2025 IEEE 33rd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) | 979-8-3315-0281-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/FCCM62733.2025.00053

Optimized Coding and Parameter Selection for Efficient FPGA Design of Attention Mechanisms

Ehsan Kabir*, Austin R.J. Downey$, Jason D. Bakos†, David Andrews*, Miaoqing Huang*
*Department of EECS, University of Arkansas, Fayetteville, Department of †CSE, $ME, University of South Carolina, USA
{ekabir, dandrews, mqhuang}@uark.edu, austindowney@sc.edu, jbakos@cse.sc.edu

***Abstract*—Efficient utilization of on-chip computational and memory resources, along with optimized high-level synthesis (HLS) coding, is vital to maximize parallelism and minimize latency. This paper demonstrates the HLS algorithms to achieve high utilization of processing elements to enhance parallelism. It also analyzes how various parameters of an attention layer impact latency, employs an efficient tiling technique, and explains the process of selecting an optimized tile size (TS).**

***Index Terms*—FPGA, Attention-based Neural Networks, High-Level Synthesis, Hardware Accelerators.**

I. Introduction

High level Synthesis (HLS) tool offers faster development, but writing efficient HLS code to maximize DSP utilization remains a challenge [1]. The attention layer is the most resource-intensive component of a deep neural network (DNN) [2]. However, FPGA resources are often limited, so the input matrices must be partitioned into tiles, and developing an optimal partitioning strategy is a significant challenge. This paper addressed these challenges by making the following contributions:

- A novel architecture designed by efficient coding in HLS and efficient tiling of weight matrices for enhancing DSP consumption in the processing elements (PE).
- A theoretical model to validate both predicted and experimental latency.

II. HLS Design Technique

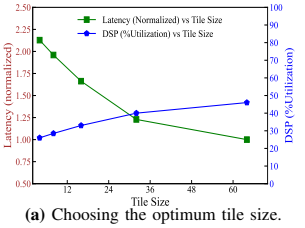
Algorithm 1 Q Matrix Calculation

```
1:  for (i = 1; i <= Sequence Length; i = i + 1) do
2:      #pragma HLS pipeline off
3:      S_q ← 0
4:      for (k = 1; k <= d_model/h; k++) do
5:          #pragma HLS pipeline II = 1
6:          for (j = 1; j <= Tiles; j++) do
7:              S_q ← S_q + x[i][j] × w_q[k][j];
8:          end for
9:          Q[i][k] ← Q[i][k] + S_q;
10:     end for
11: end for
```

There are loading units and computing modules generated by the HLS design technique on Vitis high-level synthesis (HLS) 2022.2.1 tool. Due to space limitations, only one HLS algorithm (1) is shown; others can be found in [3]. It outlines the computations of query matrix, where the 2nd loop (line 6) is pipelined causing the innermost loop (line 8) to be fully unrolled. This generates ($\frac{d_{model}}{TS}$) PEs.

III. Tile Size and Head Number Determination

The tiling technique is described in [3]. The graph in Fig. 1a indicates that the optimal tile size for achieving the highest DSP consumption and the lowest latency was 64. This combination would achieve 46% DSP utilization. The graph in Fig. 1b indicates that the optimal number of heads to achieve the highest LUT consumption (98%) and the lowest latency was 8.



(a) Choosing the optimum tile size. **(b)** Choosing the optimum head number.

IV. Analysis and Conclusion

The resource utilization and performance of the accelerator are influenced by parameters such as tile size or the number of tiles, the number of attention heads, sequence length, and embedding dimension, assuming a fixed bit width. An analytical model was created to define the relationship between these parameters and both latency and resource utilization.

TABLE I: Validation of Experimental and Analytical Results

Method	Sequence Length	Embedding Dimension	Number of Heads	Tile Size	DSPs	Frequency (MHz)	Latency (ms) Attention (SA)
Experimental	64	768	8	64	4157	400	0.94
Analytical					4168		0.98
Experimental	32	768	8	64	3898		0.534
Analytical					3656		0.579
Experimental	64	512	8	64	3887		0.597
Analytical					3656		0.592

Table I presents a comparison between the experimental data and theoretical results. It shows the combined latency and DSP usage of the computing and loading units for selected configurations. The experimental results closely align with the theoretical predictions. In conclusion, a maximum tile size of 64 and 8 parallel attention heads achieve optimal utilization of compute and memory resources on the Alveo U55C FPGA platform.

References

[1] E. Kabir et al., "Accelerating lstm-based high-rate dynamic system models," in *FPL Conference*, 2023.
[2] B. Li et al., "Ftrans," in *International Symposium on Low Power Electronics and Design*, 2020.
[3] E. Kabir et al., "Protea," in *SC24-W: Workshops*, 2024.

2576-2621/25/$31.00 ©2025 IEEE
DOI 10.1109/FCCM62733.2025.00053