

Low-latency Jet Tagging for HL-LHC Using Transformer Architectures

Lauri Laatu, Arianna Cox, Abhijith Gandrakota, Benedikt Maier, Jennifer Ngadiuba, Zhiqiang Que, Chang Sun, Alexander Tapper

09.07.2025



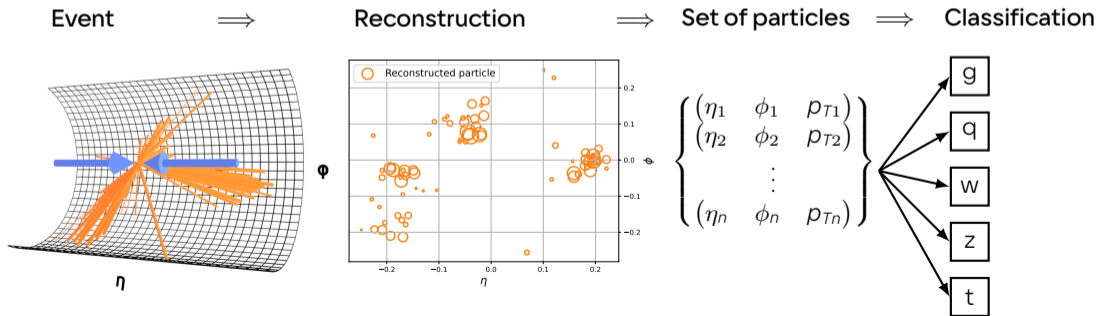
Caltech



IMPERIAL

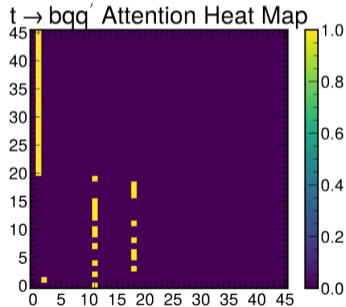
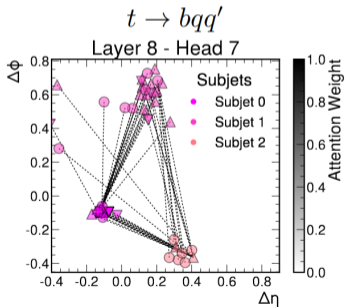
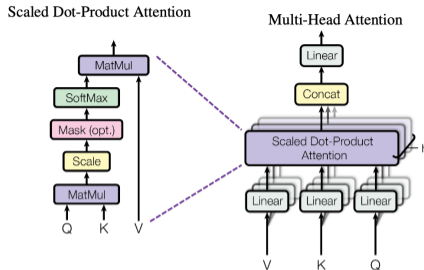
Jet Tagging

- Jet tagging is the process of identifying the type of particle that initiates a jet
- This work is targeting trigger and scouting applications, latency requirement $O(1 \mu s)$
- Dataset of high- p_T jets from simulations of LHC proton-proton collisions (10.5281/zenodo.3602259)
- Common benchmark dataset for FastML studies
- Up to 150 particles with three features p_T , η and ϕ



Attention for Jet Tagging

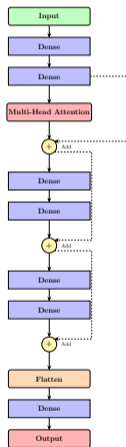
- Models based on attention perform well in HEP
 - Particle transformer (ParT) arXiv:2202.03772
 - Pileup mitigation 10.1088/2632-2153/ac7198
 - Previous transformer implementations for FPGA: 10.1109/ICFPT56656.2022.9974463 & arXiv:2402.01047
- In the case of a jet tagging, the attention matrix captures the particle-to-particle correlations arXiv:2412.03673



