FAMOUS: Flexible Accelerator for the Attention Mechanism of Transformer on UltraScale+ FPGAs

Ehsan Kabir, Md. Arafat Kabir, Austin R. J. Downey, Jason D. Bakos, David Andrews, Miaoqing Huang

UNIVERSITY OF ARKANSAS

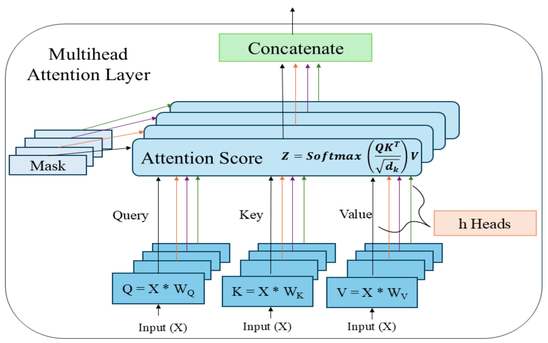
UNIVERSITY OF South Carolina

Introduction

- Transformer neural networks (TNN) have demonstrated significant advancements in natural language processing, machine translation, computer vision.
- A remarkable feature named multi-headed attention (MHA) mechanism enables a high level of computational parallelism for both the training and inference phases.
- Suitable for acceleration on hardware like FPGAs, due to FPGA's high degree of parallelism, low latency, and energy efficiency.
- Most of the FPGA or ASIC-based accelerators for TNN have specialized sparse architecture for a specific application. Thus, they lack the flexibility to be reconfigured for a different model during runtime.
- We applied efficient tiling and wrote efficient high level synthesis (HLS) code to increase parallelism for dense computations of MHA.

Background

The input sequence X is linearly mapped into Query (Q), Key (K), Value (V) matrices using weights and biases. The parameter dk = dmodel/h is the 2nd dimension of Q and K. dmodel is a hyperparameter called embedding dimension and h is number of heads.

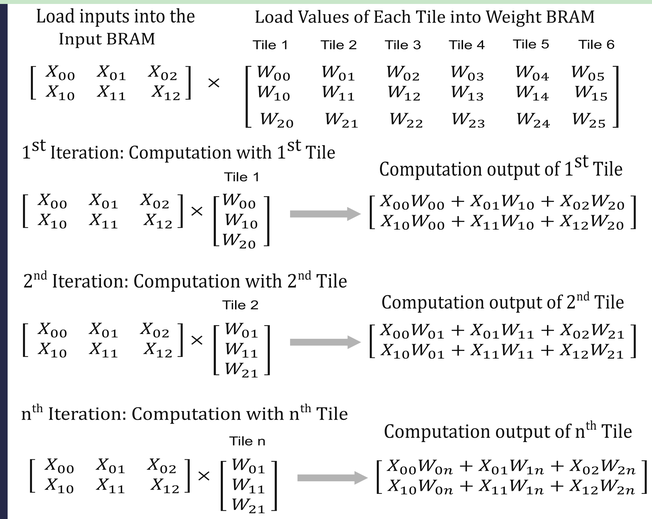


Accelerator Architecture

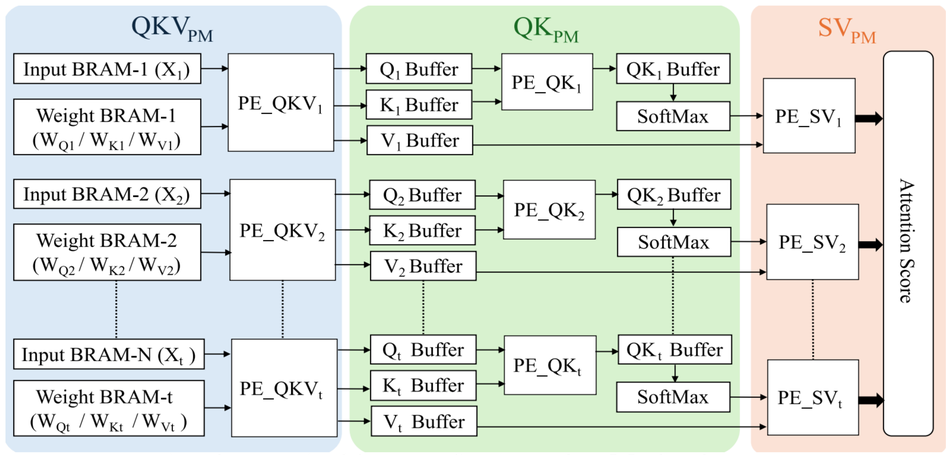
- Designed with C in Vitis HLS.
- Three main processing modules:QKV_{PM} , QK_{PM} and SV_{PM}. QKV_{PM} : Generates Q, K, V matrices. QK_{PM} : Matrix-matrix multiplication operations between the Q and K matrices.
- SV_{PM} : Matrix-matrix multiplication operations with V and the output from QK_{PM}.

