

Row-wise Accelerator for Vision Transformer

Hong-Yi Wang, and Tian-Sheuan Chang, *Senior Member, IEEE*
Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Abstract—Following the success of the natural language processing, the transformer for vision applications has attracted significant attention in recent years due to its excellent performance. However, existing deep learning hardware accelerators for vision cannot execute this structure efficiently due to significant model architecture differences. As a result, this paper proposes the hardware accelerator for vision transformers with row-wise scheduling, which decomposes major operations in vision transformers as a single dot product primitive for a unified and efficient execution. Furthermore, by sharing weights in columns, we can reuse the data and reduce the usage of memory. The implementation with TSMC 40nm CMOS technology only requires 262K gate count and 149KB SRAM buffer for 403.2 GOPS throughput at 600MHz clock frequency.

Index Terms—vision transformer, hardware design, accelerators

I. INTRODUCTION

Deep learning in computer vision has long been dominated by convolutional neural networks (CNNs) based backbone [1] for its revolutionary performance. On the other hand, natural language processing (NLP) use Transformer [2] that no longer relies on convolution but self-attention to model long-term data dependencies. Its enormous success makes it a viable competitor in computer vision tasks as well. As a result, various vision transformer models [3]–[6] have been developed that show competitive or even better performance than current CNN based models. However, real-time executions of these models suffer from high computational complexity and memory bandwidth, which demands hardware acceleration. However, existing deep learning hardware accelerators [7]–[9] are optimized for CNN based models, which is not suitable for vision transformer models due to the significant difference in the favored architecture structure and parameters. Directly executing vision transformer models on these accelerators will result in low hardware utilization.

To address the above issue, this paper proposes a hardware accelerator for vision transformers. To tailor the hardware design to vision transformers, we first analyze the common features of different models and compare them with existing CNN models. Based on this analysis, we propose a row-wise scheduling that uses the dot product as a primitive for major operations in the model to attain high hardware utilization. Besides, the proposed hardware broadcast weight for computations that can share weights for higher data reuse and smaller buffer size. The implementation on TSMC 40nm CMOS process can achieve 403.2GOPS throughput when running at 600Mhz clock frequency.

The remainder of the paper is structured as follows. Section II goes over the associated tasks. Section III examines the

common characteristics of several vision transformer models. The proposed design is shown in Section IV. Section V presents the experimental data as well as comparisons with other studies. Finally, Section VI brings this paper to a close.

II. RELATED WORK

A. Transformer

Transformer adopts the encoder and decoder architecture. The encoder consists of multiple encoding layers to generate the relationship between the inputs, while the decoder consists of multiple decoding layers that do the opposite. Fig. 1 shows an example of one layer of encoder. In which, an encoding layer consists of three kinds of blocks, self-attention, feed forward and normalization. The self-attention block will figure out the relationship between different inputs to generate output encodings as shown in (1).

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d_k})V \quad (1)$$

Then, each output encoding is further processed by the feed forward block that uses two successive fully connected layers with Gaussian error linear units (GELU) [10] as its activation function. Its output is normalized by the normalization block that uses layer normalization. Besides, they would use the skip connection in every block.

The fully connected layers in the vision transformer share their weight along the spatial dimension and apply fully connection on the channel dimension. Thus, these fully connected layers are the 1×1 convolutions. However, to keep naming consistency with other papers, the name of the fully connected layer is still used below.

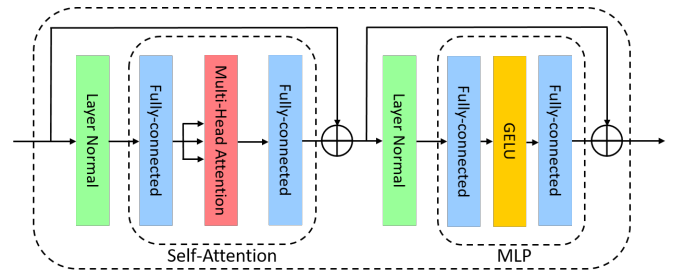


Fig. 1. Transformer encoder architecture

B. Vision Transformer

With Transformer's success in nature language processing, Transformer has also been introduced to the vision application [3], which gains competitive or better performance than previous convolutional neural network based models. For the

classification task on ImageNet, lots of vision transformers such as ViT [3], TNT [5], Swin [4] and PVT [6] use the encoder of Transformer. These models differ on the detailed structure of the encoder layer.