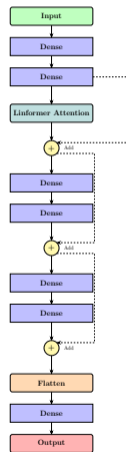
Models Benchmarked in These Studies

- Models with attention use either
 - **Multi-Head Attention**
 - **Linformer Attention** (arXiv:2006.04768)
- Comparison to Deep Sets
- Allows for message passing at particle level

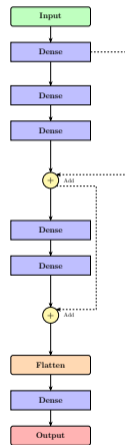
Multi-Head



Linformer



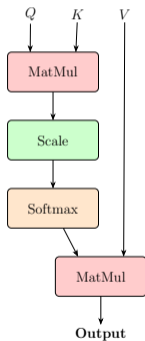
Deep Set



Attention Layers

- Multi-Head Attention scales $\mathcal{O}(n^2)$, where n is the sequence length
- Linformer Attention scales $\mathcal{O}(n \cdot k)$, where k is the projection dimension

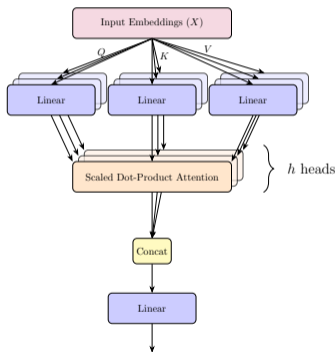
Scaled Dot-Product Attention



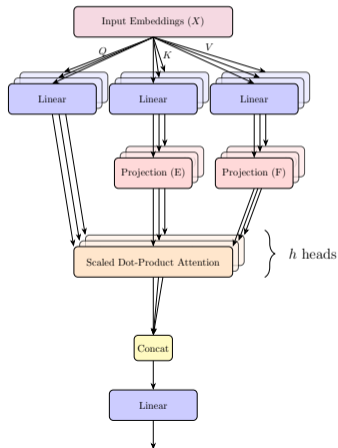
Scaled Dot-Product Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Multi-Head

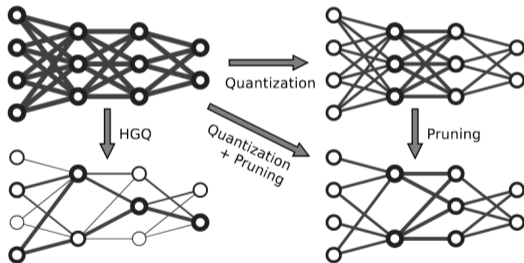


Linformer



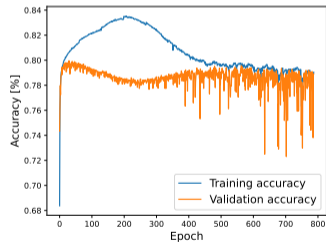
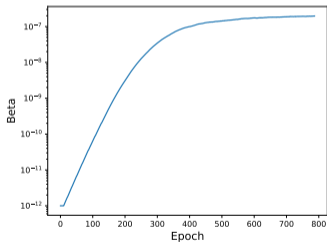
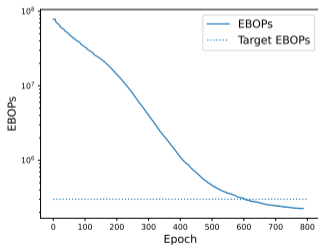
High Granularity Quantization

- HGQ is able to perform per-weight and datalane quantization (arXiv:2405.00645)
- Estimates Effective Bit Operations (EBOPs)
 - EBOPs is minimized with a regularizer parameter **Beta**



