Fast Vision Transformer via Additive Attention

Yang Wen
College of Electronics and Information Engineering
Shenzhen University
Shenzhen, China

Samuel Chen
School of Artificial Intelligence
Xidian University
Xi'An, China

Abhishek Krishna Shrestha
School of Artificial Intelligence
Xidian University
Xi'An, China

***Abstract*—The Vision Transformer has been a more effective architecture for computer vision tasks than convolutional neural networks (CNN). However, it is time-consuming due to the quadratic complexity of the input sequence length. In this paper, a Fast Vision Transformer (FViT) is proposed based on an additive attention module, which reduces computation complexity to linearity. The experiment results show that the proposed model achieves faster inference with less memory.**

***Index Terms*—Fast Vision Transformer, Additive Attention**

I. Introduction

Recently, a vision transformer (ViT) [4] has been proposed as a backbone for image classification as well as object detection, where the attention mechanism [1] is the core module leading to outstanding performance compared with convolutional neural networks (CNNs) [5]. However, the attention module is time-consuming due to quadratic computational complexity. It is quite important to strike a balance between performance and computational efficiency. To solve this issue in transformer architecture, a Fastformer is proposed based on additive attention, which compacts the query sequence into a global query vector. This exploits the interaction between the global queries and attention keys with the element-wise product and a linear transformation is applied to learn global context-aware attention values.

Inspired by additive attention, in this paper, a fast vision transformer is proposed based on the additive attention module to strike a trade-off between the computational efficiency and performance of the vision transformer. The proposed model exploits the global contexts of a sequence and obtains token representation via its interaction with the global context. The experimental results on the ImageNet dataset show that the proposed model gets comparable performance with the ViT in less inference time.

II. Proposed Method

A. Vision Transformer

The ViT [4] model is proposed based on multi-head self-attention [1], which exploits the correlations between inputs at a pair of positions and captures the contexts within a sequence.

Given an image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, a sequence of N flattened 2D patches $x_p \in \mathbb{R}^{N\times(P^2.C)}$ are extracted, where (H, W) is the original image's resolution, C is the number of channels, (P, P) is the size of each image patch. A linear projection is applied to embed them into D-dimension space. These projected image patches will go through a multi-head self-attention module, where each head can be expressed as follows.

$$\mathbf{head}_i = \text{Attention}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V) \quad (1)$$

$$\text{Attention} = \text{softmax}(\frac{\mathbf{QW}_i^Q(\mathbf{KW}_i^K)^T}{\sqrt{d}})\mathbf{VW}_i^V \quad (2)$$

It is shown from Eq.(1) and (2) that the self-attention module is $O(n^3)$ complexity. In this paper, an additive attention module is proposed to reduce the computational complexity.

This work was supported by the National Natural Science Foundation of China (No. 62301330,62101346), the Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515110101, 2023112110380701).

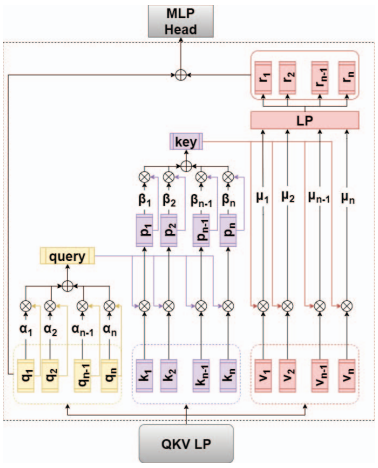

Fig. 1: Fastformer Architecture

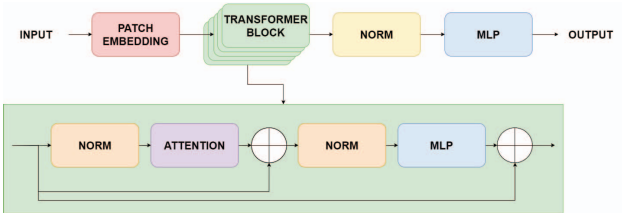

Fig. 2: The architecture of Additive attention-based Encoder

B. Additive Attention Module

In this section, the additive attention module is introduced and its architecture is illustrated in Fig.1. As the ViT, the

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00113

sequence of transformed feature vectors from image patches, known as image tokens, as well as the prediction token will go through three encoding blocks which generates three matrices $\mathbf{Q} = [\mathbf{q_1}, \mathbf{q_2}, ..., \mathbf{q_N}], \mathbf{K} = [\mathbf{k_1}, \mathbf{k_2}, ..., \mathbf{k_N}]$ and $\mathbf{V} = [\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_N}]$, known as the query, key and value, respectively. They exploit the context information of the input image.