C. Hardware Accelerators for CNNs and Transformer in NLP

Various hardware accelerators have been proposed for CNN model execution. They exploit the high parallelism of CNN computation with hundreds of PEs for massive computation, and reuse input/weights/activation in different ways to reduce the external memory bandwidth requirement [7], [8], [11], [12]. The PEs are arranged in a 1D or 2D array, optimized for widely used 3×3 convolutions and reconfigured for other convolution types if needed but with lower hardware utilization. In contrast, 3×3 convolutions are seldom used in vision transformers. In vision transformers, their main structures don't involve CNN. They only use CNN in the beginning as a kind of feature extraction method for image pre-processing. Besides, the kernel size is 4×4 in vision transformers. Thus, if directly executing these models on these accelerators, their hardware utilization would be lower than 66%. Not to mention that these designs do not support multi-head attention.

Several works [13]–[15] have also proposed for transformer accelerator in NLP. No data reuse would make it perform poorly in vision. Additionally, the configuration of parameters in NLP ($sequence\ length \times channel$) is different from that in vision ($height \times width \times channel$). Thus, existing designs are not suitable for vision transformers.

III. MODEL ANALYSIS FOR VISION TRANSFORMERS

A. Common Features in Vision Transformer

Table I shows the common configuration parameters for different vision transformer models. First, since these models are for ImageNet classification, the input sizes for different layers are a multiple of 7 due to the 224×224 image size in ImageNet. Besides, their channel numbers were almost always a multiple of 96 except for PVT, which is a multiple of 64. These configuration numbers are quite different from CNN based classification models.

For the model structure, their backbones are the encoder of the transformer, which composes of multi-head attention blocks, residual connection, and fully connected layers. In addition, compared with the traditional CNN model as shown in Table II, activation function of vision transformers is GELU instead of ReLU. Moreover, their normalization functions were layer normalization instead of batch normalization.

Based on these analysis results, the hardware design will optimize for these common configurations and structures.

B. Analysis of FLOPs and Parameters Distribution in Vision Transformers

To facilitate the analysis, we decompose the model, Swin-T, into three layers according to their actual functions, including the convolution layer, fully connected layer, and multi-head attention layer. Fig. 2 shows the distribution analysis of FLOPs and parameters for different layer types .

TABLE I
COMMON FEATURES IN VISION TRANSFORMER MODELS

	features	channels	input size	conv size
ViT	first vision transformer	768	14	16
TNT	extra patches transformer	768	14	16
Swin	windows transformer shifted windows	96×	7×	4
PVT	image pyramid	64×	14	4×

TABLE II
COMPARISON BETWEEN VISION TRANSFORMER AND CNN BASED MODEL

	Vision Transformer	CNN
Activation function	GELU	ReLU
Normalization	layer normal	batch normal
Main calculation	fully-connected	convolution
Similarity	short cut (residual)	

From the viewpoint of layer types as shown in Fig. 2, more than 97% of the FLOPs and more than 83% of number of parameters are occupied by the fully connected layer. Both the convolution layer and multi-head attention layer only occupy a very small fraction. Thus, the design of computation unit and buffer size should optimize for the fully connected layer.

IV. PROPOSED ARCHITECTURE

A. Row-wise Scheduling

For the three major layers, multi-head attention and convolution are matrix multiplication, while the fully connected layer is a matrix vector product. To unify these computation types and fit different configuration parameters, we decompose all these operations as a more basic dot product to build our processing element (PE) block as shown in Fig. 3.

