[반도체전공트랙 성과발표회]

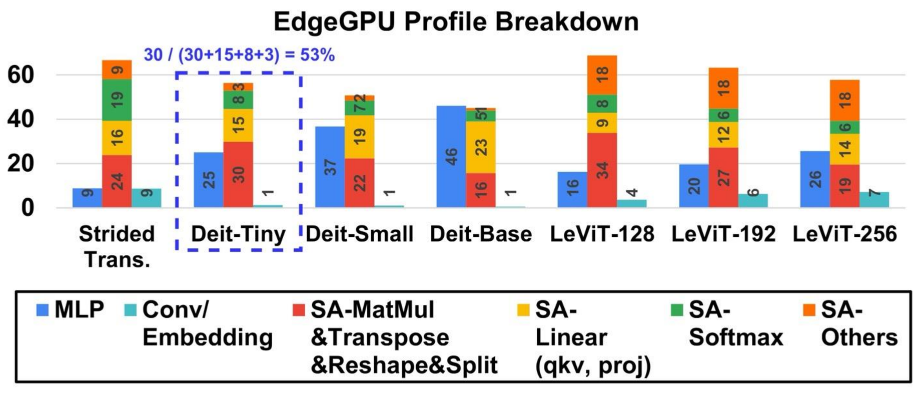
Data-aware Sparsity for ViT: Efficient Algorithm-HW Co-design with Windowed Attention

Author : Yeonsik Park, Inwook An, Jaeyoon Lee, Changhyun Kim Advisor : Seungkyu Choi

Department of Electronic Engineering, College of Electronics & Information, Kyung Hee University

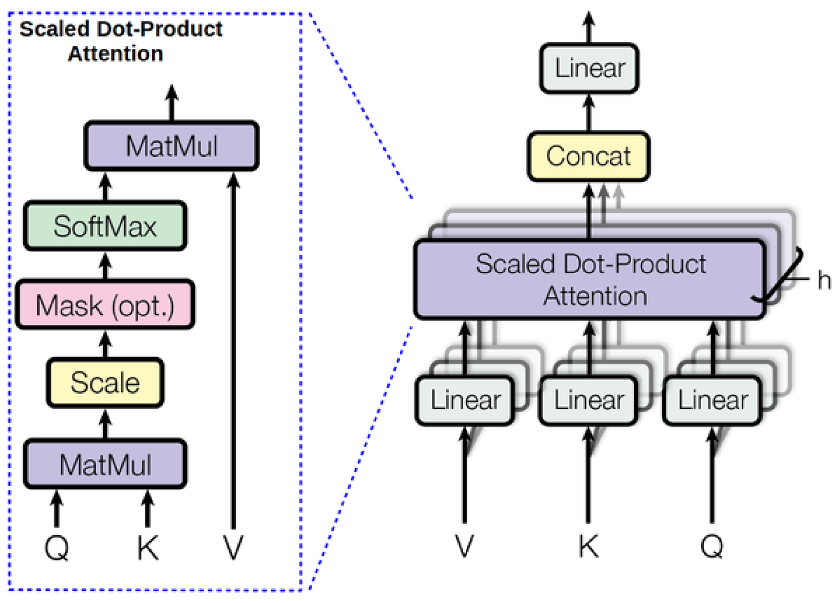
Introduction

After revolutionizing the field of Natural Language Processing (NLP), Transformer models have recently demonstrated performance surpassing CNNs in the domain of Computer Vision (CV). While the attention mechanism provides significant performance improvements, the computational complexity of the core operation of Transformers, Attention, increases quadratically with the number of input tokens. As shown in Fig. 1, the Self-Attention module accounts for over 50% and up to 69% of latency when executed on edge devices. Existing attention accelerators have primarily focused on NLP Transformers. Therefore, we aim to conduct research on algorithm-accelerator co-design optimized for the CV domain.

