The Investigation of Data-Parallelism Strategy Based on VIT Model

Zhengtao Feng

Software Engineering, Nanhai Campus, South China Normal University, Taoyuan East Road, Nanhai District, Foshang, 528200, China.

Email:20202005347@m.scnu.edu.cn

Abstract. With the advent of advanced techniques for training large models, the research community has become increasingly interested in exploring various methods to enhance the efficiency of model training. The Vision Transform (VIT) model represents a novel approach in the field of image processing, being the first attempt to apply the Transformer model in this domain. This study employs a repeated experimental methodology with data-parallelism to investigate the conditions and settings that optimize the training efficacy of the VIT model. Data-parallelism, a distributed parallel approach, is utilized to evenly distribute training tasks across multiple GPUs, allowing for a comparative assessment of the training effects. By manipulating fundamental configurations and adjusting the number of GPUs, the objective of achieving the most favorable training outcomes is pursued. Subsequently, this research endeavors to determine the optimal GPU configuration for training the CIFAR10 dataset using the VIT model. The experimental findings suggest that employing three GPUs yields the best results when training the CIFAR10 dataset with the VIT model. Specifically, the employment of three GPUs results in the most notable decrease in loss value and the highest accuracy in image classification. Consequently, the training effectiveness surpasses that of alternative experiments. Furthermore, in comparison to training the VIT model without data-parallelism, utilizing data-parallelism with any number of GPUs proves to enhance the efficiency of VIT model training.

Keywords: Vision Transformer Model, Data Parallelism, Distributed Model Training

1. Introduction

In light of the advent of the big data era, the development of large scale models has gained significant attention as a means to effectively train extensive datasets. Following the introduction of the Transformer model, researchers found that the Transformer, based solely on attention mechanisms [1], can dispense with recurrence and convolutions entirely and this

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_61

models is superior in quality while requiring significantly less time to train and being more parallelizable by experiments on two machine translation. The Transformer model has emerged as the preferred choice in Natural Language Processing (NLP), where it is commonly pre-trained on large-scale text corpora and subsequently fine-tuned on task-specific datasets [2]. This success has spurred significant interest in the field of Computer Vision with the introduction of the Vision Transformer (VIT) model. Recognized as a cutting-edge model, VIT has achieved remarkable performance in image classification tasks by leveraging a self-attention mechanism to extract and represent features with exceptional proficiency, surpassing the capabilities of traditional Convolutional Neural Networks (CNNs). By replacing convolutional layers with self-attention layers, VIT enables the capture of contextual relationships among image patches, facilitating the acquisition of more effective image representations. Consequently, VIT represents a transformative adaptation of the Transformer model for image training, wherein images are partitioned into numerous patches, and a sequence of linear embeddings of these patches is employed as input to the VIT model.

The VIT model has been applied in many fields. For instance, the VIT model can be used for screen image recognition for recapture [3]. This study explores the distinct characteristics of CNNs and VITs in feature extraction and proposes a cascaded network architecture that integrates local-feature and global-feature extraction modules for the purpose of detecting recaptured screen images from original images, regardless of the presence or absence of dismantle operations [3]. Furthermore, the VIT model has been successfully applied to Weakly Supervised Object Localization (WSOL), where the objective is to predict object locations in an image using only image-level category labels [4]. The applicability of the VIT model extends to the medical domain as well. A recent investigation [5] introduces a computer-aided diagnosis (CAD) method utilizing the Vision Transformer for the analysis of optical coherence tomography (OCT) images, enabling the automatic discrimination of Age-related Macular Degeneration (AMD), Diabetic Macular Edema (DME), and normal eyes. The results demonstrate that the Vision Transformer presents a superior alternative for more accurate diagnosis of retinal diseases. Moreover, the VIT model demonstrates its utility in the field of biology. Building upon the foundation of the Vision Transformer, a deep learning approach [6] is proposed to identify viral diseases in cassava leaf images. Nevertheless, as the complexity of the fields in which the VIT model is employed increases, the importance of efficient model training becomes paramount. Thus, this paper considers the potential for enhancing the training efficiency of the VIT model to facilitate its broader application across diverse domains.

