Research on the performance of hybrid vision models based on ViT

Bowen Chai

Shanghai University of International Business and Economics, School of International Business, Shanghai, 201620, China

balwyn134821589341@163.com

Abstract. ViT is a model proposed by the Google team in 2020 to apply a transformer in image classification, although it is not the first paper to apply a transformer in visual tasks, because of its model is "simple" and effective, and scalable, it has become a milestone work in the application of transformer in CV field, and has also triggered the subsequent related research. The core conclusion of the original ViT paper is that when there is enough data for pre-training, ViT outperforms CNN, breaks through the limitation of the transformer's lack of inductive bias, and can achieve better migration results in downstream tasks. However, when the training dataset is not large enough, ViT usually performs worse than ResNets of the same size, because Transformer lacks inductive bias, a kind of a priori knowledge, assumptions made in advance, compared with CNN. Through its innovative architecture and powerful performance, the visual representation transform (ViT) model continues to advance the field of computer vision, while facing some challenges and room for improvement. With the deepening of research and the continuous development of technology, ViT is expected to play a greater role in more practical applications. The article aims to explore the advantages and applicability of the ViT model and tries to construct a hybrid visual model to improve its generalization ability for different types of datasets, demonstrating the hybrid model's significance in improving the performance of the ViT model.

Keywords: Hybrid Vision Models, Vision Transformer, CIFAR10, CNN+ViT, Performance

1. Introduction

The development of ViT can be traced back to 2017 when the Attention Is All You Need paper proposed the transformer structure for realizing machine translation tasks.[1] Later, the transformer structure was widely used in speech recognition, natural language processing and other fields, and it was also explored and attempted in the image field. In 2018, some scholars proposed to use a transformer instead of CNN, which was called Image Transformer, but the effect was not as good as CNN. [2]Until 2020, the ViT model proposed by the Google team[3] used the transformer structure to realize the image classification task, and achieved comparable performance with the CNN model, which is called "Vision Transformer".[4]

According to He et al, ResNet50 is a convolutional neural network model with 50 convolutional layers. It was proposed by researchers at Microsoft and won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2016. The model adopts the idea of residual learning to avoid the gradient vanishing problem of deeper networks while ensuring the expressive power of deeper networks.[5]The principle of ResNet50 is mainly to fuse the feature maps of the previous layers with those of the later layers through direct connections across channels, so that the whole model can be deeper and maintain a certain gradient flow, avoiding the deep network's gradient vanishing problem. Meanwhile, techniques such as batch normalization, pre-activation and residual block are also used to further improve the expressive ability and training speed of the model.

The CIFAR-10 dataset is open-access and is often used to train and evaluate computer vision algorithms. Since the images in the dataset are small, the algorithms can be trained and tested quickly, and it is suitable as an entry-level dataset for training and testing some basic computer vision algorithms, such as object recognition, classification, localization, tracking, and other tasks. Another important significance of the CIFAR-10 dataset is that it has also become a standard benchmark dataset for the evaluation of some deep learning algorithms, such as ResNet50, Inception, etc. Part of CIFAR-10 is shown in Fig. 1:

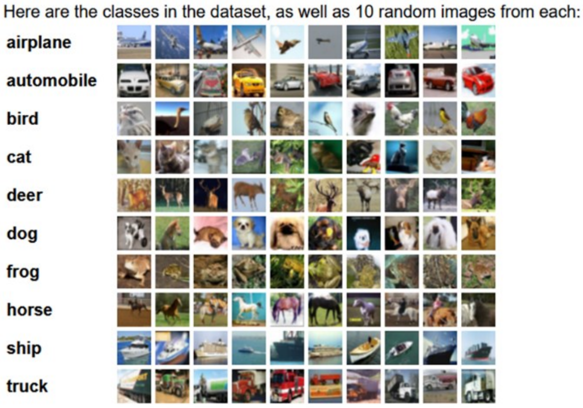
Here are the classes in the dataset, as well as 10 random images from each:
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

Figure 1. Selected images in CIFAR-10

This paper explores the performance of ViT model in small-scale image classification tasks by comparing the performance of the traditional ViT model and ResNet50 on the CIFAR10 dataset. Since the performance of ViT is slightly inferior in dealing with small-scale datasets, this paper will improve the traditional ViT model by introducing a variable convolutional layer to construct a CNN-ViT model, and further compare the performance of the three models on the CIFAR10 dataset.