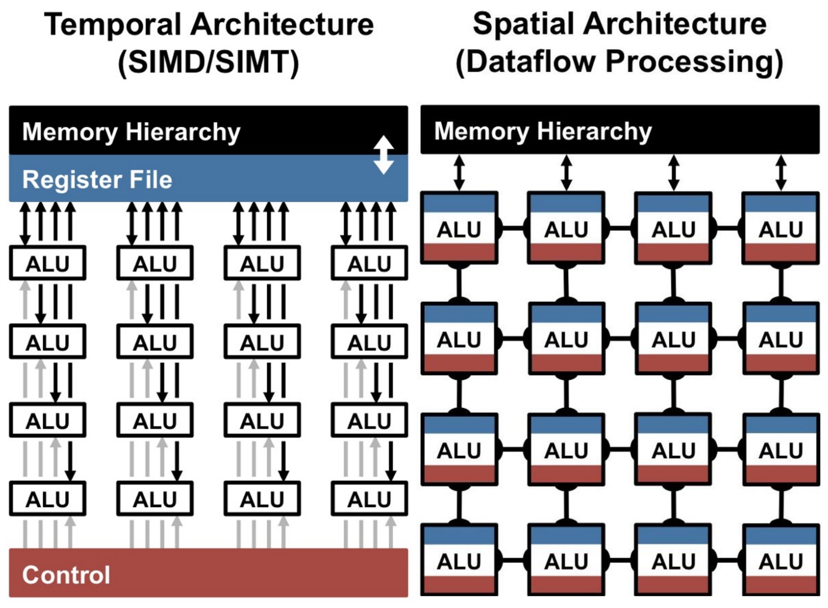


[Fig. 1] Measured Latency Breakdown of Various ViTs

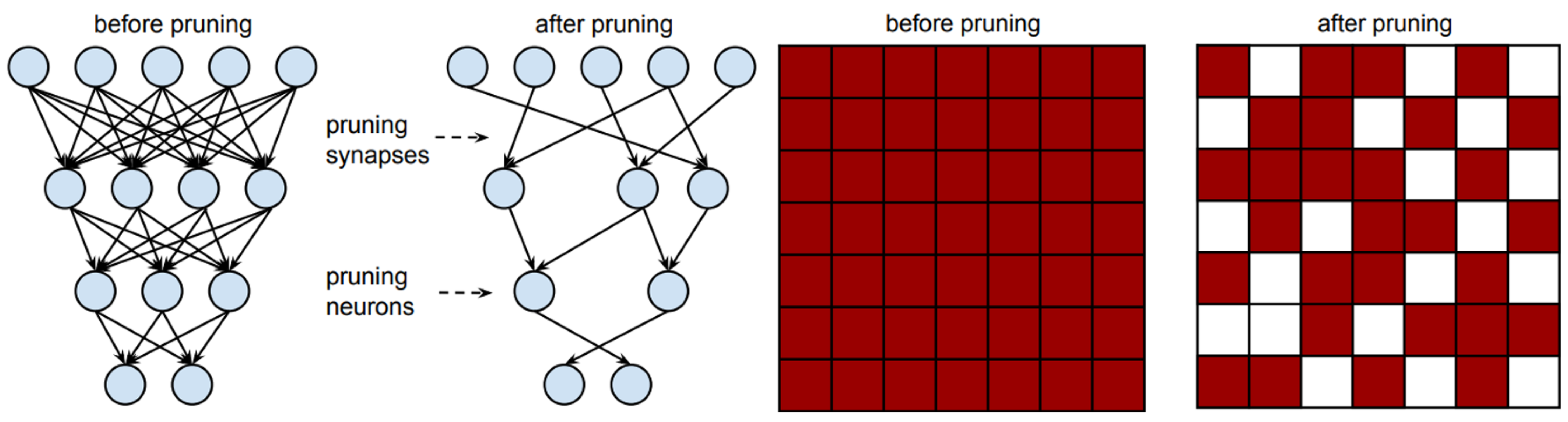
Background



[Fig. 2] Multi- Head Self Attention (MHSA)



