Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/41/20230739

Plant disease recognition based on parallel Vision Transformer

Haoxin Liu

Aulin College, Northeast Forestry University, Harbin, China

Liuhaoxin6@nefu.edu.cn

Abstract. Ensuring food security and increasing crop yields are significant challenges facing the world today as the global population grows. In this context, the seasonable and accurate perception of plant diseases is essential. This paper proposes to combine Vision Transformer (ViT) and Gpipe to identify disease pictures of plant leaves. Among them, ViT is used to ensure the identification accuracy of the model, and Gpipe is used to improve the running speed of the model. This experiment uses the PlantVillage dataset, which includes diseased pictures of various plant leaves, to evaluate model performance. It uses the method of controlled experiments to Identify the model's performance and efficiency of parallelism in the pipeline. After many experiments, the recognition accuracy of the color picture part of ViT on this data set has reached 93%. In addition, the memory requirements for a single GPU are significantly decreased using pipeline parallelism. This study provides a low-cost and efficient plant disease identification tool for the agricultural field, which can detect plant diseases correctly and in time. It is beneficial to increase food production and ensure food safety.

Keywords: Plant Diseases, Pipeline Parallel, Image Recognition, Vision Transformer, Gpipe.

1. Introduction

With the rapid growth of the world's population, the demand for food around the world is increasing year by year, and some countries or regions are still facing severe food shortages [1]. One of the main reasons for these current problems is that plant diseases and insect pests have significantly impacted global agricultural production. These plant diseases have caused a significant decline in crop yields. The most typical example is Panama disease, which once made Gros Michel extinct [2,3]. Because these diseases have seriously affected the world's food supply and the livelihood of farmers. With the development of botany and other disciplines, humans have found ways to cure some plant diseases, but many types of plants and plant diseases exist worldwide. About 50,000 known species are caused by various factors, including fungi and bacteria. According to calculations, the annual crop loss caused by plant diseases accounts for about 20% to 40% of the global output [4,5]. Therefore, correctly identifying the disease and giving the correct treatment is very important.

Traditional methods of detecting plant diseases and insect pests require people with specialized knowledge. Still, using artificial areas in large-scale farms to detect whether plants are diseased is impractical, requiring a lot of workforce and money. In recent years, with the development of computers, deep learning models have shown great potential in detecting plant pests and diseases [6,7]. However, most deep learning methods require considerable computer resources for calculation, and they also need

DOI: 10.54254/2755-2721/41/20230739

to release corresponding data according to different types of identified diseases. Therefore, developing a method for detecting plant diseases and insect pests that can operate efficiently is significant.

This paper proposes to use a Transformer-based deep learning model for image recognition—Visual Transformer (ViT) and combine it with Gpipe to realize the parallelism of training and recognition pipelines. The experiment was conducted on PlantVillage, a dataset of images of plant pests and diseases. Experimental findings indicate that the ViT performs well on PlantVillage, with a recognition accuracy of 93%.

2. Methods

2.1. Visual Transformer

The model used in this experiment is the Vision Transformer. It is a new deep-learning model proposed by Google in 2022 [8]. The Linear Projection of Flattened Patches, Transformer Encoder, and Multi-Layer perceptron (MLP) Heads are the three components that make up the ViT model. The image data is a three-dimensional matrix, which cannot be directly input into the Transformer. Therefore, the data needs to be converted through the first module. First, a picture is divided into Patches of a specific size. Then, each Patch is mapped to a one-dimensional vector through linear mapping, thus completing the preliminary image data processing. The Transformer Encoder uses the self-attention mechanism to capture long-distance relationships in images, allowing the model to assign different attention weights to each image block. In addition, positional encoding provides the model with information about the position of image patches, ensuring that spatial structure is considered. Stacked encoder layers and residual connections further enhance the model's depth and capacity, enabling it to capture complex image features and relationships. The MLP Head acts as a classifier in the model, and it will accept the output of the Encoder for processing to generate classification results.