In order to improve the efficiency for training images in vit models, this paper uses the technology called data-parallelism to explore in which situation vit model will achieve the best image training effect. This paper compared the multiple experimental data when VIT model uses the data-parallelism and not use it to find out if data-parallelism is useful for image training. This paper also compares the experimental data which VIT model in data-parallelism training by different numbers of GPUs to explore how many GPUs can improve the efficiency

of VIT training best. Meanwhile, the VIT model is also a very suitable model for parallel training. The principle of the research is that by comparing the experimental results of the VIT model with and without data parallelism, as well as data parallelism achieved by multiple GPUs to observe the training experiment effect.

2. Method

2.1 Dataset Description and Preprocessing

The dataset utilized in this research comprises the CIFAR-10 dataset, which consists of curated subsets extracted from the vast collection of 80 million tiny images. The CIFAR-10 dataset encompasses 60,000 color images with dimensions of 32×32 pixels, meticulously categorized into 10 distinct classes, each containing 6,000 images. Specifically, the dataset is segregated into two main subsets: a training set, consisting of 50,000 images, and a test set, comprising 10,000 images. The training set is further subdivided into five training batches, while the test set constitutes a single batch. Notably, the test batch is composed of 1,000 randomly selected images from each class. The training batches, on the other hand, consist of the remaining images, arranged in a random order, albeit with the possibility of slight class imbalance. Each training batch encompasses an equal distribution of 5,000 images from each class, yielding a total of 10 RGB color image categories. Before commencing the formal experiments utilizing the CIFAR-10 dataset, preprocessing steps were implemented, including the transformation of the dataset into tensors and normalization of the data.

2.2 Deep Learning Model

The VIT model represents an extension of the Transformer architecture tailored for Computer Vision applications. The VIT model employs a distinctive approach by dividing an image into smaller patches and subsequently utilizing the sequence of linear embeddings representing these patches as input to the Transformer architecture. This process treats image patches in a similar manner to how tokens are handled within the domain of Natural Language Processing (NLP). Through this methodology, the VIT model can train an image classification model in a supervised manner, leveraging the inherent capabilities of the Transformer architecture. The VIT model consists of three modules which are Embedding layer, Transformer encoder and MLP Head. For image data, VIT model needs a Embedding layer to transform the image data and divides a picture into a bunch of Patches according to a given size. The embedding layer will eventually transform the image to a two-dimensional matrix which is VIT model need. Transformer Encoder is repeated stack Encoder Block which can transform the data form embedding layer. The MLP Head serves to extract the token information and generate the corresponding output. When there is enough data for pre-training, the performance of VIT will exceed that of CNN, breaking through the limitation of transformer lack of inductive bias, and can obtain better migration effect in downstream tasks. Nevertheless, the VIT model lacks certain inherent inductive biases exhibited by CNNs, such as translation equivariance and

locality [7, 8]. Consequently, when training with limited data, VIT may struggle to generalize effectively.

2.3 Data Parallelism

Data parallelism represents a straightforward and accessible approach for achieving parallelism in training models [9, 10]. Data parallelism involves the replication of identical model weights across multiple devices, allowing for the distribution of distinct portions of data to each device for concurrent processing. This strategy effectively parallelizes the training process along the Batch dimension, enabling simultaneous computation on multiple devices. By leveraging data parallelism, the computational workload is efficiently distributed, thereby facilitating faster training and optimization of the model. Data-parallelism relies on parallel processors, which are parallelism within SIMD systems. This technique can divide the training tasks into many parts and distribute those training tasks to many GPUs. The different GPU receive the piecemeal training tasks and start independent training. The function integrates the data together after each GPU completes the training task. Using the Data-parallelism in model training can reduce the burden on a single GPU.A complete data-parallelism operation of data will be divided into three parts which are Task segmentation, Parameter synchronization and Forward/Reverse Calculation. In task segmentation module, the training task is divided into multiple processes (devices), each maintaining the same model parameters and computing tasks, but processing different batch data. Data parallelism can improve training throughput by adding parallel training devices. Afterwards, synchronizing the model parameters and the process will be gradient updated. After calculating the reverse gradient by Loss and updating the model parameters, it is necessary to ensure correct synchronization of model parameters between processes. Finally, each process independently calculates forward based on its own input data and each process independently performs backward calculations based on its own forward calculations. This ensures the acquisition of a consistent global gradient, enabling independent parameter updates across processes.

