Homework #3

Deep Learning for Computer Vision

Due: 110/12/14 (Tue.) 03:00 AM

電機所 R09921102 曾翊維

Problem 1: Image Classification with Vision Transformer

1. Report accuracy of your model on the validation set. (TA will reproduce your results, error ±0.5%) (10%)

a. Discuss and analyze the results with different settings (e.g. pretrain or not, model architecture, learning rate, etc.) (8%)

I found that small batch size and small lr lead to better result.

Model	Input size	Pretrained	Batch size	Lr	Weight decay	Transforms	Acc
B_16	224 * 224	TRUE	32	1E-05	1E-04	tfm1	0.9193
B_16	224 * 224	TRUE	16	1E-05	1E-04	tfm1	0.93
B_16_imagenet1k	384 * 384	TRUE	8	1E-05	1E-05	tfm2	0.95

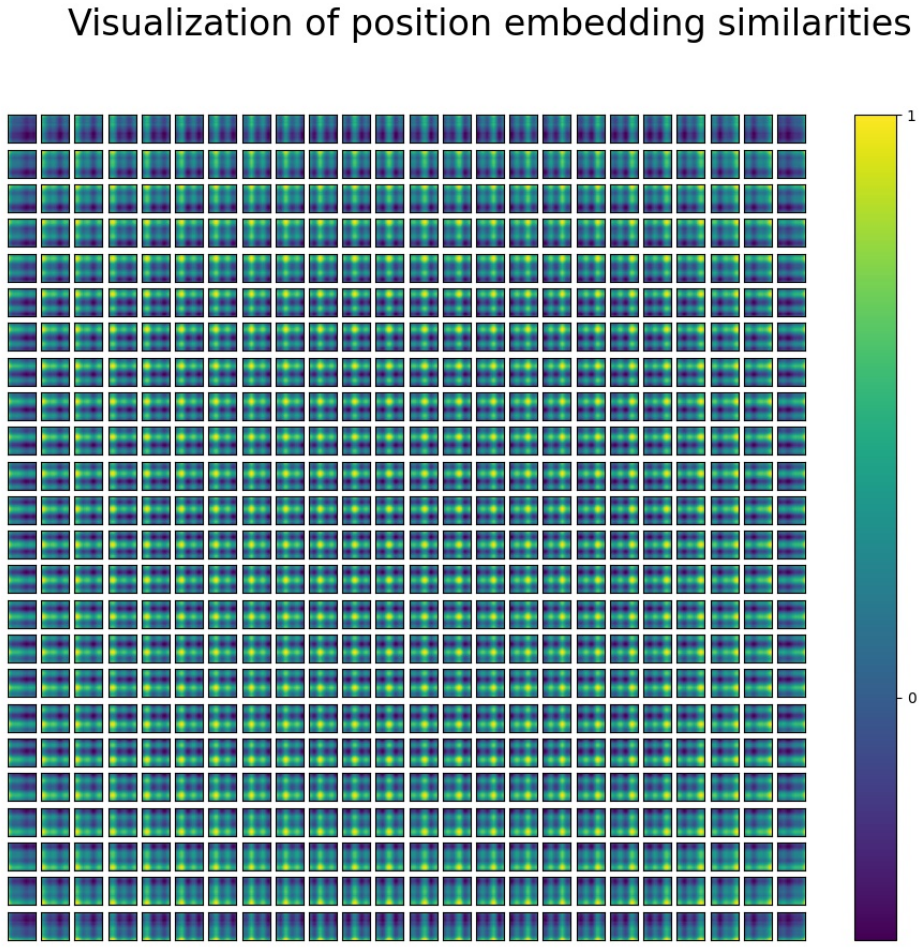
```
tfm1 = transforms.Compose([
transforms.RandomRotation(20),
transforms.RandomResizedCrop(224, scale=(0.7, 1.0)),
transforms.ColorJitter(brightness=0.3),
transforms.ToTensor(),
transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))])

tfm2 = transforms.Compose([
transforms.RandomRotation(30),
transforms.RandomResizedCrop(384, scale=(0.8, 1.0)),
transforms.ColorJitter(brightness=0.3),
transforms.ToTensor(),
transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))])
```

b. Clearly mark out a single final result for TAs to reproduce (2%)
Final Accuracy = 0.95

