Mini-Project 2-3 Mid-Report

Shunyi Zhu, Zhen Wang, Haoze He
Department of Eletrical and Computer Engineering
Tandon School of Engineering, New York Univeristy
sz3719, zw2655, hh2537@nyu.edu

1 Methodology

1.1 Vision Transformer

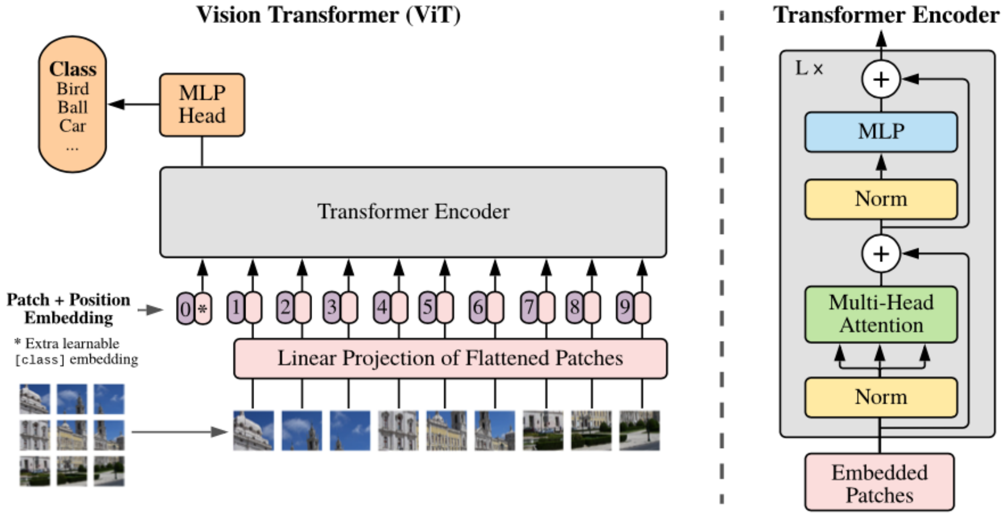

Figure 1: Base ViT Model [1]

From the original paper of the ViT model we test, this model split an image into fixed-size patches, linearly embed each of them, add position embedding, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, they use the standard approach of adding an extra learnable ”classification token” to the sequence.[1]
They provide models pre-trained on ImageNet-21k for the following architectures: ViT-B/16, ViT-B/32, ViT-L/16 and ViT-L/32. They provide the same models pre-trained on ImageNet-21k and fine-tuned on ImageNet.

Their work showed that ViT uses self-attention to integrate information across the image, and the average distance spanned by attention the average distance spanned by attention weights at different layers[1]. This distance is relative similar to the field sizes in CNN, if the depth(layer) is larger, the distance going to be larger. Their work showed that the ViT has really a huge power to reach the high accuracy and by splitting the image, the attention is well learned.

1.2 Convolution Neural Network

The Convolution Neural Network is a deep learning tool that can take in an input image and be able to differentiate object from one another by assigning weights to different characteristics in the figure. For now, CNN has earned great success in the field of computer vision and video analysis. One advantage given by the CNN is that is can learn the spatial and temporal dependencies.