2.4 Implementation details

Throughout the experimental process, this research employs several libraries, namely torch, torchvision, and einops. The parameter settings remain constant across all experiments. Specifically, the batch size is set to 64, the number of workers is 2, and the number of training epochs is 10. The image size is fixed at 32x256, while the patch size is 32x32. The dropout rate utilized in the VIT model is 0.1. To optimize the model, an optimizer is employed, initialized with a learning rate of 0.001 and a momentum value of 0.9.

3. Results and Discussion

The principle of the research is that by comparing the experimental results of the VIT model with and without data parallelism, as well as data parallelism achieved by multiple GPUs to observe the training experiment effect.

3.1 The Training Results of the Vit Model Under Normal Training and Data Parallelism Training

The VIT model was initialized with the requisite parameters of image_size, patch_size, num_classes and channels parameters. The training outcomes were duly recorded. After keeping all the settings same and using the data-parallelism technique based on the Pytorch to retrain the VIT model. Based on the depiction in Fig. 1, it is discernible that when employing data-parallelism for training the VIT model, the second GPU exhibited an occupancy rate of 31%. Notably, the first GPU displayed an occupancy rate of 57%, surpassing the occupancy rate in conventional training by an additional 4%. It is clear that the loss value in data-parallelism training is lower than which in common training from Fig. 2. In addition, the accuracy of training results in data-parallelism is slightly higher than which in common training according to the Fig. 3.

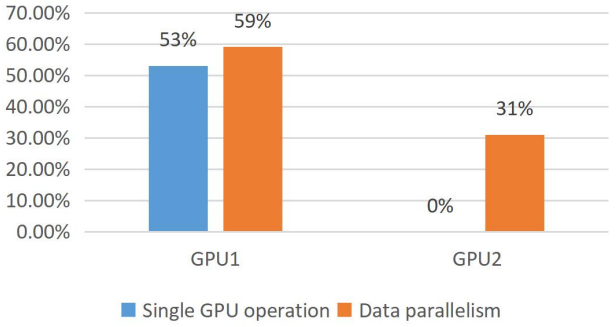

Fig. 1. Vit model GPU usage rate in different training way (Photo/Picture credit: Original).

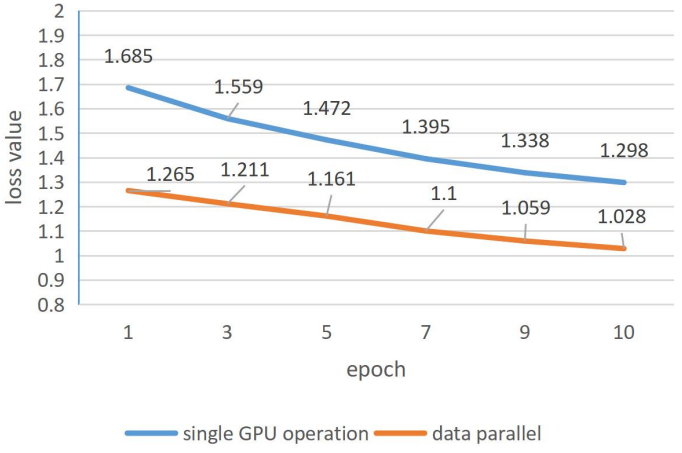

Fig. 2. The change of loss value of vit in different training methods (Photo/Picture credit: Original).