2. Visualize position embeddings (20%)

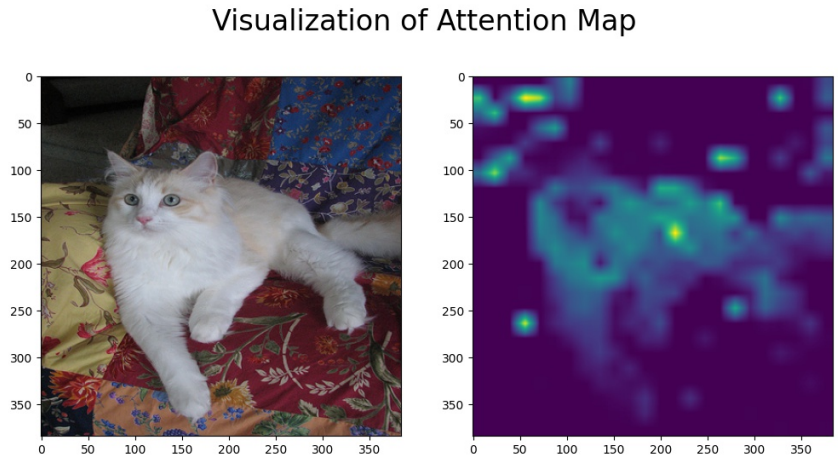
a. Visualize cosine similarities from all positional embeddings (15%)

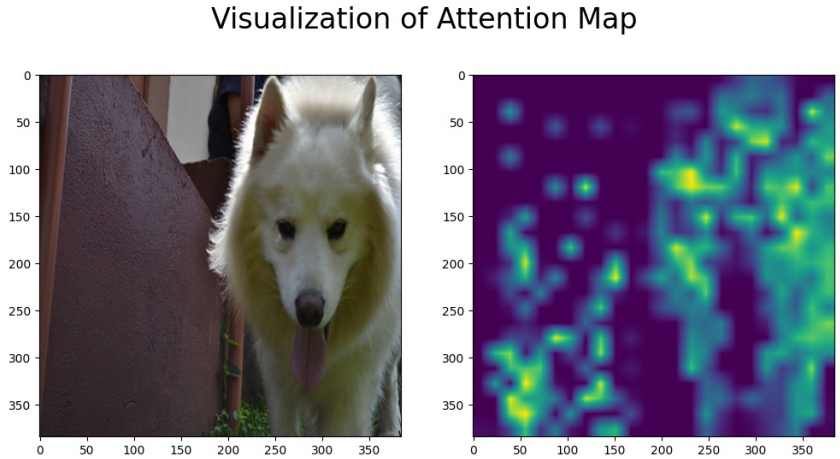


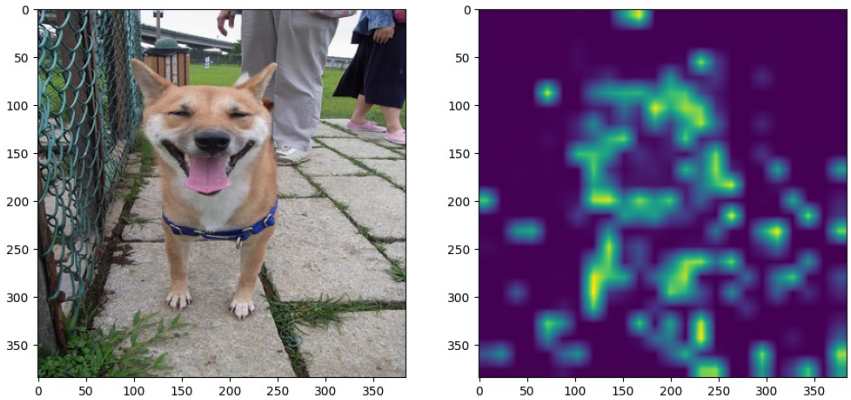
b. Discuss or analyze the visualization results (5%)
My input size is (384, 384) and patch size is (16, 16), So the number of patches is 24*24 = 576. As shown above, the position of yellow spot with larger pixel value locates according to the patch index(i.e. patch[i, j] with larger pixel values around row i and col j), which indicates that the position embeddings are well learned after training.

3. Visualize attention map of 3 images (p1_data/val/26_5064.jpg, p1_data/val/29_4718.jpg, p1_data/val/31_4838.jpg) (20%)

   a. Visualize the attention map between the [class] token (as query vector) and all patches (as key vectors) from the LAST multi-head attention layer. Note that you have to average the attention weights across all heads (15%)







b. Discuss or analyze the visualization results (5%)

As shown above, 26_5064.jpg and 31_4838.jpg yield expected results with larger pixel values at the location of main object.

29_4718.jpg, however, shows that some attentions are located at the brown object in the left of the photo, which may result from the fact that the brown object with large region of similar pixel values just looks like an animal with brown fur.