An additive attention module is applied to exploit the query matrix into a global query vector $\mathbf{q} \in \mathbb{R}^{\mathbf{d}}$ in linear complexity. The attention weight $\alpha_\mathbf{i}$ of the i-th query vector is calculated in the following way:

$$\alpha_i = \frac{\exp(\mathbf{w}_q^T \, \mathbf{q}_i/\sqrt{d})}{\sum_{j=1}^{N} \exp(\mathbf{w}_q^T \, \mathbf{q}_j/\sqrt{d})} \tag{3}$$

where $\mathbf{w}_q$ is a parameter vector that can be calculated as

$$\mathbf{q} = \sum_{i=1}^{N} \alpha_i \mathbf{q}_i \tag{4}$$

Then an element-wise product between the global query vector and each key vector integrates them into a global context-aware key matrix. The i-th vector in this matrix is denoted by $\mathbf{p_i}$, written as $\mathbf{p}_i = \mathbf{q} * \mathbf{k}_i$. For computational efficiency, the i-th vector's additive attention weight is calculated as follows:

$$\beta_i = \frac{\exp(\mathbf{w}_k^T \, \mathbf{p}_i/\sqrt{d})}{\sum_{j=1}^{N} \exp(\mathbf{w}_k^T \, \mathbf{p}_j/\sqrt{d})} \tag{5}$$

where w_k is a parameter vector that can be learned. The following is how the global key vector is calculated:

$$\mathbf{k} = \sum_{i=1}^{N} \beta_i \mathbf{p}_i \tag{6}$$

Finally, for better context exploitation, we model the relationship between the attention value matrix and the global key vector. We calculate a key-value interaction vector $\mathbf{u}_i$ by performing an element-wise product between the global key and value vector, which is expressed as $\mathbf{u}_i = \mathbf{k} * \mathbf{v}_i$, similarly to query-key interaction modeling. We add a linear transformation layer to each key-value interaction vector to learn its hidden representation. The output from this layer is denoted as $\mathbf{R} = [\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_N}] \in \mathbb{R}^{N \times d}$ which is further added with query matrix to form the final output of the model.

TABLE I: Model Comparison

Model	Top-1	Param (M)	GFLOPs
ViT-B/16	77%	86	49.3
Fastformer-B/16	63%	79	45.2
ViT-B/32	73%	88	12.6
Fastformer-B/32	59%	81	11.6
Swin-Transformer(32)	81%	88	47

III. Experiments

A. Experiment Configuration

In this paper, the ILSVRC-2012 ImageNet [3] dataset is used for the experiments. The validation dataset of ImageNet consists of 50000 images across 1000 categories.

Fastformer (b_32 and b_16 variations) is compared to the ViT with the variants of B/16 and B/32. B/16 and B/32 have a hidden dimension size of 768, MLP dimension of size 3072. According to the configuration, the patch size for the model is 16×16 and 32×32 respectively. The number of heads is 12 and the depth is set as 12. Training configurations included a batch size of 64, a learning rate of 0.0005, and training for 50 epochs.

B. Experiment Analysis

The comparison results are listed in Table I. The B/16 variant achieves 77% Top-1 accuracy, which is better than Fastformer-B/16 with 63%. But Fastformer-B/16 with 79M has less number of parameters than ViT-B/16 with 86M parameters. The Fastformer-B/16 with 45.2 GFLOPs has less computational complexity than ViT-B/16 with 49.3 GFLOPs.

In the B/32 variants, ViT-B/32 achieves 73% Top-1 accuracy, while Fastformer-B/32 65%. But Fastformer-B/32 with 81M parameters has fewer parameters than ViT-B/32 with 88M parameters. The computational cost of Fastformer-B/32's 11.6 GFLOPs is less complex than ViT-B/32 with 12.6 GFLOPs.

IV. Conclusion

In this paper, an additive attention module was proposed to accelerate the computation of self-attention. The experimental results show that the proposed method with fewer FLOPs obtains comparable performance with the ViT.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," arXiv (Cornell University), vol. 30, pp. 5998–6008, Jun. 2017.

[2] C. Wu, F. Wu, T. Qi, Y. Huang, and X. Xie, "Fastformer: additive attention can be all you need," arXiv (Cornell University), Aug. 2021.

[3] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[4] 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," arXiv (Cornell University), Oct. 2020.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386. Available: https://doi.org/10.1145/3065386

[6] Y. Liu et al., "A survey of visual transformers," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–21, Jan. 2023, doi: 10.1109/tnnls.2022.3227717. Available: https://doi.org/10.1109/tnnls.2022.3227717