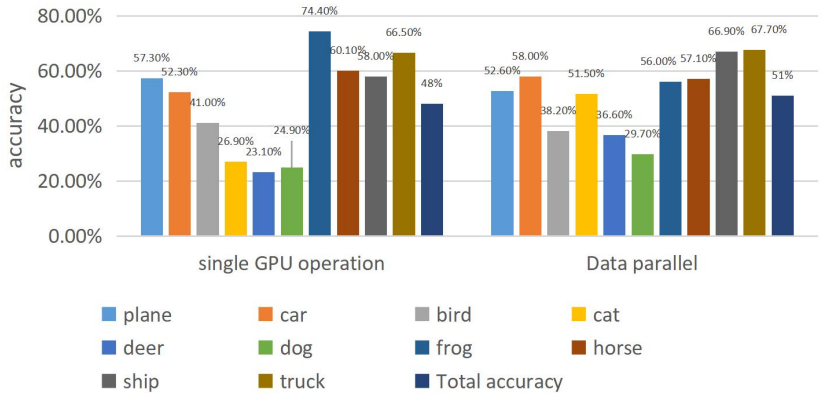

Fig. 3. The accuracy of Vit under different training methods (Photo/Picture credit: Original).

This study conducted multiple experiments to investigate the characteristics of VIT models under various conditions, and aimed to compare the experiment result of different conditions to illustrate which experiment is the most effective. Interestingly, it was observed that employing data-parallelism for training the VIT model not only improved the occupancy rate of GPUs but also yielded significant reductions in the loss value during training and improved the accuracy of image classification. In terms of training effectiveness, as long as data parallel methods are used to train the VIT model, the training results will definitely be optimized.

3.2 The Experiment Results of the Vit Model with Different Numbers of GPUs Under Data Parallelism Training

Continuing from the aforementioned experiment, using different numbers of GPUs for data parallel experiments on the VIT model. According to Fig. 4, when the number of GPUs reached three the occupancy rate maximizes and in the same kind of data-parallelism training, the occupancy rate of the GPU1 always the most with the subsequent GPU occupancy rate is relatively close. The loss value of the experiments will gradually decrease in every kind of experiment and the decrease speed of the loss value is close, but the loss value in three GPUs experiment is far lower than the other two experiments. About the accuracy of the image training by VIT model, the accuracy does not increase or decrease with the number of GPUs in individual image types. Notably, when employing three GPUs, the accuracy remains the highest, while the accuracy achieved with six GPUs slightly surpasses that attained with two GPUs, as depicted in Fig.5 and Fig. 6.

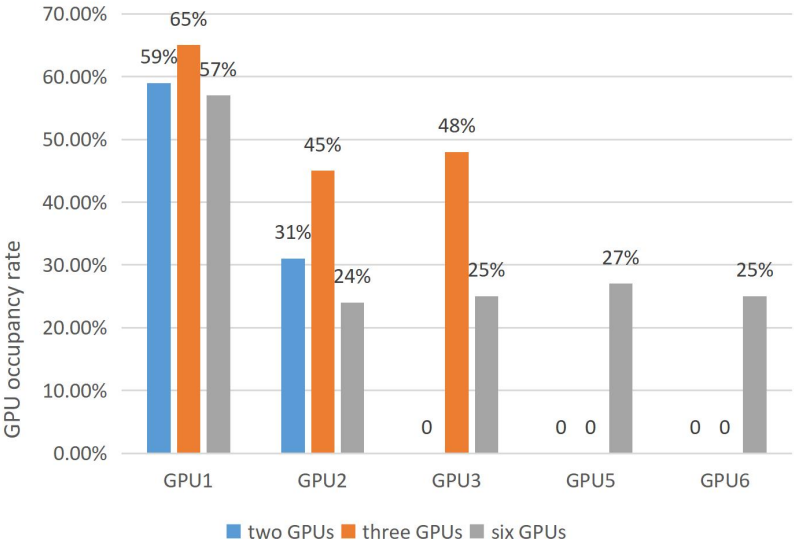

Fig. 4. GPU occupancy rate of vit data parallelism (Photo/Picture credit: Original).