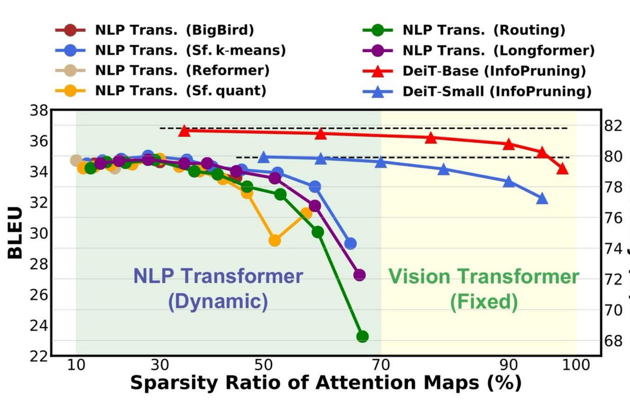
[Fig. 3] SIMD/SIMT vs Systolic Array



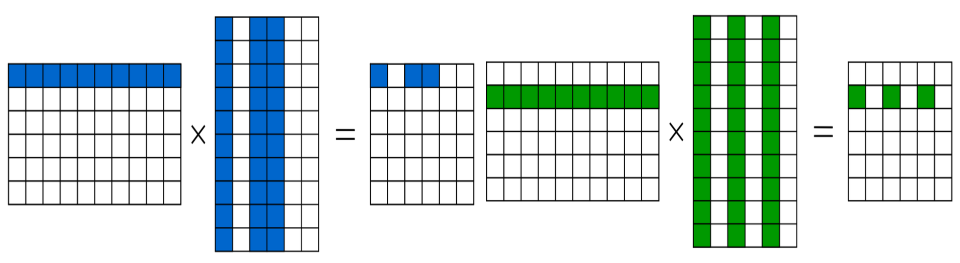
[Fig. 4] Pruning and resulting matrix sparsity

Motivation

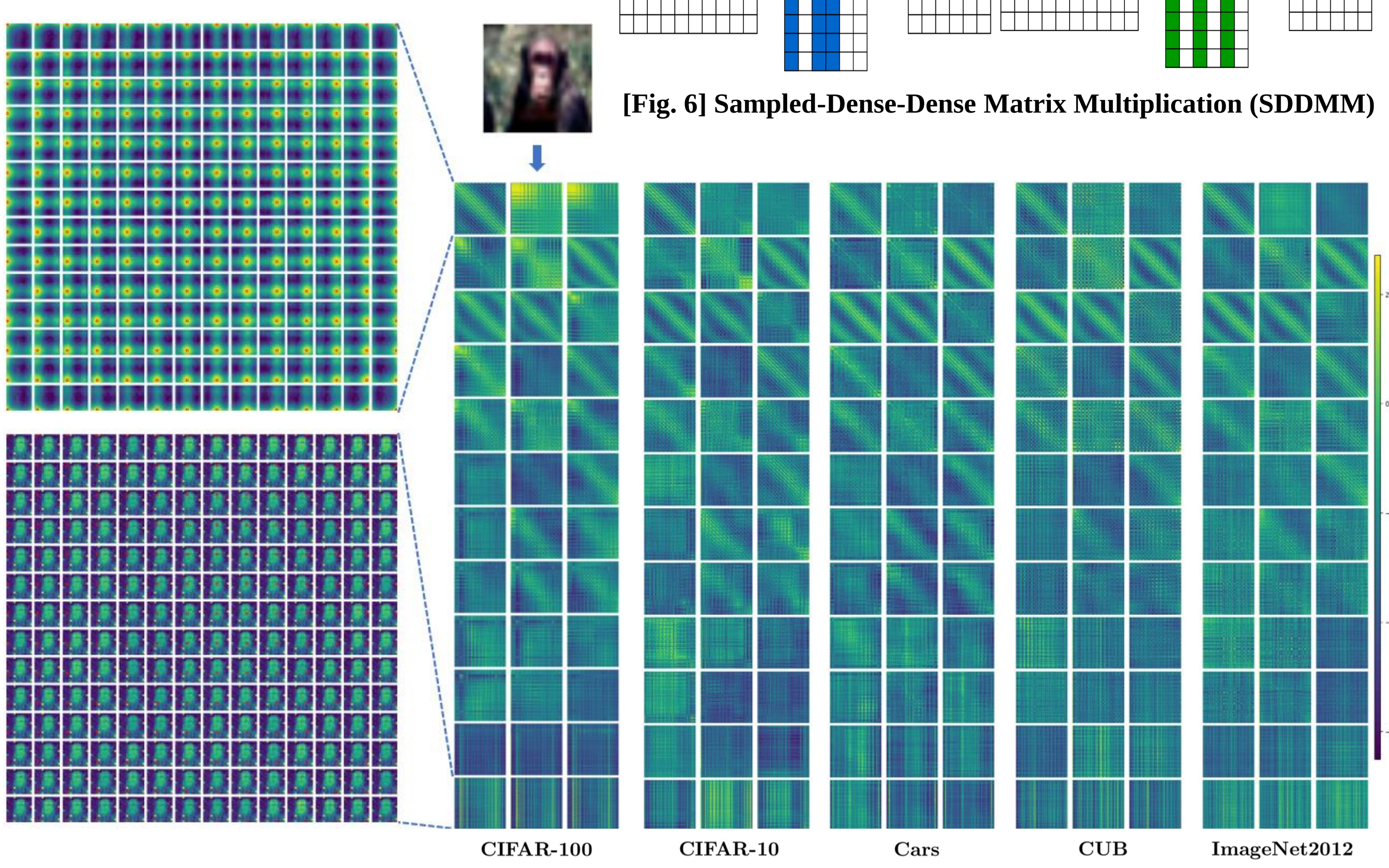
As shown in Fig. 5, CV tasks exhibit higher pruning ratios compared to NLP tasks, motivating the application of hardware-friendly structured pruning. Fig. 7 presents the analysis of attention maps across various datasets, showing that in the upper layers, attention scores are higher for self-connections and neighboring components. The core operation of the proposed algorithm, Fig. 6 Sampled-based Dense Dense Matrix Multiplication (SDDMM), is inefficient when implemented with systolic array structures. Therefore, we utilized parallel Mac Lines and adopted an output-stationary data flow.



