VRDL Homework 1 Report

0710788 劉怡廷

- ## GitHub link

  https://github.com/tina-1007/Bird-Images-Classification

- ## Reference

  TransFG: A Transformer Architecture for Fine-grained Recognition

  - Paper - https://arxiv.org/abs/2103.07976
  - Github - https://github.com/TACJu/TransFG

- ## Introduction

The homework is a task of fine-grained visual classification (FGVC), which is to classify sub-classes of object category, considered as a challenging task because the high similarity between classes. I apply TransFG to the task. It is the first study using vision transformer in FGVC, and performed second best now in CUB-200-2011 dataset.

- ## Methodology

1. Data pre-process

Since all file names and corresponding class labels have already wrote in training_labels.txt, all I need to do is to read the text file and make a list of image paths and labels, and use it to establish dataset and data loader. There are 3000 training images, I split it into 8:2, 2400 for training and 600 for validation.

To perform data augmentation, I set different data transfer for two data loaders. In the training part, I resized images into 600x600 and randomly cropped to 448x448, every epoch during training may cut out distinct image to use. After that, the image went through randomly horizontal flipped and normalized. As for validation, just resized, center cropped, and normalized.

2. Model architecture

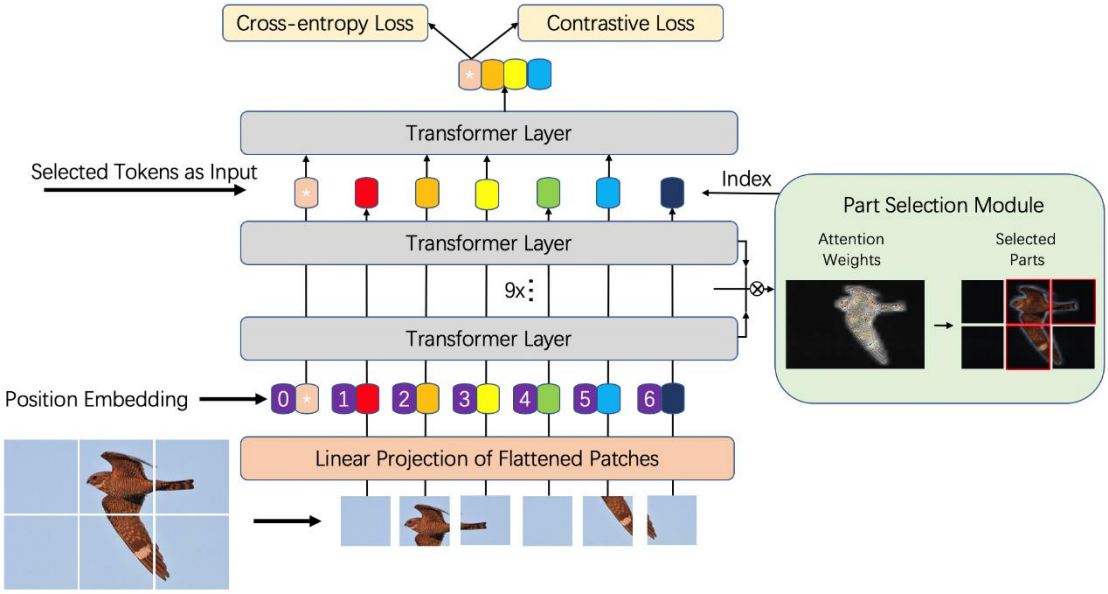

Figure 1. TransFG Model Architecture

Vision Transformer series models proposed a different way to process input images. Unlike traditional CNN, they split image into small patches as a sequence then put them in Transformer based model. Before the last layer, TransFG added a Part Selection Module to choose only important image patches as input.

3. Hyperparameters

- Pretrained model: ViT-B_16 pretrained on ImageNet21k
- Training Batch size: 4
- Optimizer: SGD with momentum = 0.9, initial learning rate = 0.03
- Scheduler: Cosine annealing
- Step size: 10000 (100 times validation)

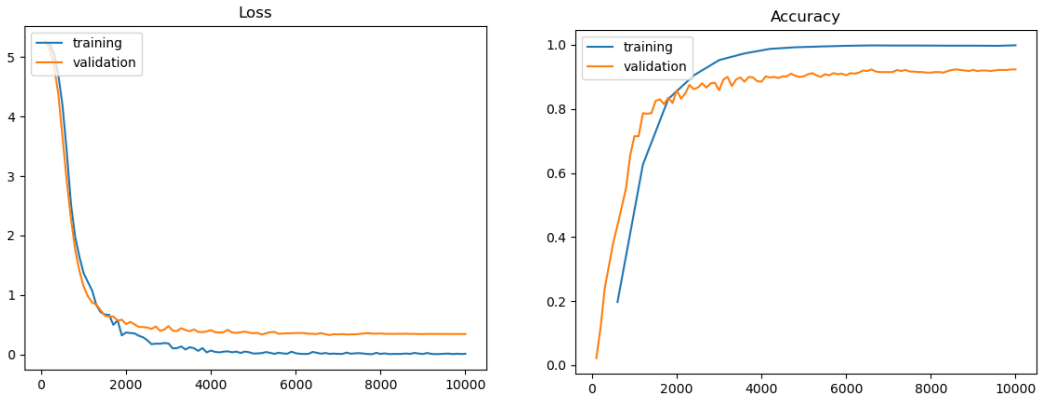

Figure 2. Training Loss and Accuracy during 1000 steps

- ## Summary

It’s my first time to deal with fine-grained visual classification, I have learned many new experience like paper searching and parameter adjusting. At first, I tried to train my model by transfer learning from pretrained ResNet model, but I failed to beat the baseline. After studying some related works, I think TransFG is the most understandable for me, so I choose this paper as my reference. With its help, I reach a better score finally.