2.2. Gpipe

Gpipe is a scalable deep-learning training method developed by Google. The core idea of Gpipe is to divide those huge models that do not fit into a single GPU into multiple parts and then put them into multiple GPUs to run. The first is model partitioning, which will divide the model into multiple consecutive parts and put each on a separate GPU. Then, during the model training process, Gpipe will train in a pipeline parallel manner. The process is like the way it works on the production line. When a part of the data is processed on one GPU, the next part of the data can be started on another GPU at the same time. This means that multiple batches of data can be processed simultaneously on different parts of the model. Also, the output of each part of the model will be saved and used as the input for the next part, which ensures a continuous flow of data throughout the model [9]. Using this method can significantly improve the training efficiency of the model and reduce resource consumption.

2.3. Training setting

This experiment uses three sets of control experiments. The first set of experiments uses 1 GPU to run the program, the second set uses 2 GPUs to run the program in parallel with the pipeline, and the third set uses the pipeline to run the program in parallel but on 4 GPUs. Take 2 GPUs as an example to illustrate the implementation principle of GPipe. First, create a Vision Transformer model and divide it into two parts. The first part is allocated to the first CPU and the second to the dual GPU. Then, create the GPipe model, specifying each part's load and operating equipment. A single GPU cannot achieve pipeline parallelism, so the first set of experiments does not take unique methods to run the program. The whole experiment is divided into three large groups, and each group also contains multiple investigations, including using the two image recognition models of VIT and ResNet to train the pictures of the three formats (Color, Gray, and Segmented) in the dataset separately and changing the learning batch Times size.

In addition, it is considered that the training results may be different due to the imbalance of data distribution in the dataset. Two additional sets of experiments will be performed using the processed

DOI: 10.54254/2755-2721/41/20230739

dataset to remove this effect. Among the 38 types of pictures in the data set, the minimum number of pictures is 152, so the pictures of each type in the data set will be randomly deleted until there are only 152 pictures in this type. The processed data sets will be input into the VIT and ResNet models that have realized pipeline parallelism, and the parameters and other conditions of the two experiments are set the same. Finally, the accuracy, throughput, video memory usage, F1, AUC_ROC, and other training data generated each time are counted to evaluate the model's performance.

3. Experimental results and analysis

3.1. Data description

This paper uses the open-source dataset PlantVillage, which contains 54,303 pictures of unhealthy and healthy leaves of plants. The dataset has three types of images: color, grayscale, and segmented. They are color images, grayscale processed images, and images with backgrounds removed. Each of these three types contains 38 different healthy and unhealthy plant leaf images, such as Apple_Apple_scab, Cherry_(including_sour)_Powdery_mildew, Strawberry_healthy, etc. The pixel size of each image is 256*256, and the format is jpg [10]. It must be preprocessed for all image data and then input into the model for training. The first is the preliminary processing of the image: the first step is to cut out a 224*224 pixel area from the center of the image to ensure that all input graphics are the same size. The second step adjusts the range of pixel values from 0-255 to 0-1. Finally, the image is normalized so that the distribution of new pixel values has a specific mean and standard deviation, which can help improve the performance and training speed of the model.

3.2. Analysis

3.2.1. ViT and ResNet in the Colour dataset. This section discusses experimental results in detail and analyses them. To reflect the performance of Vision Transformer intuitively, the Vision Transformer and ResNet were trained on the Colour dataset under the same conditions. ViT has achieved outstanding results in training the color data in the PlantVillage dataset. According to Table 1, the training is performed when the batch_size is set to 128 using 2 GPUs. After training for five epochs, the test set's accuracy of the ViT reached 93%, and values of F1 and ROC_AUC also showed that the model's performance was excellent. The initial training of the ViT resulted in a test accuracy of 75%. In contrast, the model's accuracy could have improved more after the fourth and fifth training. Therefore, depending on the curve's trend, the model's accuracy should remain around 95% even after more training. According to the trend graph of the ResNet training result data, the accuracy rate on the test set after the first training is only 61%, and the accuracy rate is stable at 83% after multiple training. Besides, it can be inferred that ResNet has been trained many times. The accuracy rate should be around 85%. It is known that the accuracy of VIT is 9% higher than that of ResNet after training five times, and it can also be shown in ROC_AUC that the performance of VIT in the Colour dataset is higher than that of ResNet.

Table 1. Using 2 GPUs under the same conditions, VIT and ResNet trained on the Colour dataset.