[Fig. 5] Performance Comparison of NLP Transformers and ViTs Based on Pruning Ratios

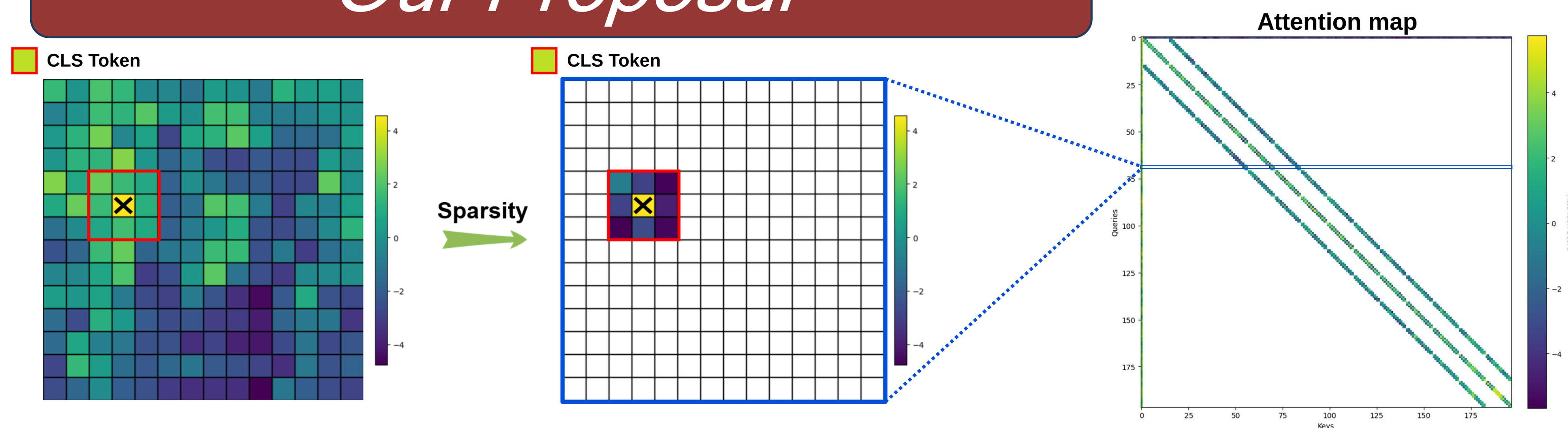


[Fig. 6] Sampled-Dense-Dense Matrix Multiplication (SDDMM)



