Transformers in FINN

Christoph Berganski
Computer Engineering Group
Paderborn University

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

SAIL
SUSTAINABLE LIFE-CYCLE OF INTELLIGENT SOCIO-TECHNICAL SYSTEMS

eki
ENERGIEEFFIZIENTE KI DURCH DNN-APPROXIMATION FÜR FPGAs

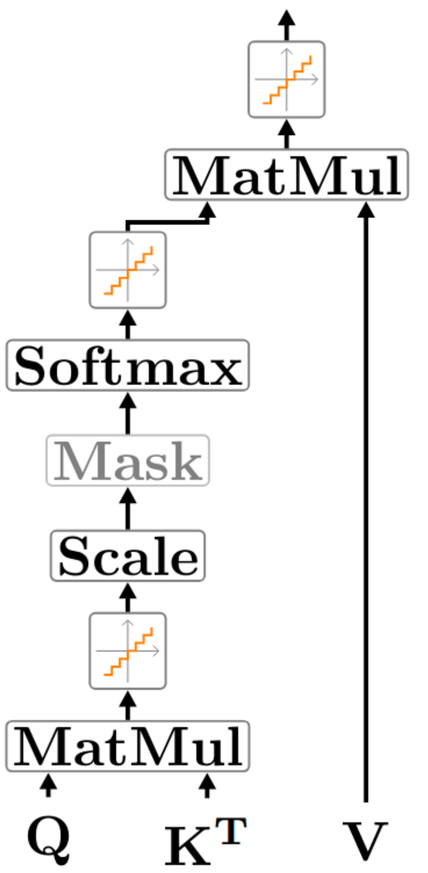
PADERBORN UNIVERSITY

Supported by:
Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection
based on a decision of the German Bundestag

Keywords: *Quantized Transformer, Scaled Dot-Product Attention, Hardware Operator, Compiler Infrastructure*

SAIL is funded under the grant no NW21-059D

Transformer Architecture



Quantized Scaled Dot-Product Attention

Since its introduction, the Transformer[1] architecture based on the scaled dot-product attention mechanism has evolved to dominate the state of the art in almost any deep leaning domain.

Applications of Transformer-based models cover natural language processing with GPTs[2-3] and BERTs[4], computer vision with ViTs[5], audio, speech and various signal processing use-cases.

Besides the typically large size, the architecture poses various challenges to a dataflow-style implementation:

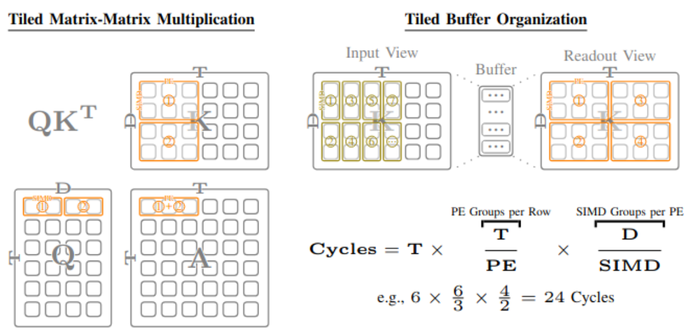
- Quadratic complexity of attention and more complex input dependencies drive buffer requirements
- Transposing cannot be done without extra buffering
- Softmax requires multiple passes over attention matrix and evaluation of exponential functions

There is no official support for scaled dot-product attention in FINN[6-7], but we will contribute this soon.

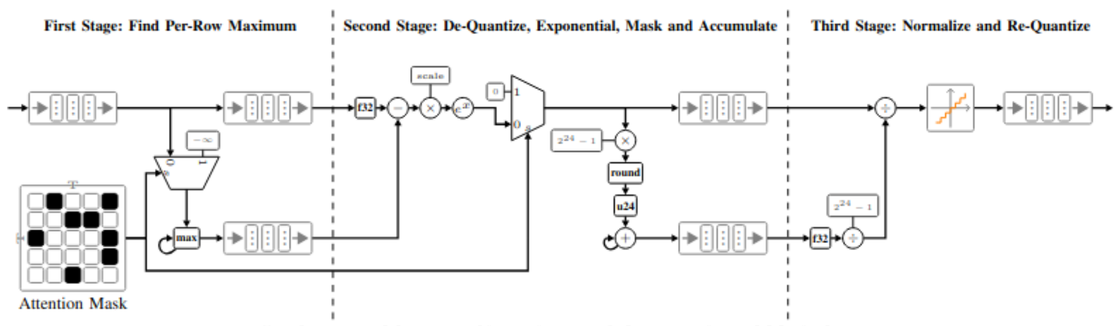
New tiled matrix multiplication operator receiving two inputs at runtime vs. existing MVU which multiplies static weights to a single input.

Transpose of the input V is handled at the readout of the internal buffer.

At the core, these are MAC units with SIMD and PE parallelism like the MVU.



Queries-keys matrix multiplication. Attention weight-values looks similar, but with transposed buffer readout.



Softmax Normalization of Attention Weights

Compiler Infrastructure

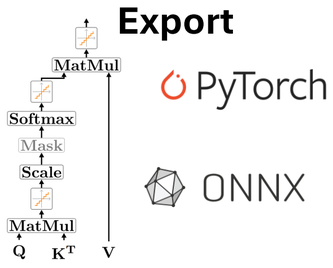
Extensions to **streamlining**, data layouts (NxTxD to TxD), graph pattern matching, code generation, and **parallelization** options (SIMD, PE and unrolled **attention heads**), yield:

$$\text{Target Throughput}\left[\frac{\text{seq}}{\text{s}}\right] = \frac{\text{Clock Frequency}}{\text{T}^2}$$

Semi-automatic FIFO-sizing allows FIFOs to reach depths of up to T elements for the attention inputs and T^2 elements for the residual branches, resulting in significant memory resource usage.

