P2 Image Classification

Name: 赵子涵
Major: Computer Science
Student ID: 2023E8013282148

Dataset

We use CIFAR10 as the data source of our project. CIFAR10 is one of the most famous dataset in the field of computer vision. It contains 60000 32*32 color images in 10 classes. The train set takes 50000 of it, and the rest are treat as test set.

Model Optimization

In this section, we show some optimization attempts as we implementing the ViT model.

The implementation adopt several ideas of [1].

Basic Model

As the first attempt, we implement the basic model of ViT. A ViT consist 3 main parts, the embedding layer, the transformer, and the MLP layer.

The embedding layer divide the whole images into patches, so we can treat it as a sequential data later. For the transformer, we only use the encoding part, which is composed of attention layer and MLP layer alternately. Finally, we use a MLP head to generate the classification result.

Sadly, the accuracy on test set barely reach 50%.

Position Embedding and CLS Token

In the first attempt, I did not embed any patch position. This makes the model unaware of the order of patches, leads to bad performance. Besides, it is recommend to use a unique classification head to produce result, instead of averaging every head.

To do this, we first generate a random vector `cls_token` and place it in the front of every projected patches. As for the position embedding, to ensure the model know the information of position, we just need to determine an order. Therefore, we generate a fixed random vector for each head (including the cls head) and add it on the embedded patches.

The accuracy improvement is not significant.

Data Augmentation

As a transformer, the model requires so much data to train properly. Sadly, the original training set of CIFAR10 only contains 50000 images. Therefore, data augmentation is quiet necessary.

To do this, in every epochs, we make a little difference on the data. Specifically, we first randomly resize and crop image to a fix size, then flip and rotate them. We also try to adopt the data augmentation implemented by torch `v2.AutoAugment(v2.AutoAugmentPolicy.CIFAR10)`.

Patch embedding

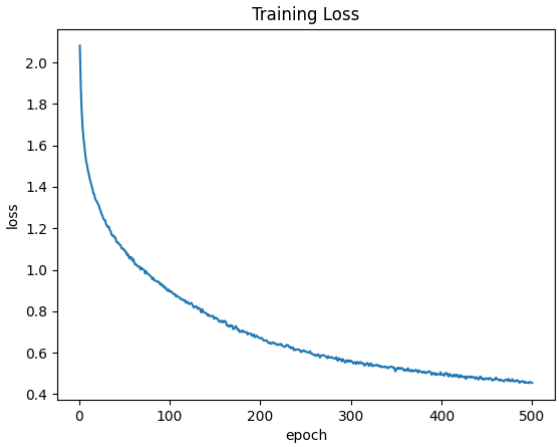
The original ViT use a linear layer to map a patch to vec. Instead, we try to replace it with a convolution layer.

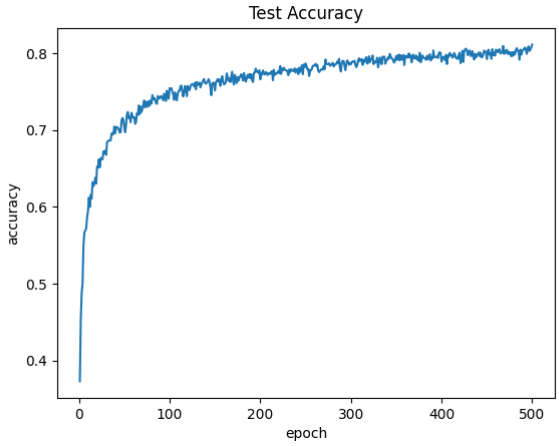
Result

Integrating the optimization tricks above, and setting the hyper-parameter as

```
BATCH_SIZE=256
LR=0.0003
patch_size=(4, 4)
embed_dim=512
transformer_depth=12
heads=8
mlp_dim=1024
dropout=0.2
```

the loss and accuracy curves are as follows.





After 500 epochs, the model finally reached 80% accuracy on test set.

---

1. https://github.com/FrancescoSaverioZuppichini/ViT ↩