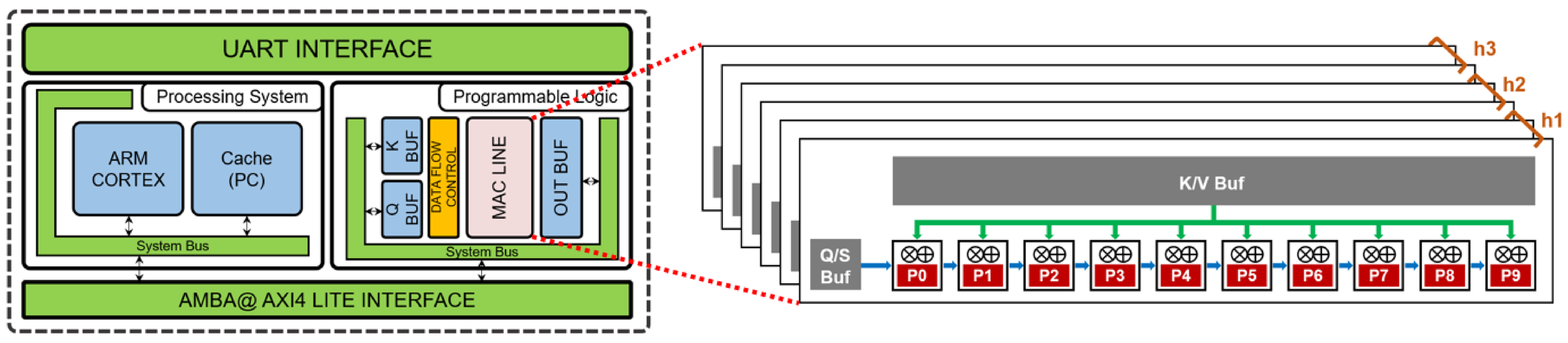
[Fig. 7] Visualizing the attention maps of all heads in DeiT-tiny on different transfer learning task

Our Proposal



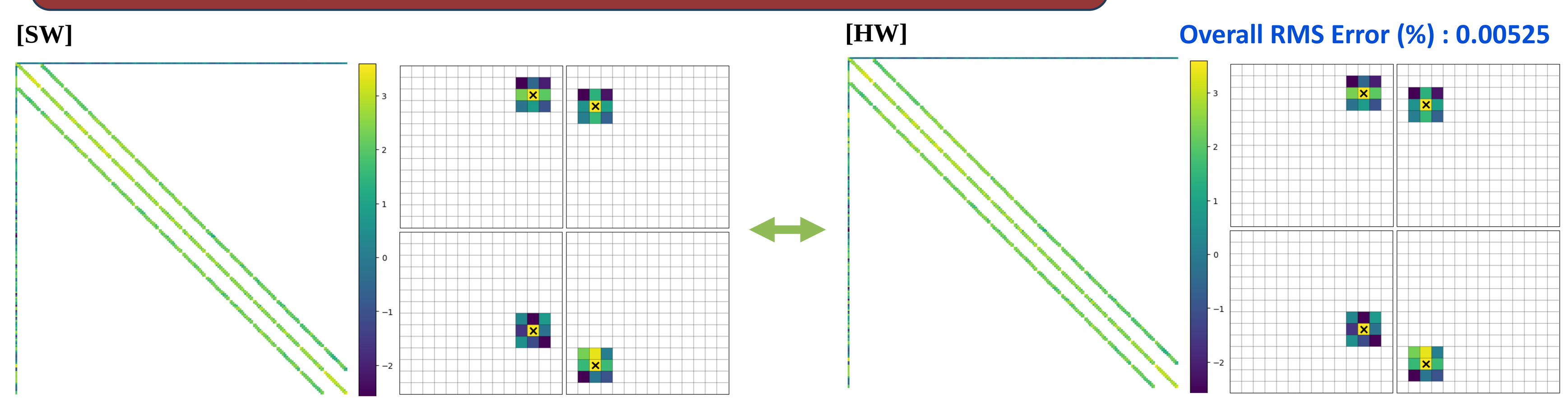
[Fig. 8] Our proposed algorithm, Windowed Attention

In our algorithm, we applied sparsity to the lower six blocks, while only the upper six blocks perform full attention computation. As shown in Fig. 8, sparsity is applied to the regions outside the 3×3 window in the 2D image for each query. Our algorithm enables up to 95% pruning of the attention map and applies hardware-friendly pruning, making it suitable for hardware acceleration.