Compared with the existing literature, the possible marginal innovations of this paper are: (a) Introducing variable convolutional layer to construct CNN-ViT model, in this paper, CNN-ViT model is constructed by introducing variable convolutional layer, which mixes convolutional neural network and ViT model, to give full play to the advantages of each of them. Compared with the traditional ViT model, the CNN-ViT model has better feature extraction capability and higher classification accuracy and shows better performance when dealing with a variety of datasets such as the CIFAR10 dataset. (b) Constructing hybrid visual models to improve generalization ability, in this paper, the generalization ability of the ViT model on small-scale datasets is improved by constructing hybrid visual models. The hybrid model fully exploits the inter-correlation between data to enhance the generalization ability of the model in dealing with small-scale datasets where ViT is not originally dominant. The results show that the hybrid visual model plays an important role in improving the generalization ability of the ViT model.

2. Research Methodology

2.1. Selection of Dataset

In this study, many factors are considered and finally, CIFAR-10 was decided to be chosen as the training dataset.

According to Chen, Deng, & Du, large-scale datasets commonly used for visual model training are ImageNet, JFT-300M, Google Landmarks v2, iNaturalist 2018, etc.[6] They contain more than one million image data. However, limited by hardware devices and network conditions, finally the relatively small-scale CIFAR-10 was chosen as the training dataset. Although the ViT model is not quite able to demonstrate its advantages on such small-scale datasets, this paper subsequently tries to improve the traditional ViT model based on this and explores whether it can enhance its performance on small-scale datasets.

2.2. Training the Vit Model

The data comes from the public dataset on the web. The code in this paper is borrowed from Chatgpt and GitHub-related open-source code.[7]

The first step is to import the PyTorch library and obtain the ViT model code.

Step 2: Define the image classifier, training function and test function.

Step 3: Load the dataset

Step 4: define hyperparameters and optimizer

Step 5: start training. The number of iterations of the training process is controlled by the epochs variable and traverses the dataset loader in each iteration for each batch of data. In each training iteration, the code also calculates the average loss value, accuracy, and total number of samples in the training set, and after using optimizer.zero_grad() to perform a zero operation on the model gradient, it uses the backwards () gradient backpropagation function,[8] which computes the gradient value corresponding to each sample point, and then uses scaler.step() to the parameter in the optimizer to update it. Finally, the code also counts the information such as the amount of correctness between the output results and the actual labels and the total amount of correctness in each batch of data to calculate the accuracy of the training, and stores the information such as the total loss degree and the accuracy of that training, respectively, in a list to visualize the training process after the training is completed.

2.3. Training the Resnet50 Model

In the first step, TensorFlow [9] and Keras [10] libraries are used for the construction of the convolutional neural network model, which performs the classification task for ten different categories of the CIFAR-10 image dataset.

In the second step, the CIFAR-10 dataset is imported.

In the third step, the labels of the dataset are uniquely thermally encoded, converted into a vector consisting of 10 binary bits, and assigned to the y_train and y_test variables, respectively. For the image data in the x_train and x_test variables, normalization is performed before model training is performed by converting the pixel values from integers from 0 to 255 to floating-point numbers between 0 and 1 and preprocessing the data into a format that meets the requirements of model training.

In the fourth step, the construction of the ResNet50 model was started, using the ResidualBlock function to implement the residual block.