	2GPUs/128/Color	1	2	3	4	5
ViT	Accuracy	0.75	0.85	0.90	0.92	0.93
	F1	0.65	0.79	0.86	0.89	0.91
	ROC_AUC	0.82	0.89	0.93	0.95	0.95
ResNet	Accuracy	0.61	0.74	0.78	0.82	0.84
	F1	0.65	0.79	0.86	0.89	0.91
	ROC_AUC	0.76	0.79	0.82	0.85	0.86

DOI: 10.54254/2755-2721/41/20230739

3.2.2. ViT and ResNet in Gray and Segmented datasets. As shown in Table 2 and Table 3, comparing the performance of the two models shows that the performance of VIT under the Segmented dataset could be better than that of the Colour dataset. The accuracy rate after training five times in the Gray data set did not reach 80%, but the accuracy rate on the Segmented data set could have been better than the Colour data set, but it also reached 87%. In contrast, ResNet's performance is relatively stable on all three datasets, with an average accuracy of 81%. Several original ones may be for the ViT not performing well in the Gray dataset. The first ViT model is designed for processing colour images and may not be suitable for processing Gray images. The images in the second observation data set show that the diseased parts of most diseased plant leaves are dark yellow, black, brown, and other colours. These colours are very similar after grayscale processing, which may affect the model Judgment on the image's label.

Table 2. Using 2 GPUs under the same conditions, ViT and ResNet trained on the Gray dataset.

	2GPUs/128/Gray	1	2	3	4	5
ViT	Accuracy	0.46	0.58	0.63	0.69	0.75
	F1	0.32	0.47	0.56	0.61	0.69
	ROC_AUC	0.32	0.47	0.56	0.61	0.69
ResNet	Accuracy	0.56	0.67	0.72	0.79	0.80
	F1	0.48	0.66	0.76	0.78	0.81
	ROC_AUC	0.71	0.74	0.82	0.84	0.85

Table 3. Using 2 GPUs under the same conditions, ViT and ResNet trained on the Segmented dataset.

	2GPUs/128/Segmented	1	2	3	4	5
ViT	Accuracy	0.62	0.78	0.81	0.85	0.87
	F1	0.49	0.70	0.75	0.81	0s.84
	ROC_AUC	0.74	0.85	0.87	0.90	0.91
ResNet	Accuracy	0.56	0.67	0.72	0.79	0.80
	F1	0.48	0.66	0.76	0.78	0.81
	ROC_AUC	0.71	0.74	0.82	0.84	0.85

3.2.3. ViT and ResNet use Gpipe. The data in Table 4 shows the benefits of pipeline parallelism achieved using Gpipe. The program running on 1 GPU does not use pipeline parallelism because pipeline parallelism requires at least two GPUs. Observing the data, it can be found that when the accuracy rates of the three are the same, the throughput of the process using pipeline parallelism is higher than that of the process not using pipeline parallelism, and as the number of parallel GPUs increases, the throughput value Gradually improving; the throughput determines the amount of data output by the model within a certain period. Such a result means that the recognition speed of the model is gradually improving. In addition, regarding the occupancy of video memory, during the program's running, calculations and processes are evenly distributed to each GPU. The observed data shows that as the number of GPUs increases, the value of the video memory occupied by each GPU gradually decreases. The video memory occupation of the 1GPU program that does not use pipeline parallelism reaches 21361MIB. However, memory utilization on a solitary GPU for a program employing pipeline parallelism with two GPUs amounts to merely 4243MIB. In contrast, the memory consumption for a single GPU in a program utilizing four GPUs is a mere 3229MIB. The correlation between video memory and the quantity of GPUs is a multifaceted aggregation and pipeline parallelism can significantly reduce the memory usage of a single GPU.

DOI: 10.54254/2755-2721/41/20230739

Table 4. The accuracy and memory footprint of the VIT model when running on different numbers of GPUs.

