Homework #3 Report

Deep Learning for Computer Vision

資工碩一　張凱庭　R10922178

Problem 1

1. Accuracy

Accuracy	0.9526

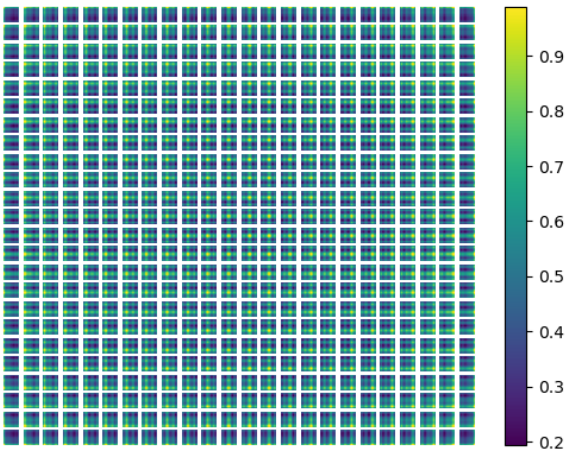
Model Setting	
Model	“B_16_imagenet1k”
Pretrained	True
Patch size	16x16
Number of patch	24x24
Num of head	12
Num of layer	12

Hyperparameters	
Epochs	9
Optimizer	SGD
Learning rate	0.001
Batch size	8

Discussion:

A. Using the model pre-trained on imagenet-1k has the best performance.
B. Smaller patch size (16×16) has better performance than (32×32), probably due the original image size of this dataset is bigger.

2. Visualize position embeddings of your model.

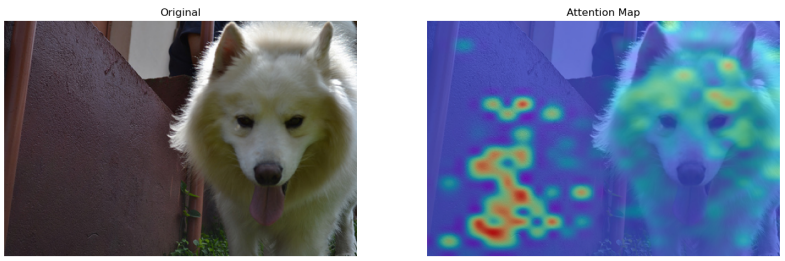


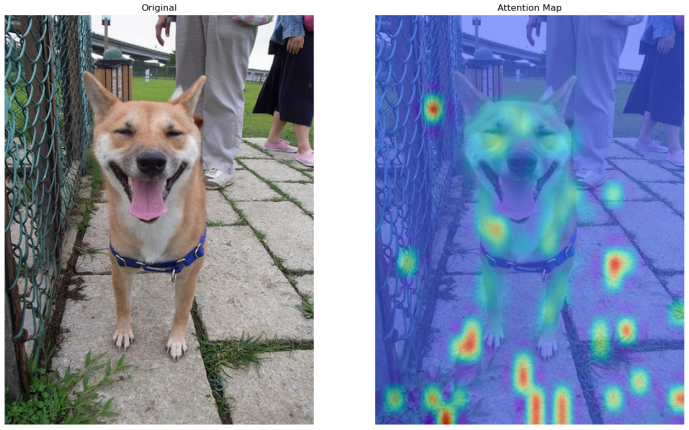
Discussion:

A. The learnable positional embedding tends to pay attention to its own position, which is reasonable because we want the flatten patches to contain positional information.

3. **Visualize the attention map of 3 images between the last multi-head attention layer.**

26_5064.jpg


29_4718.jpg


31_4838.jpg


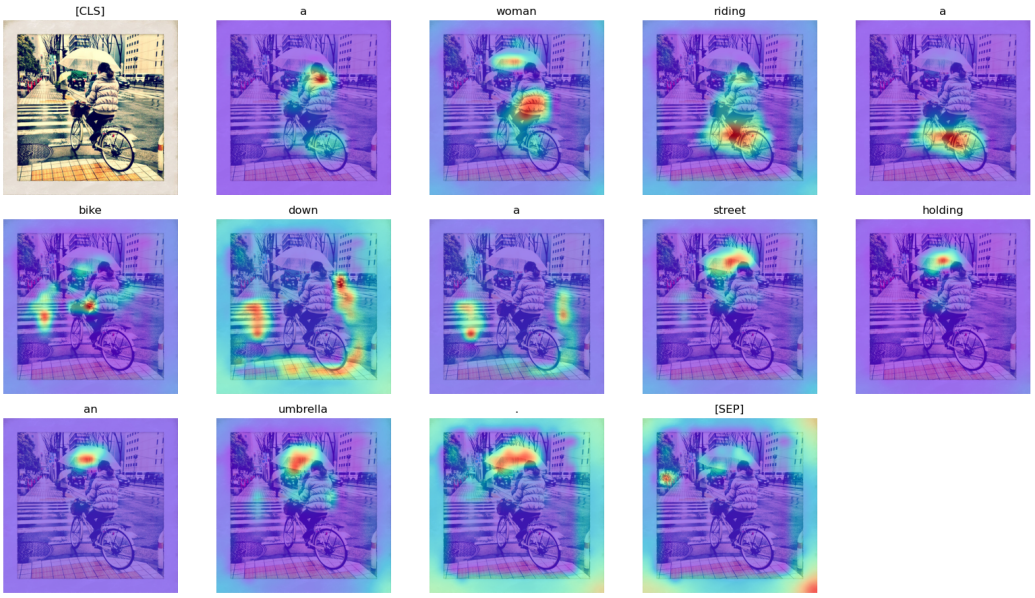
Discussion:

A. The predictions of these images are all correct. Some attention goes to the background. The reason could be that when it comes to categorizing different dogs and cats, the place where the picture was taken also plays an important role. Take 26_5064.jpg for example, cats mostly stay indoors so the model focuses more on the background object. Not only the dog or cat itself, but also the **background context** matters.

Problem 2

1. **visualize the predicted caption and the corresponding series of attention maps**

bike.png



Analyze:

A. The caption and the attended region is resaoble. Model recognized the umbrella and the bike well. The action parts were also described correctly(riding and holding).

Discussion:

A. For image captioning, the encoder part is familiar with the vision transformer. For the decoder part, it takes in the output of the encoder and the word embedding. The output feature dimension of the last decoder is equal to vocabulary size in order to predict the next word.

B. In the original work, the input image is resized according to the longer side. For the convenience of the visualization of cross attention map, I resized all the image to same size: (299×299), so the cross attention become in same dimension: (128×361). Then I reshape the second dimension to (19×19) and resize to the same size of input image.

Reference

1. Vision Transformer: https://github.com/lukemelas/PyTorch-Pretrained-ViT
2. Attention map visualization: https://github.com/jeonsworld/ViT-pytorch
3. Caption Transformer paper: https://arxiv.org/abs/2101.10804
4. Caption Transformer implementation: https://github.com/saahiluppal/catr