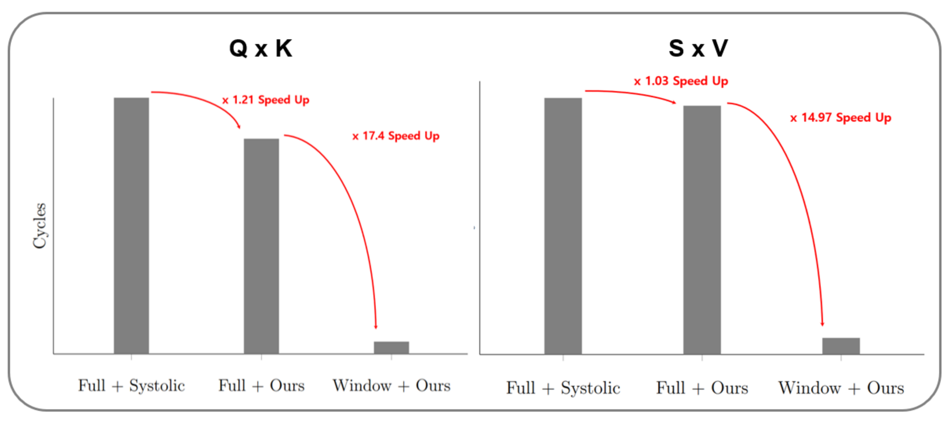


[Fig. 9] Overall Hardware Architecture for the Proposed System

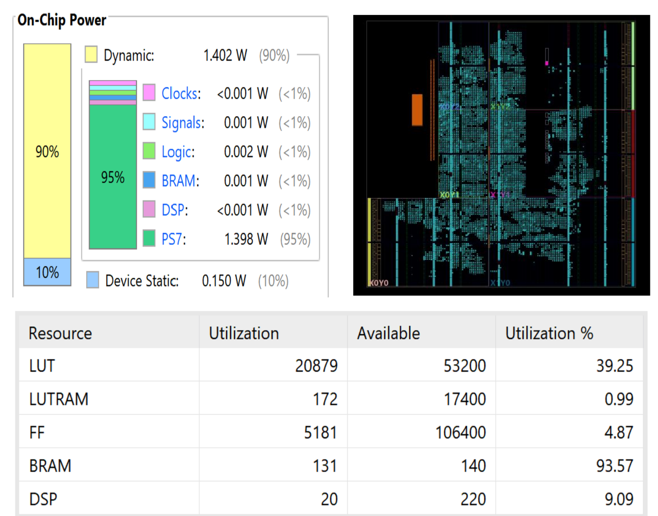
Experimental Result



[Fig.10] Comparison of Attention map: SW vs HW



[Fig. 11] Cycle Count: Full Attn vs Windowed Attn using systolic array and our accelerator with equal PEs



[Fig.12] FPGA Results

Table 1: Table 1: Accuracy on Various Transfer Learning Tasks

Method	CIFAR-100	CIFAR-10	Cars	CUB	ImageNet2012
Base	86.90%	97.90%	89.30%	82.40%	72.20%
Ours	86.00% (-0.9)	97.30% (-0.6)	88.50% (-0.8)	80.50% (-1.9)	70.10% (-2.1)

Conclusion

Despite their powerful performance, Transformers face challenges in hardware acceleration due to the high computational complexity of attention operations. However, unlike NLP, ViT offers opportunities for acceleration thanks to their fixed input token sizes and pruning ratios of up to 90%. Based on this, we analyzed attention maps across various transfer learning tasks and discovered that tokens in the lower layer tend to exhibit high attention scores with themselves and their neighboring patches in 2D image dimensions. Utilizing this observation, we designed an efficient algorithm to reduce computational overhead by performing attention operations only with adjacent elements in the image dimension. Furthermore, we designed a hardware accelerator to optimize the core computation of SDDMM. To validate the performance of this accelerator, we implemented our accelerator on the Zybo Z7 FPGA and conducted experiments in a real hardware environment, achieving significant improvements in computation speed. This approach is expected to make a significant contribution to computer vision applications where inference speed is crucial, particularly for edge devices with limited hardware resources.

Reference

[1] Y. -H. Chen, et al., "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," in IEEE, 2017.
[2] H. You, et al., "ViTCoD: Vision Transformer Acceleration via Dedicated Algorithm and Accelerator Co-Design," in IEEE , 2023.
[3] Ashish Vaswani, et al. "Attention is all you need," in NIPS, 2017.
[4] A. Dosovitskiy, et al., "An image is worth 16x16 words: Transformers for image recognition at scale", in ICLR, 2021.

경희대학교
KYUNG HEE UNIVERSITY