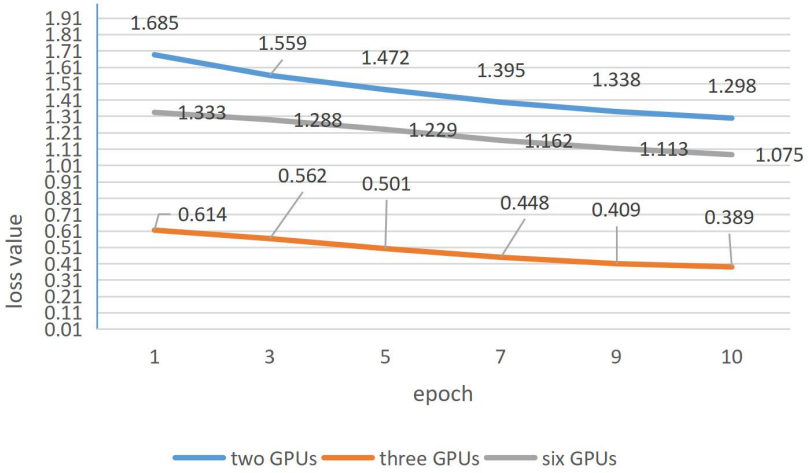

Fig. 5. Change of loss value in vit multi GPUs data parallel training (Photo/Picture credit: Original).

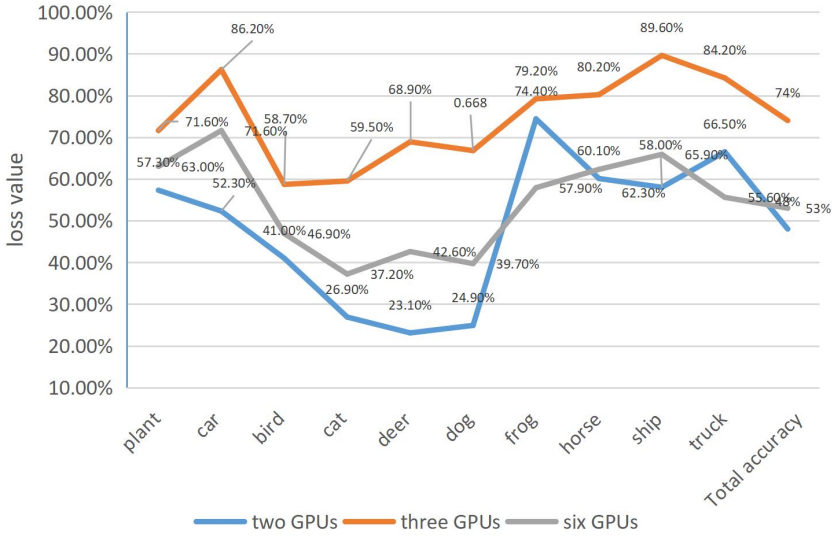

Fig. 6. Change of loss value in Vit multi-GPUs data parallel training (Photo/Picture credit: Original).

About the number of GPUs selected to use in training VIT model, according to the charts, the best choice of training VIT model is three GPUs. When VIT model is using multiple GPUs for data parallelism model training, the occupancy rate of GPU1 always maximum proportion, that is because the Data-Parallelism needs to choose a GPU as the main card, responsible for summarizing output, calculating loss, and updating weights, and the memory and utilization will be higher than other GPUs, resulting in unbalanced load between GPUs. Since the primary card is responsible for communicating with other GPUs, the primary card also has communication bottlenecks. Generally, the higher occupancy rate of GPUs in VIT data-parallelism training will bring higher accuracy and less loss value which mean better training results.

3.3 The Completion Time of Each Experiment for Ten Iterations

The running time is the most intuitive way to reflect the effect of VIT model training. The number of training experiment iterations set up under different conditions is 10 times. Through Fig. 7, it is apparent that the running of setting six GPUs to conduct the experiment is far exceed other experiment. Compared to other experiments, the running time has grown several times. When use three GPUs to conduct the experiment, The model training duration will reach the minimum.