A PE block contains 7 PE rows in a block and each row has 4 multiplier-and-accumulator (MAC) units. In this paper, we use 12 PE blocks for model execution. These numbers are selected to fit the common configuration parameters as shown before. Thus, one row can compute a dot product on 2 vectors with both sizes of 4, which can fit different operation requirements due to the decomposition. This also enables a very high hardware utilization and simple dataflow to optimize the overall computing time.

In a PE block, each weight data are broadcast to all multipliers from top to bottom to share weight. Each MAC will receive

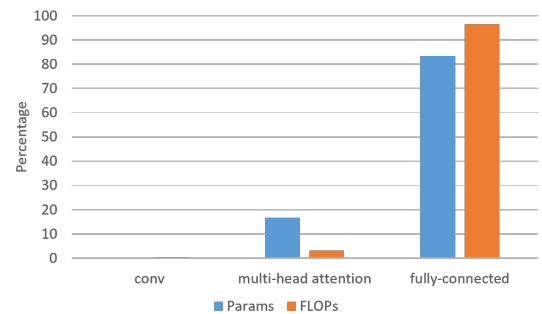


Fig. 2. Distributions of parameters and FLOPs in Swin-T for different layers.

different inputs from the input SRAM to support the fully connected layer. The multiplication results are accumulated in a horizontal direction and saved to the local buffer for later more accumulation.

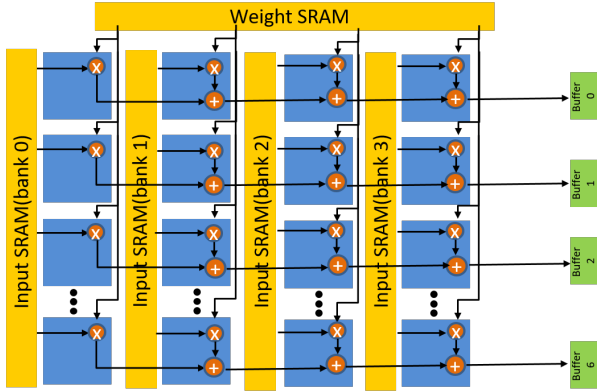


Fig. 3. A PE block

B. Overall Architecture

Fig. 4 shows the proposed overall architecture. The result from each PE block is accumulated in an accumulator block and further summed together with the adder tree for the large number of input channel cases. Then the result will be post-processing by layer normalization or softmax depending on the layer types. The final result will be sent to off-chip memory.

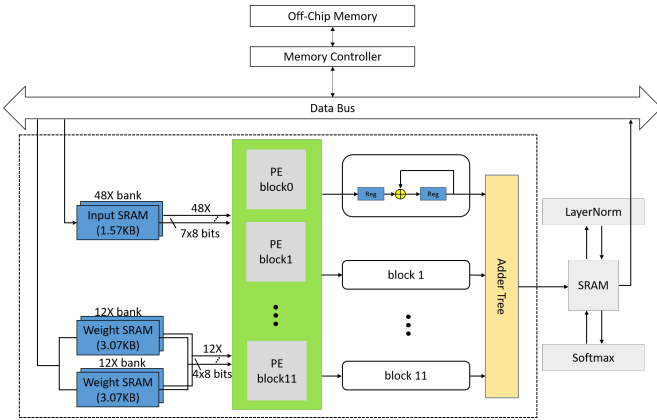


Fig. 4. Proposed overall architecture for vision transformer

C. Convolution Layer

The only convolution in vision transformers is at the first layer, which uses 4x4 convolution with stride 4 to reduce the input size into 1/4 of the original. The size of each kernel is 4x4x3, which can be perfectly placed into each PE weight block, as shown in Fig. 5(a). The input is RGB images. Each input channel and its corresponding weight will be placed on four PE blocks, respectively. During each computing cycle as in Fig. 5(b), all 7 input rows will be activated at the same time to process all input to generate 7 outputs in a cycle. The computing order of convolution will process the first output

channel and then the second one until the final 96th channel. For the 224x224x3 input size in ImageNet, our design would calculate a 28x4x3 input in a cycle and take 448 cycles to finish one output channel.