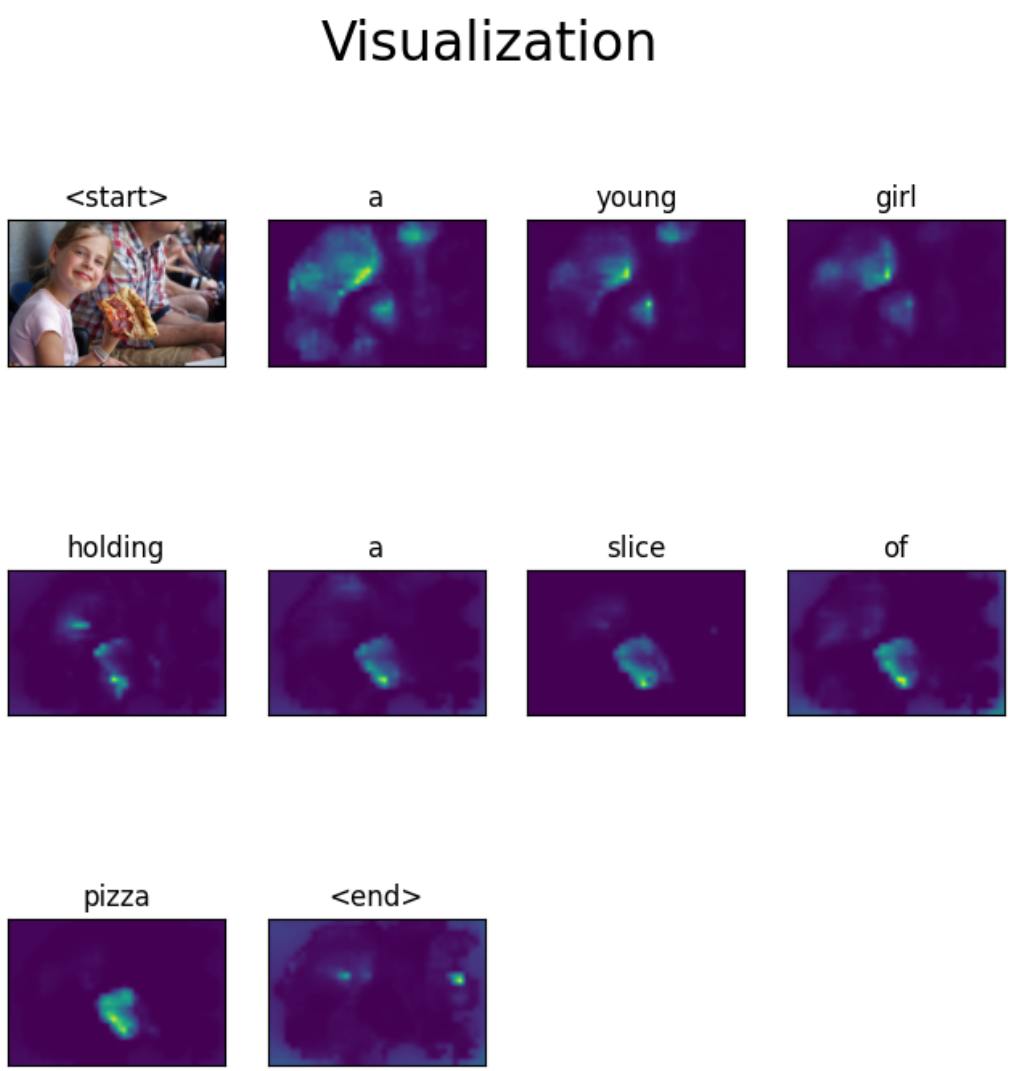
Problem 2: Visualization in Image Captioning

1. For the five test images, please visualize the predicted caption and the corresponding series of attention maps in a single PNG output. TA will reproduce your visualization results with your bash script. (10%)

   Done!

2. Choose one test image and show its visualization result in your report.

   a. Analyze the predicted caption and the attention maps for each word. Is the caption reasonable? Does the attended region reflect the corresponding word in the caption?



As shown above, the attention maps of words related to a girl with are actually highlighted in the corresponding region. Likewise, the last few attention maps indeed reflect the corresponding word "pizza" in the caption.

b. Discuss what you have learned or what difficulties you have encountered in this problem.

   One of the difficulties I encountered was that the model was downloaded to ./user/.cache/… , and I hadn't notice that until I found that the result stayed the same no matter how I modified the model source code in my local directory.
   Then, I used the following code to specify the download path.
   "torch.hub.set_dir(<path>)"

Ref:

1. https://colab.research.google.com/github/hirotomusiker/schwert_colab_data_storage/blob/master/notebook/Vision_Transformer_Tutorial.ipynb#scrollTo=QvmVuX38ZzQF
2. https://www.tensorflow.org/tutorials/text/image_captioning