2024 IEEE Conference on Artificial Intelligence (CAI)

Optimized Vision Transformer Training using GPU and Multi-threading

Jonathan Ledet[†], Ashok Kumar[†], Dominick Rizk[‡], Rodrigue Rizk[⋆], KC Santosh[⋆]
[†]*Center for Advanced Computer Studies*, *University of Louisiana at Lafayette*, Lafayette, LA, USA
[‡]*Department of Electrical Engineering*, *and Computer Science*, *Catholic University of America*, Washington, DC, USA
[⋆]*Applied AI Research Lab*, *Department of Computer Science*, *University of South Dakota*, Vermillion, SD, USA
jonathan.ledet1@louisiana.edu, ashok.kumar@louisiana.edu, rizkd@cua.edu, rodrigue.rizk@usd.edu, kc.santosh@usd.edu

***Abstract*—Traditional Convolutional Neural Networks (CNNs) often struggle with capturing intricate spatial relationships and nuanced patterns in diverse datasets. To overcome these limitations, this work pioneers the application of Vision Transformer (ViT) models which have gained significant attention in the field of computer vision for their ability to capture long-range dependencies in images through self-attention mechanisms. However, training large-scale ViT models with a massive number of parameters poses computational challenges. In this paper, we present an optimized approach for training ViT models that leverages the parallel processing capabilities of Graphics Processing Units (GPUs) and optimizes the computational workload distribution using multi-threading. The proposed model is trained and tested on the CIFAR-10 dataset and achieved an outstanding accuracy of 99.92% after 100 epochs. The experimental results demonstrate the effectiveness of our approach in optimizing training efficiency compared to existing methods. This underscores the superior performance of ViT models and their potential to revolutionize image classification tasks.**

***Index Terms*—CIFAR-10 dataset, convolutional neural networks (CNN), GPU, image classification, multi-threading, vision transformer (ViT), attention mechanism**

I. Introduction

Traditional Convolutional Neural Networks (CNNs) have been pivotal in the domain of image classification, demonstrating significant success. However, their inherent limitations in capturing long-range dependencies and global context have spurred exploration into innovative architectures. Vision Transformers (ViTs), a novel paradigm in computer vision, aim to overcome these limitations by adapting the transformer architecture to process image data. ViTs have demonstrated state-of-the-art performance in various computer vision tasks, but their training demands substantial computational resources. Previous research [1] - [4] has explored parallelizing deep neural network training using GPUs. However, adapting these techniques specifically to ViTs requires novel considerations due to the unique self-attention mechanisms employed in ViT architectures. In this work, we address the challenge of efficiently training large-scale ViT models by leveraging the parallel processing power of GPUs and optimizing the training pipeline with multi-threading. Our proposed approach aims to reduce training time while maintaining or even improving the model's accuracy.

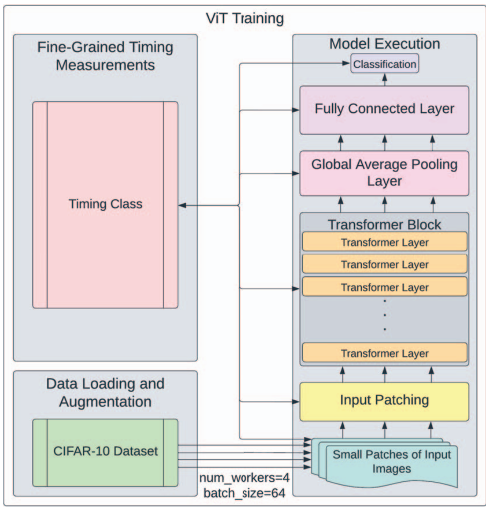

Fig. 1. Proposed Optimized Vision Transformer Training Pipeline.

II. Multi-threading in ViT Training

The incorporation of multi-threading into the Vision Transformer (ViT) training program represents a strategic optimization to elevate the efficiency of key components in the training pipeline. By embracing concurrent execution strategies, the training process is enhanced, leading to accelerated convergence and improved overall performance. Fig. 1 illustrates our proposed approach for efficient Vision Transformer training, combining GPU acceleration and multi-threading techniques. Data parallelism across different modules optimizes computation, while multi-threading streamlines tasks, collectively reducing training time.

Data Loading and Augmentation: One crucial phase where multi-threading plays a pivotal role is during the data loading and augmentation process. Leveraging concurrent execution, the `torch.utils.data.DataLoader` efficiently manages the retrieval and preprocessing of batches from the CIFAR-10 dataset. Through the use of multiple worker threads, data loading becomes an asynchronous operation, mitigating potential bottlenecks and substantially enhancing the throughput of the training process. This optimization ensures that the

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00189

TABLE I
VIT LAYER PROCESSING TIMES

Layer	Processing Time (ms)
Input Patching	0.97
Positional Encoding	0.99
Transformer Encoder	15.96
Linear Embedding	0.99
Global Average Pooling	0.02
Fully Connected Layers	0.01
Output Layer	0.11
Total	19.05

model is consistently fed with a stream of diverse and augmented data, contributing to the robustness and generalization capabilities of the Vision Transformer.