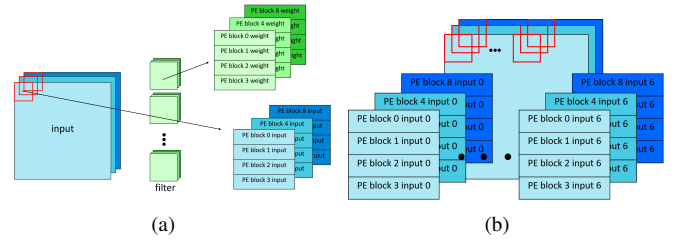


Fig. 5. (a) Input and weight mapping to PE blocks for one output convolution channel. (b) Detailed mapping for input of PE blocks

D. Fully-Connected Layer

By row-wise scheduling, different sizes of fully connected layers can become a dot product in a single row. However, since the number of input channels are multiple of 96 and exceeds the hardware capability in one cycle, its output will be accumulated in the accumulator for the final output.

Fig. 6(a) and Fig. 6(b) show the detailed calculation for 96 channels. In the first cycle, the 1st to 48th channels of input and filters would be placed at the corresponding position from PE block 0 to 11. The second cycle will process the rest of channels. Besides, with 7 input rows in a PE block, we would finish 7 outputs in every 2 cycles for 96 channels of input. Then, the computing order of the fully connected layer will process the first output channel and then the second one until the final output channel.

PE Block 0				PE Block 0			
$w_{0,0}$	$w_{0,1}$	$w_{0,2}$	$w_{0,3}$	$w_{0,48}$	$w_{0,49}$	$w_{0,50}$	$w_{0,51}$
$I_{0,0,0}$	$I_{1,0,0}$	$I_{2,0,0}$	$I_{3,0,0}$	$I_{48,0,0}$	$I_{49,0,0}$	$I_{50,0,0}$	$I_{51,0,0}$
$I_{0,0,1}$	$I_{1,0,1}$	$I_{2,0,1}$	$I_{3,0,1}$	$I_{48,0,1}$	$I_{49,0,1}$	$I_{50,0,1}$	$I_{51,0,1}$
$I_{0,0,2}$	$I_{1,0,2}$	$I_{2,0,2}$	$I_{3,0,2}$	$I_{48,0,2}$	$I_{49,0,2}$	$I_{50,0,2}$	$I_{51,0,2}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$I_{0,0,6}$	$I_{1,0,6}$	$I_{2,0,6}$	$I_{3,0,6}$	$I_{48,0,6}$	$I_{49,0,6}$	$I_{50,0,6}$	$I_{51,0,6}$

(a) cycle 1

(b) cycle 2

Fig. 6. Detailed input and weight mapping of 7 inputs for a fully-connected layer with 96 channels at (a) the first cycle and (b) the second cycle.

E. Multi-Head Attention Layer

With the row-wise scheduling and the window multi-head self-attention(W-MSA) concept as in Swin-Transformer, we can make the computation of multi-head attention layers regular and simple. In this case, for the multiplication of Q and K as in (1), Q could be regarded as the weight and K as the input for PE blocks. In this case, since the size of the

TABLE III
COMPARISON TO OTHER CNN ACCELERATORS

	Ours	IECA [17]	Eyeriss v2 [8]
Technology(nm)	40	55	65
Model Types	Trans.	CNN	CNN
PE Number	336	168	92
Clock Rate(MHz)	600	250	200
Peak Throughput(GOPS)	403.2	84	153.6
Area (KGE)	186	344.7	2695
SRAM(KB)	149	109	192

matrix is smaller than other layer types, it will only use 8 PE blocks. Thus, for Q matrix mapping, four columns of Q matrix will be allocated to one PE block and processed in a row-by-row manner. Each row will take 7 cycles. For K^T matrix mapping, they are allocated to a group with 7 input rows \times 8 PE blocks for computation, and processed in a group-by-group manner. During this process, since only 8 of the 12 PE blocks, the hardware utilization will be lower. However, the flops of the multi-head attention layer would take no more than 3% in whole model, and such hardware utilization impact would impact no more than 1% of total cycles.

