.Lobanov(UH), V.Loncar(MIT), W.Luk(ICL), J.Ngadiuba(FNAL), M.Pierini(CERN), P.Rincke(Uppsala), A.Seksaria(Uppsala), S.Summers(CERN), A.Tapper(ICL), T.K.Arrestad(ETH)

andre.sznajder@cern.ch