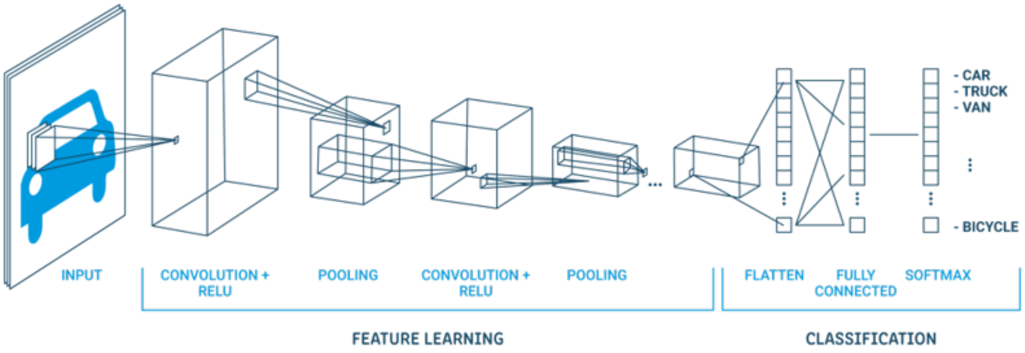

Figure 2: CNN Model

Both convolution neural network(CNN) and vision transformer(ViT) are success in past few years. They are two dominant frameworks in the field of computer vision. However, these two frameworks have their own drawbacks. As for CNN, the convolution layer has limited respective field, which makes it hard to capture the associated information between global and local area. As for ViT, it can capture the associated information using self-attention. But without similarity comparisons among all tokens also leads to high redundancy. In addition, the low-efficiency flaw of ViT transformer caused by high computational/space complexity in Self-Attention. To combine the advantage of both ViT and CNN, we plan to reduce the number of tokens using submodular optimization, merge tokens with high similarity into one, and use it as input for CNN architecture.

2 Current work

We successfully run and test the original framework on the given datasets. However, due to the limit of the hpc resources(GPU Memory resources limit), we cannot use the default/recommended training parameters to train our model to get the same accuracy as the model's from our based code by far. Therefore, we adjust the batch size, reducing the batch size from 512 to 32, and test on several datasets with the pretrained model.

Training Parameters	Value
dataset	cifar100
pretrained model type	ViT-B_16
img size	224
train batch size	32
eval batch size	32
learning rate	0.03
weight decay	0
num steps	10000
decay type	cosine
warmup steps	500

From our training-test result, we can see that there is an accuracy dropping from the original best accuracy. After a rough calculation, the average running time of this model on cifar-100 or cifar-10

Table 1: Test result on CIFAR-100

Training Result	Value
Valid Loss	0.25000
Valid Accuracy	0.92420
Best Accuracy	0.924700
Original Best Accuracy	0.9390±0.0005

Training Parameters	Value
dataset	cifar10
pretrained model type	ViT-B_16
img size	224
train batch size	32
eval batch size	32
learning rate	0.03
weight decay	0
num steps	10000
decay type	cosine
warmup steps	500

Table 2: Test result on CIFAR-10

Training Result	Value
Valid Loss	0.985
Valid Accuracy	0.98820
Best Accuracy	0.988300
Original Best Accuracy	0.9942±0.0003

on hpc is 2.5 hours, which is a suitable length of time for us at present, and we decided to use these two relatively small datasets for subsequent training and testing.

3 Next Work

Our goal is to combine CNN with ViT and effectively utilize the advantages of both. In order to achieve this goal, we are splitting the original input images through CNN. We believe that CNN has the advantage of dividing pictures through feature maps, and then further improves learning efficiency through transformers.
The performance of our algorithm will be evaluated by comparing with the baseline provided by Transformer and Convolution Neural Network in different perspectives. We will take the conformer as our baseline to compare the results. We will also try to reduce the complexity of self-attention. On the other hand, if time is permitted, we will test the performance of certain parameters to see how the different value for different parameters will affect the final result.

References

[1] Dosovitskiy, Alexey (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Computer Vision and Pattern Recognition (cs.CV), Artificial Intelligence (cs.AI), Machine Learning (cs.LG), FOS: Computer and information sciences, FOS: Computer and information sciences*.

[2] Kunchang Li, Yali Wang (2022) UniFormer: Unifying Convolution and Sqelf-attention for Visual Recognition *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[3] Lu, Jiachen (2021) SOFT: Softmax-free Transformer with Linear Complexity. *NeurIPS*.

[4] Ma, Teli (2021) Oriented Object Detection with Transformer. *Computer Vision and Pattern Recognition (cs.CV), FOS: Computer and information sciences, FOS: Computer and information sciences*, arXiv.

[5] Naseer, Muzammal (2021) Intriguing Properties of Vision Transformers. *Computer Vision and Pattern Recognition (cs.CV), Artificial Intelligence (cs.AI), Machine Learning (cs.LG), FOS: Computer and information sciences, FOS: Computer and information sciences*. arXiv.