V. EXPERIMENTAL RESULTS

This design has been implemented with TSMC 40nm CMOS technology and needs 262K NAND gate count and 149KB SRAM buffer to execute vision transformer models, especially optimized for Swin Transformer. The precision of weights and activations are both 8 bits. For area, SRAM occupies 93.2% of area, and PE, SRAM controller, and accumulator occupy 4.6%, 1.5%, 0.6% of area respectively. The peak performance is about 403.2GOPS at 600MHz clock frequency. The average time to process an 224×224 image with Swin-T is about 22.4 ms, equivalent to 44.5 images/s throughput. Theoretically, overall hardware utilization could be as high as 99% or higher.

Comparisons with other designs are not easy due to different target models. For reference, we list other CNN hardware accelerators in Table III. Our peak throughput is higher than others due to more PE numbers and higher clock frequency. However, our area cost is smaller than others due to its simple PE structure.

Table IV shows the throughput comparison to execute Swin-T with different processors and designs. GPU is a NVIDIA GeForce RTX 2080 Ti, which has 4352 CUDA Cores and a total of about 9.8MB of cache memory. Compared to GPU, our throughput is $1.1 \times$ than that of GPU, respectively. This higher throughput is achieved by our high hardware utilization. Another design is based on FPGA. Vis-Top [16] is $1.9 \times$ faster than us, but little information has been disclosed in this paper. This higher throughput is possible due to larger PEs based on its DSP slices utilized.

VI. CONCLUSION

This paper proposes an efficient design for vision transformers that is optimized for encoding layer operations instead of CNN. This design can efficiently execute different layer

TABLE IV
COMPARISON TO OTHER PROCESSORS AND DESIGNS.

	GPU	Ours (ASIC)
Throughput (image/s)	41.5	44.5
Relative Speedup	1	1.07
Cache Memory(KB)	9852	149
Throughput per MACs	0.0095	0.1324

operations by row-wise scheduling for high hardware utilization. Our design retains high throughput while maintaining area efficiency. The final implementation with TSMC 40nm CMOS process needs 262K NAND gate counts and 149KB SRAM buffer and achieves 403.2 GOPS peak throughput when running at 600MHz clock frequency.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [2] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [3] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [4] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.
- [5] K. Han *et al.*, "Transformer in transformer," *arXiv preprint arXiv:2103.00112*, 2021.
- [6] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.
- [7] Y.-H. Chen *et al.*, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE journal of solid-state circuits*, vol. 52, no. 1, pp. 127–138, 2016.
- [8] Y.-H. Chen *et al.*, "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 292–308, 2019.
- [9] A. Biswas and A. P. Chandrakasan, "Conv-sram: An energy-efficient sram with in-memory dot-product computation for low-power convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, 2018.
- [10] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
- [11] Y. Chen *et al.*, "Dadiannao: A machine-learning supercomputer," in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, 2014, pp. 609–622.
- [12] T. Chen *et al.*, "Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," *ACM SIGARCH Computer Architecture News*, vol. 42, no. 1, pp. 269–284, 2014.
- [13] S. Lu *et al.*, "Hardware accelerator for multi-head attention and position-wise feed-forward in the transformer," in *2020 IEEE 33rd International System-on-Chip Conference (SOCC)*. IEEE, 2020, pp. 84–89.
- [14] B. Li *et al.*, "Ftrans: energy-efficient acceleration of transformers using fpga," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, 2020, pp. 175–180.
- [15] T. J. Ham *et al.*, "A³: Accelerating attention mechanisms in neural networks with approximation," in *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2020, pp. 328–341.
- [16] W. Hu *et al.*, "Vis-top: Visual transformer overlay processor," *arXiv preprint arXiv:2110.10957*, 2021.
- [17] B. Huang *et al.*, "IECA: An in-execution configuration CNN accelerator with 30.55 GOPS/mm² area efficiency," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 11, pp. 4672–4685, 2021.