Finally, the Adam optimizer is set up with a loss function of categorical_crossentropy and an evaluation metric of accuracyaccuracy; the best model is saved periodically for the validation set during model training through the ModelCheckpoint and EarlyStopping callback functions, and training is stopped early to prevent overfitting Use the fit function to train the model, pass in the training set and test set, specify the batch size as 128 and the number of training periods as 20 and call the callback function during the training process to complete the training of ResNet50 for CIFAR-10 and output the training results.

2.4. CNN+ViT Model Construction and Training

The idea of constructing a hybrid CNN+ViT model to improve the performance of ViT comes from an article on ViT parsing posted on the web by a scholar from Zhejiang University.[11] According to Zhang, J., the core conclusion of the original ViT paper is that when there is enough data for pre-training, ViT outperforms CNN, breaks through the limitation of the transformer's lack of inductive bias, and can obtain better migration results in downstream tasks. However, when the training dataset is not large enough, ViT usually performs worse than ResNet of the same size, because Transformer lacks inductive bias, a kind of a priori knowledge, and assumptions made in advance, compared with CNN, which makes CNN have a lot of a priori information and need relatively less data to learn a better model.

Then, since CNN has the property of inductive bias, and the Transformer has strong global inductive modelling ability, perhaps a hybrid model using CNN+Transformer can get better results. So, this paper constructs a CNN+ViT hybrid model and continues to use it to train CIFAR-10, and finally compares the training performance of these three models.

Again, the PyTorch open-source library Timm was imported first, so that the ViT model could be imported later.

Next, a three-layer convolutional neural network (CNN) was imported to work with the original ViT model.

The ViT encoder uses the Transformer to process the features generated by the CNN encoder, which splits the image into blocks of a pre-set “patch_size”, in this case, 64, and embeds each block into a low-dimensional vector representation. These vectors are then processed under the multi-head self-attention mechanism and sent to the fully connected network for final classification.

After this again the CIFAR-10 dataset is imported and preprocessed; and the Adam optimizer and loss function are defined and training is started to record the results.

3. Experimental Results Analysis

This project used Google Colab to complete the training of the models, ran in Python 3 Google Compute Engine backend (GPU) mode, and used Tableau to visualize the training results in data. All models were trained with Epoch preset to 20.

Among them, the ViT model was used to train CIFAR-10 20 times, ResNet50 3 times, and CNN+ViT 10 times; the more complete data of them was recorded, as shown below:

3.1. Training Results of the Vit Model

As shown in Table 1 and Figure 3, the training results of the ViT model are presented as follows, and the training duration is 7 hours.

Table 1. ViT training CIFAR-10

Epoch	Training Set Loss Degree	Training Set Accuracy	Test Set Loss Degree	Test Set Accuracy
1	1.6349	27.53%	0.0136	35.71%
2	1.5205	38.20%	0.0128	41.08%
3	1.2647	45.46%	1.4291	46.88%
4	1.3445	49.24%	0.0109	50.12%
5	1.4079	51.70%	0.0107	51.36%
6	1.2204	52.97%	0.0101	54.35%
7	1.2554	54.05%	0.0099	54.11%
8	1.2286	55.30%	0.0098	54.59%
9	1.2731	56.35%	0.0097	56.13%
10	1.2527	56.85%	0.0091	57.03%
11	1.2284	57.61%	0.009	58.54%
12	1.3317	58.12%	0.0088	59.17%
13	1.0816	58.97%	0.009	58.41%
14	1.192	59.26%	0.0085	60.56%
15	1.2002	60.30%	0.0085	60.97%
16	1.2811	61.17%		

