Munier Antoine

Homework 4

Objective :

To implement and compare the performance of Vision Transformer and SWIN models on the CIFAR-10 dataset for object classification, and to analyze their decision-making processes using Grad-CAM visualization.

1) Data Preparation

In this part we load the CIFAR-10 dataset with the Pytorch framework. We normalize, resize the images and split them between training and testing.

2) Model Selection and Fine-tuning

We create a function that prepares the model by calling ViT and SWIN Transformer. The function also modifies the classification header of each model to generate 10 class. We also choose Option 2 to freeze the earlier layers and refine only the classification header and subsequent layers to reduce computational demand.

(I made this choice because I think the goal of the exercise is to code a functional tool rather than the best possible tool that requires a lot of time and resources to generate.)

3) Training

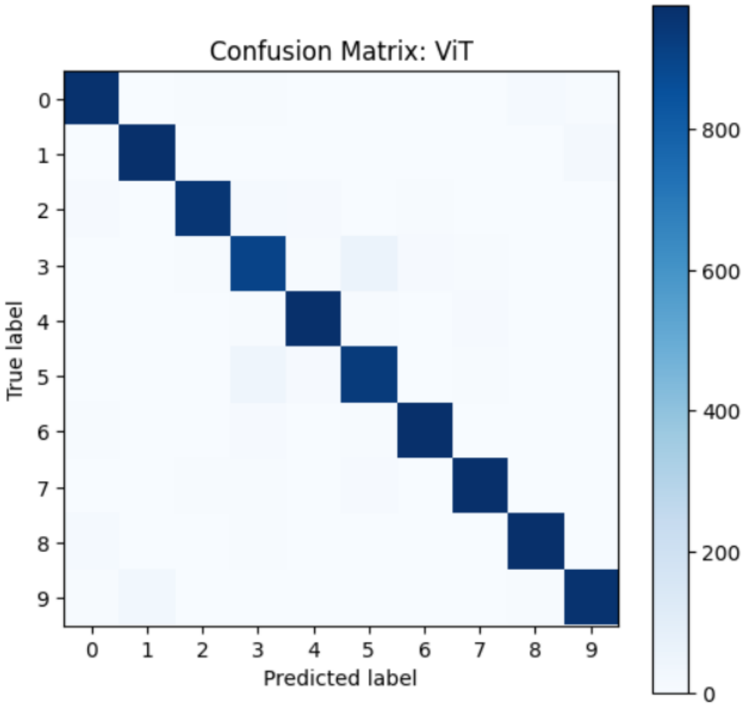
We train our models with the AdamW optimizer, and the learning rate schedules cosine annealing. We also track the training progress by looking at loss and accuracy. We display and save the model checkpoints with the best validation accuracy.

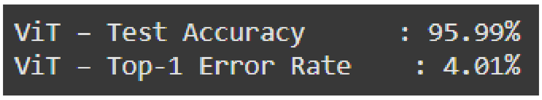
We also add a progress bar to follow the progress of the training.

I chose to use 5 epochs because after several attempts with this number, you're sure to achieve good results. With fewer, you can be unlucky and "underperform," but with more, it never increases.

4) Evaluation

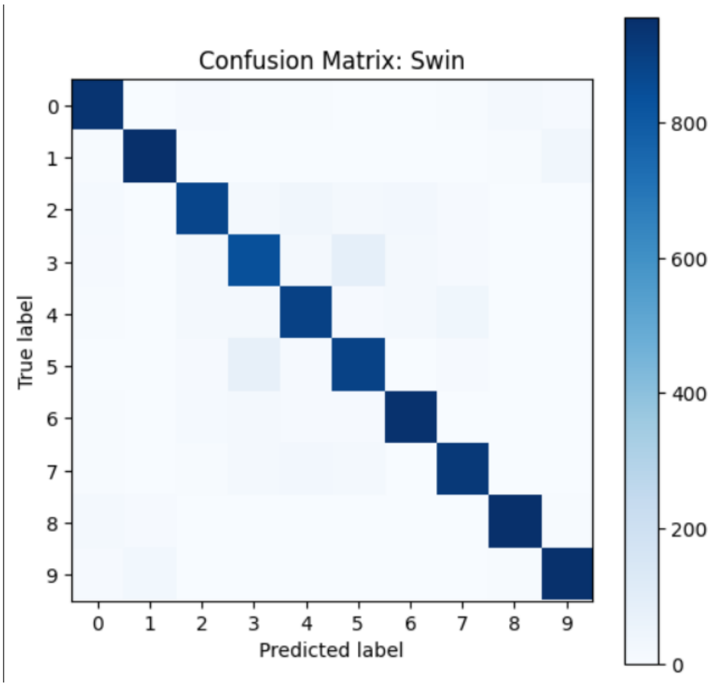
Vit result:

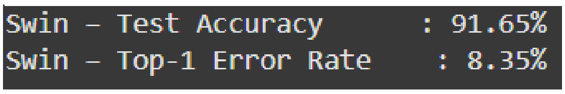




```
ViT – Test Accuracy      : 95.99%
ViT – Top-1 Error Rate    : 4.01%
```

Swin result:





```
Swin – Test Accuracy       : 91.65%
Swin – Top-1 Error Rate     : 8.35%
```

As a reminder, here are the 10 classes: 0 airplane, 1 automobile, 2 bird, 3 cat, 4 deer, 5 dog, 6 frog, 7 horse, 8 ship, 9 truck.