1GPUs/128/Colour	1	2	3	4	5
Accuracy	0.75	0.84	0.89	0.90	0.93
Throughout	170.98	167.73	173.91	176.79	176.56
VRAM	21361	21361	21361	21361	21361
2GPUs/128/Colour	1	2	3	4	5
Accuracy	0.75	0.85	0.90	0.92	0.93
Throughout	213.02	212.85	212.78	206.12	208.46
VRAM1	4243	4243	4243	4243	4243
VRAM2	4013	4013	4013	4013	4013
4GPUs/128/Colour	1	2	3	4	5
Accuracy	0.78	0.86	0.90	0.92	0.93
Throughout	294.74	283.94	293.11	291.48	288.96
VRAM1	3461	3461	3461	3461	3461
VRAM2	3229	3229	3229	3229	3229
VRAM3	3191	3191	3191	3191	3191
VRAM4	3249	3249	3249	3249	3249

3.2.4. ViT and ResNet in limited datasets. This paper also tests the performance of the two models on limited data sets, as shown in Table 5. Upon careful examination of the data, it becomes evident that the accuracy rate of both models on the randomly deleted data set is approximately 50%. However, the accuracy rate of ViT is 10% higher than that of ResNet. This result is like the previous accuracy rate on the original data set. For the problem of low accuracy of the two models, the analysis believes that since there are only 152 pictures of each category in the randomly deleted data set, the model needs more data to adjust the parameters, leading to the model's accuracy. In contrast, ViT shows better performance under this experimental setting.

Table 5. Accuracy of VIT and ResNet on processed datasets.

	2GPUs/128/Color	1	2	3	4	5
ViT	Accurcy	0.29	0.38	0.53	0.55	0.56
ResNet	Accurcy	0.05	0.1	0.29	0.38	0.46

4. Conclusion

The recognition rate is significant for detecting plant diseases and insect pests. The planting area of crops is vast, and one percent of errors may cause several tons of crops to fail to produce typically. The improvement of VIT's accuracy in identifying plant leaf diseases and insect pests will significantly impact crop yields. This paper finds that first, the accuracy of VIT in identifying diseased plant leaves is very high, reaching 93%, which is higher than ResNet's 84%, and the performance of other aspects of the model is also excellent. Besides, Pipeline parallelism based on GPipe can improve the model's throughput and reduce the memory usage of a single GPU. The data set used in this paper only contains 38 species, but in fact, about 7,000 species of crops are planted by humans. With the increase in data volume, it is still being determined whether ViT will maintain such good performance. In addition, the experiment found that the performance of ViT in the recognition of gray images did not meet the expected standards. Compared with color images, the volume of gray images will be smaller, which will be in the case of tight storage space due to the large amount of data. So, improving VIT's performance in gray image recognition is an important direction.

DOI: 10.54254/2755-2721/41/20230739

References

[1] Goncharova, N. A., & Merzlyakova, N. V. (2021). Food shortages and hunger as a global problem. Food Science and Technology, 42, e70621.

[2] Warmington, R. J., Kay, W., Jeffries, A., O'Neill, P., Farbos, A., Moore, K., ... & Studholme, D. J. (2019). High-quality draft genome sequence of the causal agent of the current Panama disease epidemic. Microbiology Resource Announcements, 8(36), 10-1128.

[3] Dita, M., Barquero, M., Heck, D., Mizubuti, E. S., & Staver, C. P. (2018). Fusarium wilt of banana: current knowledge on epidemiology and research needs toward sustainable disease management. Frontiers in plant science, 9, 1468.

[4] Ristaino, J. B., Anderson, P. K., Bebber, D. P., Brauman, K. A., Cunniffe, N. J., Fedoroff, N. V., ... & Wei, Q. (2021). The persistent threat of emerging plant disease pandemics to global food security. Proceedings of the National Academy of Sciences, 118(23), e2022239118.

[5] Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. Nature ecology & evolution, 3(3), 430-439.

[6] Liu, J., & Wang, X. (2021). Plant diseases and pests detection based on deep learning: a review. Plant Methods, 17, 1-18.

[7] Saleem, M. H., Potgieter, J., & Arif, K. M. (2019). Plant disease detection and classification by deep learning. Plants, 8(11), 468.

[8] Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., ... & Wu, Y. (2019). Gpipe: Efficient training of giant neural networks using pipeline parallelism. Advances in neural information processing systems, 32.

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2010). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020. arXiv preprint arXiv:2010.11929.

[10] Hughes, D., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv preprint arXiv:1511.08060.