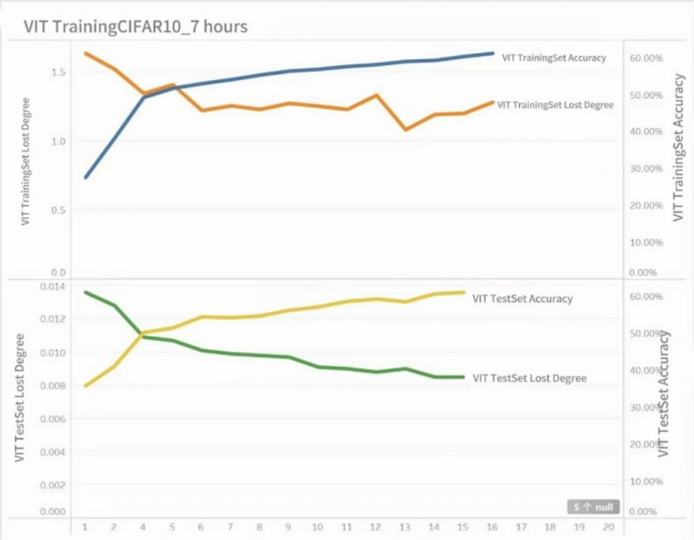

Figure 3. Visualization of ViT model training results

3.2. *Training Results of ResNet50 Model*

As shown in Table 2 and Figure 4, the training results of the ResNet50 model are presented as follows, with a training duration of 6 hours and 40 minutes.

Table 2. ResNet50 training CIFAR-10

Epoch	Training Set Loss Degree	Training Set Accuracy	Test Set Loss Degree	Test Set Accuracy
1	1.8700	43.07%	2.1633	26.16%
2	1.3869	57.82%	5.1761	22.96%
3	1.2296	63.59%	2.025	32.55%
4	1.0266	63.59%	1.2364	60.98%
5	0.8755	72.25%	0.862	69.77%
6	0.7357	75.79%	0.9031	70.63%
7	0.6806	77.34%	1.0124	66.27%
8	0.5295	81.63%	0.7192	76.26%
9	0.4534	84.19%	0.8596	73.39%
10	0.3787	86.76%	0.715	77.48%
11	0.3157	88.83%	1.3259	67.83%
12	0.3037	89.51%	0.9077	73.80%
13	0.2145	92.38%	1.0554	73.75%
14	0.1727	93.87%	0.9612	75.96%
15	0.153	94.55%	0.8413	79.21%
16	0.1267	95.54%	1.0632	77.02%
17	0.1129	96.00%	1.2959	74.89%
18	0.1037	96.39%	1.0033	77.04%
19	0.0957	96.70%	0.9779	78.35%

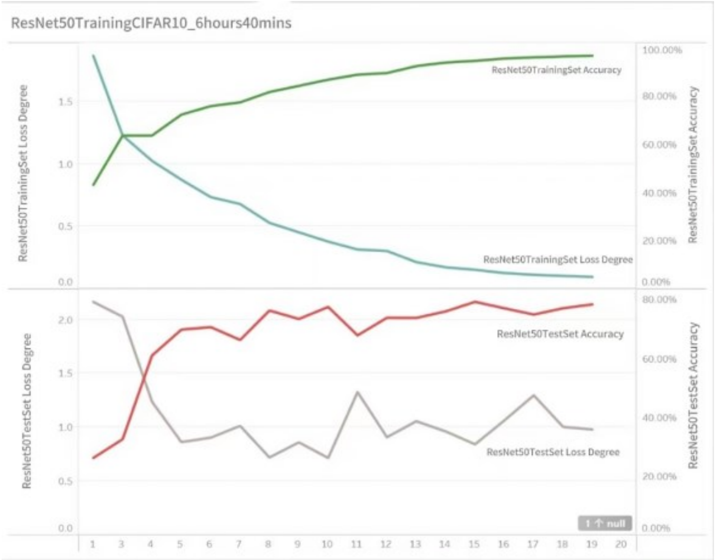

Figure 4. Visualization of ResNet50 training results

3.3. Training Results of CNN+Vit Model

As shown in Table 3 and Figure 5, the training results of the ResNet50 model are presented as follows, and the training time is 4 hours and 17 minutes.