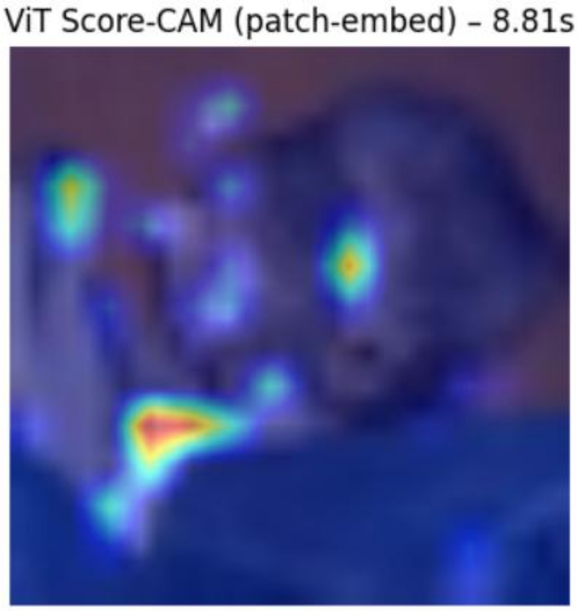
We notice from the confusion matrices that where the two models are most mistaken is in classes 3 and 5. These two classes are quite similar (cat and dog). It is therefore quite logical to obtain these results.

5) Grad-CAM Visualization

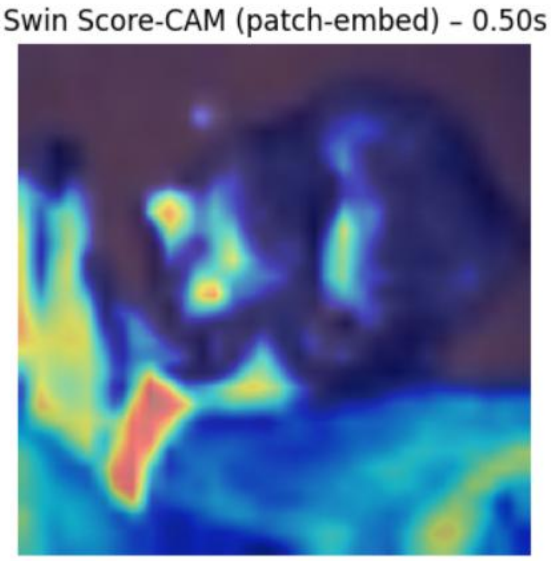
In this part we use Grad-CAM to generate the heatmap. Unfortunately, I encountered problems when implementing it. Indeed, Grad-CAM needs a tensor of dimension 4 and the layer I was looking at was only 3. I tried many things to solve this problem, but I failed. So, I used the Score-CAM library which does the same thing. In a slower way because it does N forward passes while Grad-CAM 1 alone with the gradient. For one image this will not affect us much.

Randomly selected image:

Heatmap ViT :



Heatmap Swin :



6) Analysis and Comparison

In terms of accuracy, ViT is 4% better than swin. This is likely due to its overall processing. Swin remains good (92%), but the local windowing seems to lack the overall context.

Heatmap comparison:

ViT: We see a very pronounced bright red peak in the bottom-center/left: this is the area of the chassis/wheel that most influences the prediction. At the wheelbase. There are a few secondary hot spots in yellow, but the focus remains very focused on specific points.

Swin: The map is more diffuse and yellow/green tinged on the left half (background and leg) and moderate yellow around the head. The only real red peak is also on the leg, but it is less isolated than with ViT, and the entire background takes on a warm relief.

We can conclude that ViT is better for the global context with an object on several parts of the image. However, the model takes more time. Swin is therefore more economical in time and resources but a little less good. But it has more difficulty when the indices are at long distance because it works locally.

Conclusion :

The heatmaps reveal that ViT distributes its attention more precisely across several key areas, reflecting its ability to understand the general context and extract only what is necessary to make a choice. Conversely, Swin, thanks to its hierarchical windowing, distributes attention across several patches. This explains the warmer heatmap. Vit is therefore the best choice for accuracy, but Swin remains very good and less resource-intensive.