Transformer Quantization

Quantization and Export

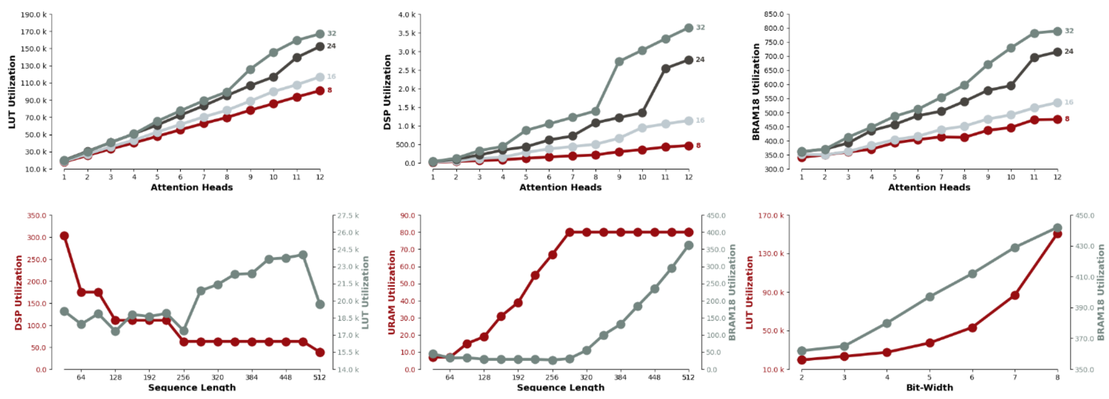


Flexible quantization aware training (**QAT**) in Brevitas[8] targeting **2 to 8 bits** for weights, biases and activations

Quantized embedding and positional encoding, trained at 8 bits precision as all model inputs/outputs

Effect on accuracy (**Acc.**) and perplexity (**Ppl.**) evaluated on small Transformers for radio signal classification and small GPTs generating child stories.

Evaluation



Resource utilization for different scaling options: All variations, if not varied throughout the scaling sweep, are quantized to 2 bits and correspond to a single head of 32-dimensional embeddings and sequence length of 512. The top row shows scaling up the number of heads while keeping the dimension per head fixed at 8, 16, 24 or 32 dimensions. The bottom-left shows scaling over increasing sequence lengths. The bottom right shows scaling over increasing quantization bit-width

Operator	Count	LUT	DSP	BRAM18	URAM
Attention	12	99421	180	156	0
Thresholding	10	606	0	0	0
MVU	6	37085	3456	238	0
Elementwise Add	3	317	0	0	0
Replicate Stream	3	161	0	0	0
Split Heads	3	2562	0	0	0
Merge Heads	1	846	0	0	0
FIFO	112	14681	0	388	80
DWC	8	1490	0	0	0
IODMA	2	2319	0	8	0
Total	**160**	**159488**	**3636**	**790**	**80**

Resource breakdown for Transformer-block of 12 heads, 512x384 inputs at 2-bit quantization

Model	Acc.	LUT	DSP	BRAM18	URAM	Latency
(A)	73.1%	231k	1840	717	56	44 μs
(B)	67.9%	266k	2344	793	80	58 μs
(C)	69.7%	213k	2352	763	56	44 μs
(D)	67.5%	274k	1824	721	80	44 μs
(0)	68.2%	181k	1584	532	56	44 μs

Radio Signal Classification Transformer at 75k seq/s

Model	Ppl.	LUT	DSP	BRAM18	URAM	Latency
(A)	17.8	864k	5244	2682	272	2.05 ms
(B)	21.0	465k	5756	2198	224	2.05 ms
(C)	16.9	776k	7292	3470	704	9.93 ms
(D)	32.5	311k	3362	1646	112	1.38 ms

Small GPT generating child stories at 1494 seq/s (A, B, D) and 307 seq/s (C)

Discussion and Future Work

Successful demonstration of Transformers in FINN currently being cleaned up for eventual contribution via pull-requests. This allows to easily explore the design space for Transformers across a range of topologies, dimensions and precisions.

Future work will address improved auto FIFO-sizing, pure integer Softmax, scaling up model sizes to explore advanced compression techniques and multi-FPGA deployment, and further use-cases like vision and audio processing.

Also planned for future work are comparison to GPUs, other FPGA accelerators and more detailed measurements of resource and power consumption.

References

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
2. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al., “Improvinglanguage understanding by generative pre-training,” 2018.
3. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., “Language models are unsupervised multitask learners,” OpenAI blog, vol. 1, no. 8, p. 9, 2019
4. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
5. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” ICLR, 2021.
6. Y. Umuroglu, N. J. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre, and K. Vissers, “Finn: A framework for fast, scalable binarized neural network inference,” in Proceedings of the 2017 ACM/SIGDA Interna- tional Symposium on Field-Programmable Gate Arrays, ser. FPGA ’17. ACM, 2017, pp. 65–74
7. M. Blott, T. B. Preußer, N. J. Fraser, G. Gambardella, K. O’brien, Y. Umuroglu, M. Leeser, and K. Vissers, “Finn-r: An end-to-end deep-learning framework for fast exploration of quantized neural networks,” ACM Transactions on Reconfigurable Technology and Systems (TRETS), vol. 11, no. 3, pp. 1–23, 2018.
8. A. Pappalardo, “Xilinx/brevitas,” 2023. [Online]. Available: https://doi.org/10.5281/zenodo.3333552