Fine-Grained Timing Measurements: To gain a nuanced understanding of the ViT model's performance, a dedicated `Timing` class is employed to conduct fine-grained timing measurements. This class records the execution time of critical sections within the model, including input patching, positional encoding, transformer encoding, linear embedding, global average pooling, fully connected layers, and the output layer. These detailed measurements offer insights into how computational resources are allocated across different components, allowing for the identification of potential optimization opportunities. The fine-grained nature of these timings is instrumental in fine-tuning the model and ensuring efficient resource utilization.

III. Experimental Setup, Results, and Discussion

To evaluate the proposed optimized ViT training approach, we conducted experiments on the well-established dataset CIFAR-10 [5], a widely used benchmark in the field of computer vision. CIFAR-10 consists of 60,000 32x32 color images across ten classes, making it suitable for assessing the performance and efficiency of our optimized ViT training approach. Harnessing the power of GPU acceleration and multi-threading, the experimental platform boasts the following hardware specifications: CPU – Intel(R) Core(TM) i5-8600K CPU @ 3.60GHz; GPU – NVIDIA GTX 1060 6GB; Operating System – 64-bit Windows 10 Home; and RAM – 32GB DDR4 RAM 3000. The training strategy encompasses key parameters for achieving optimal results: Optimizer – Adam; Learning Rate – 0.0001; Loss Function – CrossEntropyLoss; Batch Size – 64; and Data Augmentation – Basic normalization. Experiments were executed over a sufficient number of epochs to allow the models to converge, with performance metrics recorded at regular intervals for comprehensive analysis. The source code for the model implementation is available on GitHub for reproducibility[1].

Table I details the processing times for each layer in the ViT model. Notably, the "*Transformer Encoder*" layer, responsible for capturing long-range dependencies, consumes the majority of the processing time (15.96 milliseconds). The "*Input Patching*" and "*Positional Encoding*" stages contribute significantly as well. The total processing time for the ViT is 19.05 milliseconds, reflecting the computation-intensive nature of self-attention mechanisms in large-scale vision models.

TABLE II
TEST ACCURACIES COMPARISON

# of Epochs	CNNs	Transformers	ViT
10	72.18%	68.44%	66.06%
25	71.17%	69.29%	84.73%
50	70.41%	71.62%	98.18%
100	70.13%	71.86%	99.92%

We compare the accuracy and convergence speed of our optimized ViT models against traditional training methods. Table II provides a comparative analysis of test accuracies for CNNs, Transformers, and ViTs at different training epochs on the CIFAR-10 dataset. At the initial 10 epochs, CNNs lead with the highest accuracy at 72.18%, outperforming both Transformers and ViTs. However, ViTs steadily outpace CNNs and Transformers as training progresses, achieving a notable 98.18% accuracy at 50 epochs and an impressive 99.92% accuracy after 100 epochs. These results highlight the consistent and substantial performance gain of ViTs, showcasing their effectiveness in capturing intricate relationships within images over an extended training duration.

IV. Conclusion

In this paper, we presented a novel methodology for optimizing Vision Transformer training using GPU acceleration and multi-threading. Our experiments revealed that, with prolonged training durations and the integration of multi-threading and GPU acceleration, ViTs outperformed both classical CNNs and Transformers in image classification tasks. Our approach significantly reduces training time, making large-scale ViT models more practical for a broader range of researchers and practitioners in the field of computer vision. Our experimental results demonstrated the efficiency of ViTs, especially with multi-threading and GPU acceleration, showcasing their potential for image classification tasks. The results underscore the adaptability and parameter efficiency of ViTs, especially in handling varying-sized inputs.

References

[1] 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. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.

[2] H. Seong-Hyeon and L. Kwang-Yeob. 2017. Implementation of image classification CNN using multi thread GPU. In 2017 International SoC Design Conference (ISOCC). 296–297.

[3] P. Sundareson, "Parallel image pre-processing for in-game object classification," 2017 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Bengaluru, India, 2017, pp. 115-116, doi: 10.1109/ICCE-ASIA.2017.8309316.

[4] R. Rizk, D. Rizk, F. Rizk, A. Kumar and M. Bayoumi, "A Resource-Saving Energy-Efficient Reconfigurable Hardware Accelerator for BERT-based Deep Neural Network Language Models using FFT Multiplication," 2022 IEEE International Symposium on Circuits and Systems (ISCAS), Austin, TX, USA, 2022, pp. 1675-1679, doi: 10.1109/ISCAS48785.2022.9937531.

[5] A. Krizhevsky, V. Nair, and G. Hinton,(2014) The CIFAR-10 Dataset.

[1] https://github.com/2ai-lab/Optimized-Vision-Transformer-Training-using-GPU-and-Multi-threading