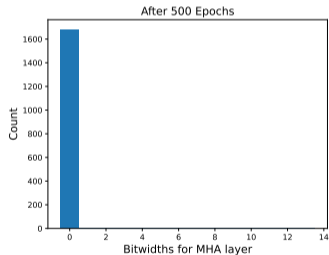
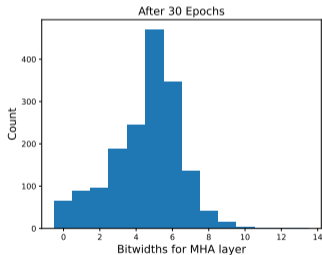
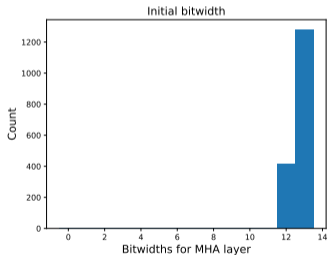
Controlling Beta

- FPGAs have a known amount of resources available
- Instead of setting **Beta**, set **target EBOPs**
- Adapt **Beta** throughout training with a technique from control engineering: PID control (BetaPID)



High Granularity Quantization with Attention

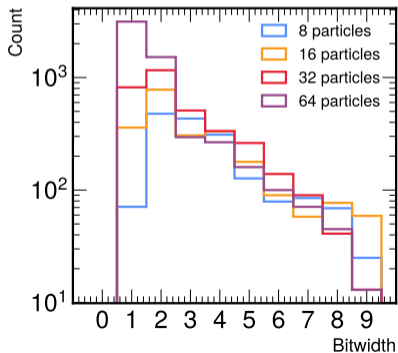
- HGQ has the tendency to prune out the attention layer
- Requires constraining bitwidth for attention to at least 1 bit



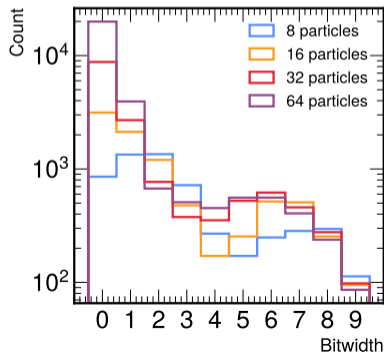
Model Bitwidths

- Bitwidths for Linformer
- Target EBOPs set at 350k (corresponds to size that fits on a single SLR)
- Longer input models with higher sparsity to achieve the target EBOPs

Attention



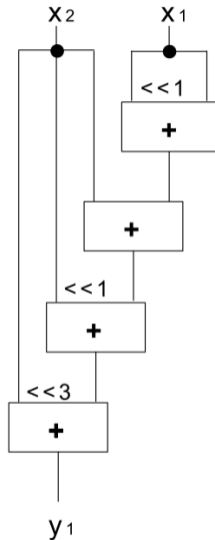
Rest of the model



Distributed Arithmetic

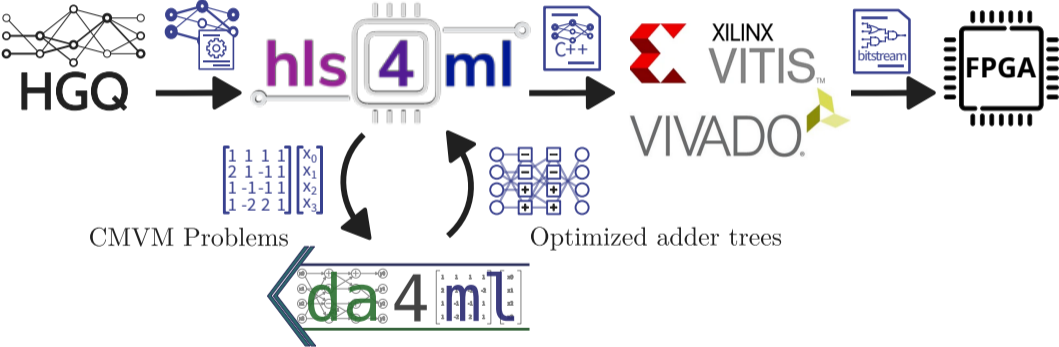
- Constant Matrix Vector Multiplication (CMVM) operations optimized using Distributed Arithmetic (DA)
- Split multiplications into additions and bitshifts
- Reduces further the resource usage
- Example:

$$y_1 = 3x_1 + 11x_2$$



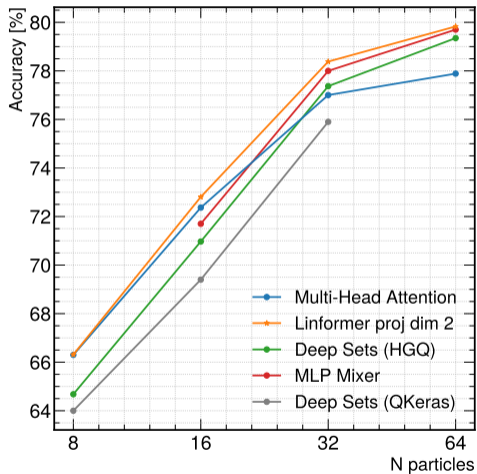
Distributed Arithmetic in hls4ml

- Enabled in hls4ml with da4ml package (arXiv:2507.04535)



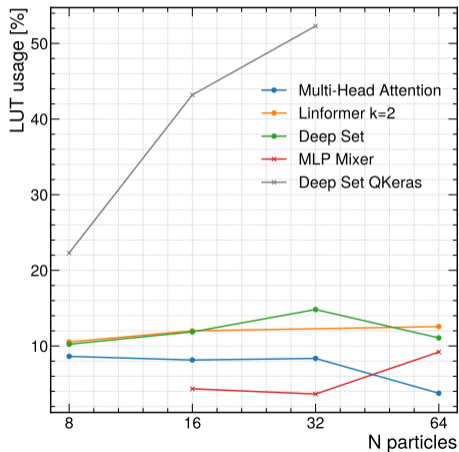
Performance

- Deepset vs Transformer performance with fixed **target EBOPs** to fit on FPGA
 - With fixed EBOPs and high input length, two MHA layers require too much compute to perform well
- Comparing to previously published work
 - Deepset with QKeras arXiv:2402.01876
 - MLP-Mixer arXiv:2503.03103



Resource Usage

- HGQ models
 - Pipelined at II=1
 - Latency ≈ 100 ns
 - FMax 150 Mhz - 200 Mhz
 - FPGA: xcu250-figd2104-2L-e
- Fixed EBOPs to fit single SLR

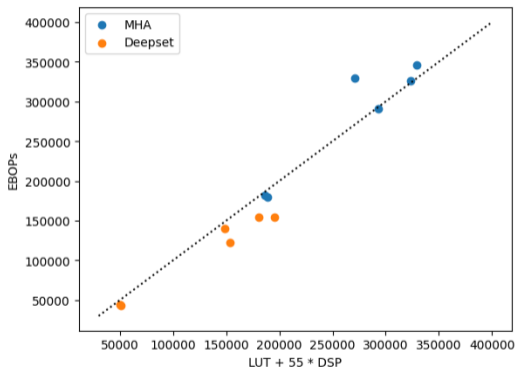


Conclusion

- HGQ can be used for quantizing Transformer models
- Multi-Head Attention doesn't scale well for FPGAs
 - Scales $\mathcal{O}(n^2)$ with input length
- Linear attention from Linformer can be used for efficient attention on FPGAs
 - Scales $\mathcal{O}(n \cdot k)$ with input length
- Training with HGQ yields up to two orders of magnitude reduction in EBOPs
- Resource usage for HGQ Deepset up to $\approx 80\%$ lower than per-layer quantization while having better performance
- Next steps
 - More complex dataset that corresponds better to HL-LHC conditions
 - Event level processing

Ebops vs resource usage

- EBOPs is a good proxy for resource usage on FPGA



Transformer vs Deepset Performance in FP

- Performance with different amount of input particles
 - Using the highest p_T particles
- Two attention heads in the MHA models
- Models with attention perform better → Scaling up Deepset is not enough