Accelerating the Computation within the Attention Layer of the Transformer

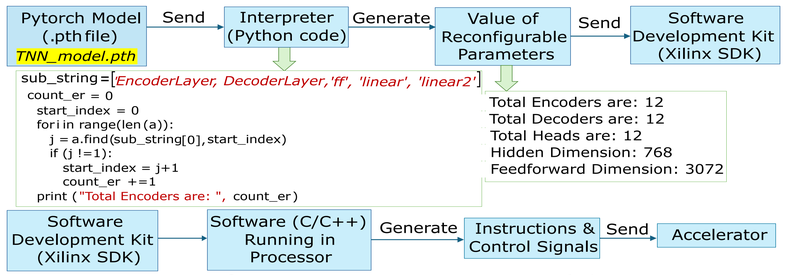
- A novel architecture ensuring high BRAM and DSP utilization for efficient parallel processing with low latency.
- An efficient tiling of weight matrices to accommodate large models in on-chip memory.
- 3.28× and 2.6× faster than the Intel Xeon Gold 5220R CPU and NVIDIA V100 GPU respectively.
- 1.3× faster than the fastest state-of-the-art FPGA-based accelerator.



Tiling Technique



Accelerator Architecture for Attention Mechanism



Process for Incorporating Programmability

Download This Poster

Comparison with Other FPGA Accelerators

Works	Calabash [6]	Lu et al. [9]	Ye et al. [8]	Li et al. [7]	Peng et al. [4]	*FAMOUS*
FPGAs	Xilinx VU9P	Xilinx VU13P	Alveo U250	Xilinx VU37P	Alveo U200	Alveo U55C
Method	HDL	HDL	HDL	HLS	HLS	HLS
DSPs	4227	129	4189	1260	623	4157
BRAMs	640	498	1781	448	–	3148
GOPS	1288	128	171	72	97	623
Latency (ms)	0.239[a]	0.8536[b]	0.642	1.5264	1.706[c]	0.494

[a] Q, K, V matrix computation time ignored.
[b] Time adjusted for 8 attention heads.
[c] Time extracted for attention mechanism from a full transformer.

Comparison with Other Acceleration Platforms

Platform	Intel E5 CPU [6]	NVIDIA V100 GPU [7]	Intel Xeon CPU [8]	NVIDIA P100 GPU [8]	*FAMOUS* (Alveo U55C FPGA)	
Topologies	64, 768, 12	64, 512, 4	64, 512, 8	64, 512, 4	64, 768, 8	64, 512, 8
GOP	0.308	0.11	0.11	0.11	0.308	0.11
Latency (ms)	1.1	1.5578	1.96	0.496	0.94	0.597
GOPS	280	71	56	221	328	184

Overall Result for MHA Accelerator

Test no.	Sequence Length	Embedding Dimension	Number of Heads	Tile Size	FPGA	Data Format	DSPs	BRAMs 18k	LUTs	FFs	Latency (ms)	GOPS
#1	64	768	8	64	Alveo U55C	8bit fixed	4157 (46%)	3148 (78%)	1284782 (98%)	661996 (25%)	0.94	328
#2			4								1.401	220
#3			2								2.281	135
#4	64	512	8	64	Alveo U55C	8bit fixed	4157 (46%)	3148 (78%)	1284782 (98%)	661996 (25%)	0.597	184
#5		256									0.352	312
#6	128	768	8	64	Alveo U55C	8bit fixed	4157 (46%)	3148 (78%)	1284782 (98%)	661996 (25%)	2	314
#7	32										0.534	285
#8	16										13	16