Table 3. Training results of CNN+VIT

Epoch	Hybrid Model Loss Degree	Hybrid model correctness
1	1.9532	40.35%
2	1.4583	50.60%
3	1.2619	56.85%
4	1.1469	60.48%
5	1.0673	62.56%
6	0.9989	63.56%
7	0.9459	67.12%
8	0.9056	68.60%
9	0.8577	69.10%
10	0.799	70.28%

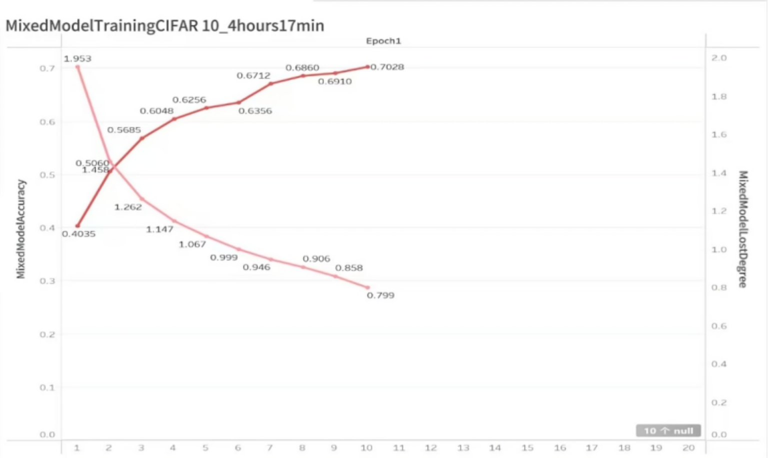

Figure 5. Visualization of CNN+ViT training results

3.4. Comparison of the Three Models

As shown in Table 4, the following results can be obtained from the comparison:

Table 4. Training results for each model

Model name	Training hours	Total traversals	Maximum accuracy (test set)
ViT	7hrs	16	60.97%
ResNet50	6hrs 40mins	19	78.35%
CNN+ViT	4hrs 17mins	10	70.28%

As we can see, CNN+ViT has exceeded the accuracy of the normal ViT model despite having the lowest number of traversals, ResNet50 has the highest number of training traversals and also has the highest correctness rate, it is believed that if CNN+ViT can achieve the same number of training layers, its performance can be close to that of ResNet50 or even exceed it.

As shown in Touvron et al., ResNet50 outperforms ViT on small datasets such as CIFAR-10, while ViT performs better on larger datasets such as ImageNet-1k with more than a million images.[12]

Figure 6 presents the correctness of the three models in the form of curves.

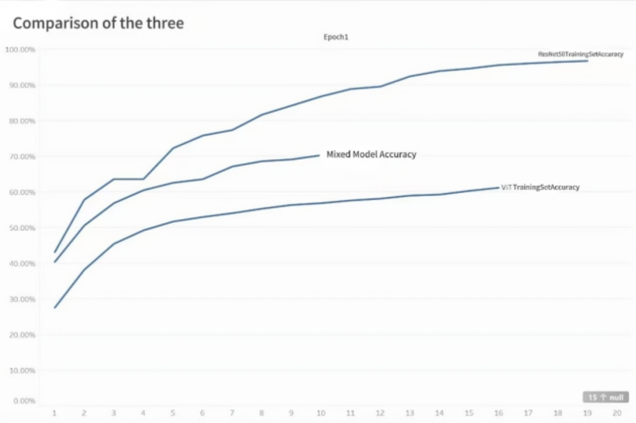

Figure 6. Comparison of the three models' accuracy

4. Conclusion

This paper explored the performance metrics such as the number of training traversals and accuracy under different neural architectures. It is found that the CNN+ViT model has a lower number of traversals (only ten), but its accuracy has surpassed that of the regular ViT model.Meanwhile, the ResNet50 model has the highest number of training traversals, but also exhibits the highest correctness rate. The CNN+ViT model is believed that if it can achieve the same number of training layers as ResNet50, its performance will keep approaching or even surpass the ResNet50 model. Our results are in line with Touvron et al. that the ResNet50 model performs well on small datasets such as CIFAR-10, while the ViT model performs better on much larger datasets (e.g., ImageNet-1k, which has more than one million image data).