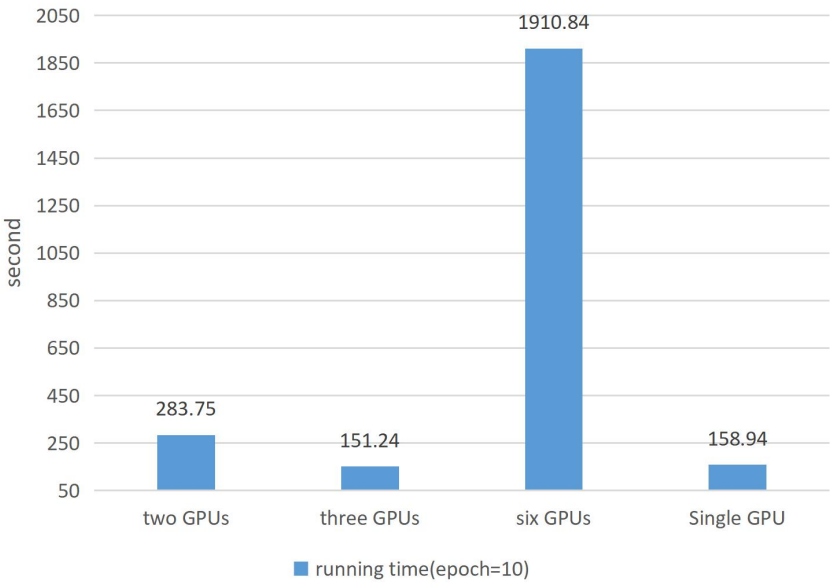

Fig. 7. Run time of vit data parallelism (Photo/Picture credit: Original).

According to the figures above, when use six GPUs to train the VIT model, the experiment results are always unsatisfactory. The effect of VIT data-parallelism training from two GPUs to three GPUs has a great improvement but when it reaches six GPU the effect has actually decreased. This is highly likely because the CIFAR10 datasets are too small. It can be clearly observed that when use the VIT model to train the CIFAR10 in single GPU, GPU usage is still some distance from the upper limit. This means that a single GPU is enough for the VIT model to train the CIFAR10 datasets and when apply multi-GPU and use the data-parallelism to train the CIFAR10 by VIT model the GPU usage rate will increase, and the training effect will also improve. Meanwhile, when there are too many GPUs applied in the model training, will cause significant performance overflow and the work the data-parallelism divide a small training task to multi GPU will waste training time and reduce training efficiency. This could potentially explain the inadequate training performance observed when utilizing six GPUs for training the CIFAR10 dataset with the VIT model.

4. Conclusion

This study uses a method of repeated experiments with data-parallelism for comparison to explore which conditions and setting can maximize the training effect of the VIT model. It is clear that data-parallelism can greatly improve the VIT model training effect by comparing the experimental results under basic training and data parallel training. No matter how many GPUs are used in data parallelism, it also can decrease the loss value and improve the accuracy of image classification, but at the same time it also will increase the GPU usage. In subsequent

experiments, this study tests how many GPU number is the best setting to train the CIFAR10 datasets by VIT model. The experimental results indicate that using three GPUs to train the CIFAR10 datasets by VIT model is the optimal solution. When three GPUs are used in the experiment, the loss value decreases the most significantly and the VIT model achieves the highest accuracy in image classification. When facing small datasets like Cifar10, too many trainings task divided will cause the waste of GPU resources. But considering that the vit model itself is a model for training large datasets, it can be inferred that multi GPU data-parallel will achieve good results when training against large datasets by VIT model.

References

1. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Advances in neural information processing systems, 30 (2017).
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
3. Envelope, G. L., Envelope, H. Y. P., Envelope, Y L., et al.: Recaptured screen image identification based on vision transformer. Journal of Visual Communication and Image Representation (2022).
4. Gupta, S., Lakhotia, S., Rawat, A., et al.: ViTOL: Vision Transformer for Weakly Supervised Object Localization (2022).
5. Jiang, Z., Wang, L., Wu, Q., et al.: Computer-aided diagnosis of retinopathy based on vision transformer. Journal of Innovative Optical Health Sciences (2022).
6. Zhuang, L.: Deep-Learning-Based Diagnosis of Cassava Leaf Diseases Using Vision Transformer. (2021).
7. Kayalibay, B., Jensen, G., van der Smagt, P. CNN-based segmentation of medical imaging data. arXiv preprint arXiv:1701.03056 (2017).
8. Qiu, Y., Wang, J., Jin, Z., et al.: Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72: 103323 (2022).
9. Shallue, C. J., Lee, J., Antognini, J., et al. Measuring the effects of data parallelism on neural network training. arXiv preprint arXiv:1811.03600 (2018).
10. Li, S., Zhao, Y., Varma, R., et al. Pytorch distributed: Experiences on accelerating data parallel training. arXiv preprint arXiv:2006.15704 (2020).