Not coincidentally, according to Zheng et al., in a small-sample image classification task, they constructed a hybrid model of a Convolutional Neural Network and Transformer called LCPN and explored the synergy between local and global features.[13] The article proposes a Local Composition Module (LCM), which has a structure very similar to a Transformer but does not use the attention mechanism to process local information. Meanwhile, the model combines LCM and CNN, which makes LCPN can classify small samples more effectively. Through extensive experimental evaluations on several small-sample image classification benchmark datasets, the article demonstrates that LCPN has excellent performance in image classification tasks with small samples and has an advantage in innovatively capturing information about local compositional patterns.

And recently there have been many new models derived based on the ViT model, such as ViT-pyramid.[14] the model adopts a pyramidal multi-scale feature representation, i.e., the input image is scaled to different scales, and then features are extracted at each scale, which are finally fused to classify the image. the main idea of ViT-pyramid is to replace the global field of view with a finer field of view, to improve the classification accuracy and efficiency of the model. thereby improving the classification accuracy and efficiency of the model. Its core structure is a set of Transformer Block-based feature extraction layers, each of which contains a self-attention mechanism for learning the relationship between the current position and other positions. These layers are characterized by the ability to dynamically adjust the size of the region of the self-attention mechanism to achieve feature extraction and combination at different scales. ViT-pyramid has achieved good performance on several datasets, especially on large-scale datasets, such as ImageNet-21K.

U-ViT, proposed in CVPR2023,[15] combines Vision Transformer, a vision model, with U-Net, and applies it in Diffusion Model (Diffusion Model) to replace the original CNN, and applies the long skip structure of U-Net in Transformer as well to realize image generation using the Transformer for the task of image generation.

References

[1] Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez & Polosukhin (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[2] Siva, Wong & Gong (2018). Improved techniques for training GANs. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden (pp. 5070-5079).

[3] Alexey Dosovitskiy. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [EB/OL]. [2023/5/27]. https://arxiv.org/abs/2010.11929.

[4] Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, & Houlsby, (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).

[5] He, Zhang, Ren & Sun, (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[6] Chen, Deng & Du (2021). Recent Advances on Vision Transformers. arXiv preprint arXiv:2106.13129.

[7] Hugging Face. (n.d.). pytorch-image-models/vision_transformer.py. GitHub. Retrieved June 3, 2 023, from https://github.com/huggingface/pytorch-image-models/blob/main/timm/models/vision_transformer.py

[8] PyTorch. (n.d.). Optim: Implementations of Gradient Descent Algorithms. PyTorch Documentation. Retrieved June 3, 2023, from https://pytorch.org/docs/stable/optim.html.

[9] TensorFlow. (2021). TensorFlow documentation. Retrieved October 25, 2021 from https://www.tensorflow.org/

[10] Keras. (2021). Keras documentation. Retrieved October 25, 2021 from https://keras.io/

[11] Zhang, J. (2021). PyTorch Lightning 1.2 - New Modules and Features. Retrieved September 7, 2021, from https://zhuanlan.zhihu.com/p/445122996?utm_id=0.

[12] Touvron, Caron, Alayrac & Misra (2021). From ResNets to Pre-Training: Revisiting ImageNet Pre-Training. arXiv preprint arXiv:2105.14333.

[13] Zheng, Wei, Yang, Zhang & Huang (2021). Exploiting Local Compositional Patterns for Few-Shot Image Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5908-5917.

[14] Tolstikhin, Houlsby, Kolesnikov, Beyer, Zhai, Unterthiner & Raffel, C. (2021). Multi-scale Vision Transformers: An Evolution towards Competitive Computer Vision Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14352-14361).

[15] Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., & Zhu, J. (2023). All are Worth Words: A ViT Backbone for Diffusion Models. arXiv preprint arXiv:2209.12152v4.