DLCV HW3 Report

B07502071 陳志臻

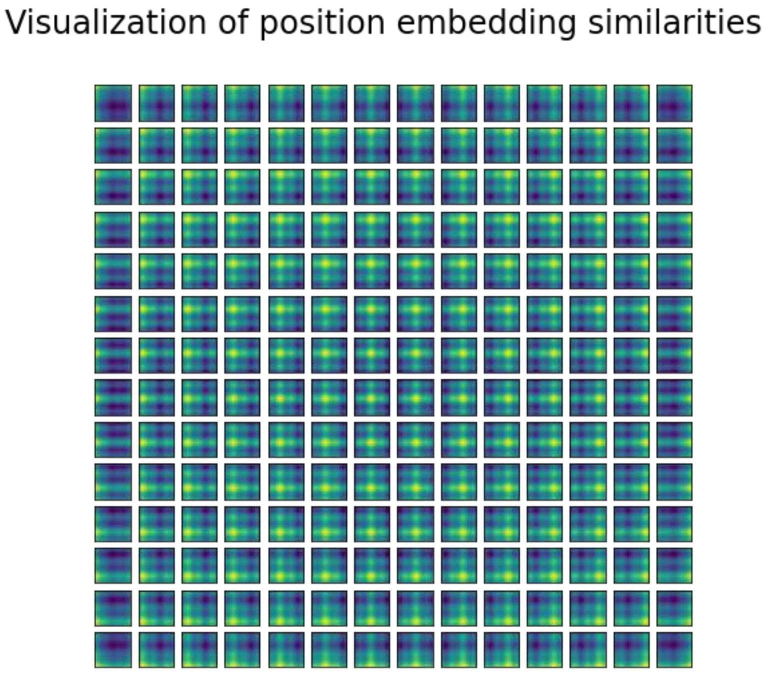
P1、Vision transformer

1. Test set: Average loss: 0.0551, **Accuracy: 1421/1500 (94.7%)**

   Analysis: In this task, I used the "B_16_imagenet1k" pretrained model from "pytorch_pretrained_vit" package. Also I add a linear layer to do the classification. For all the training data, I resized them into (384, 384), and I also use AutoAugment in transforms. I have tried several kinds of optimizer such as SGD, Adam, AdamW. I have also tried several kinds of learning rate scheduler such as get_linear_schedule_with_warmup, get_constant_schedule and get_cosine_schedule_with_warmup. The best selection set is:

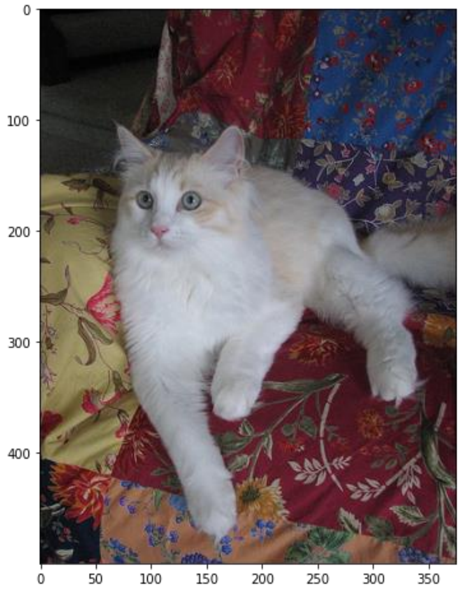
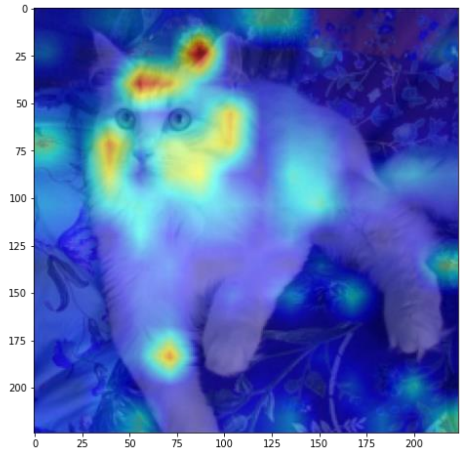
   optimizer = optim.SGD(model.parameters(), lr = 0.0003, momentum = 0.9)
   scheduler = get_cosine_schedule_with_warmup(optimizer, 5, 20)

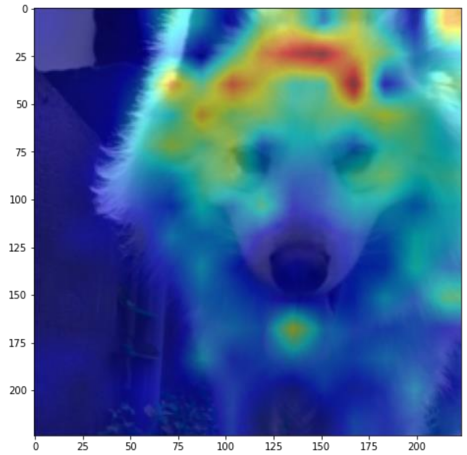
2. Position embedding

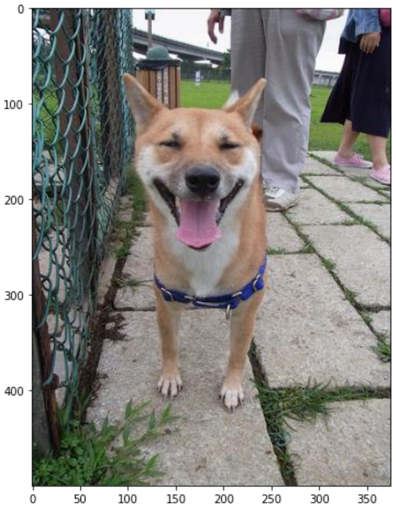
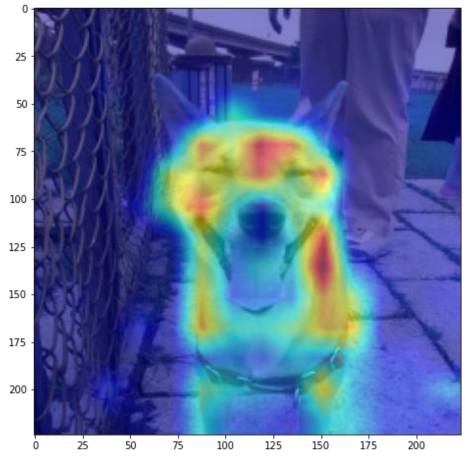


   Analysis: In this task, I applied cosine similarity on position embedding, from the visualization graph of position embedding, we can find that the position embeddings of nearby patches are relatively similar, and they are very close in the same column or same row. That is, patterns across rows or columns have similar representations.

3. Attention maps

Original pictures	With mask(After cropped)
 0 100 200 300 400 0 50 100 150 200 250 300 350	

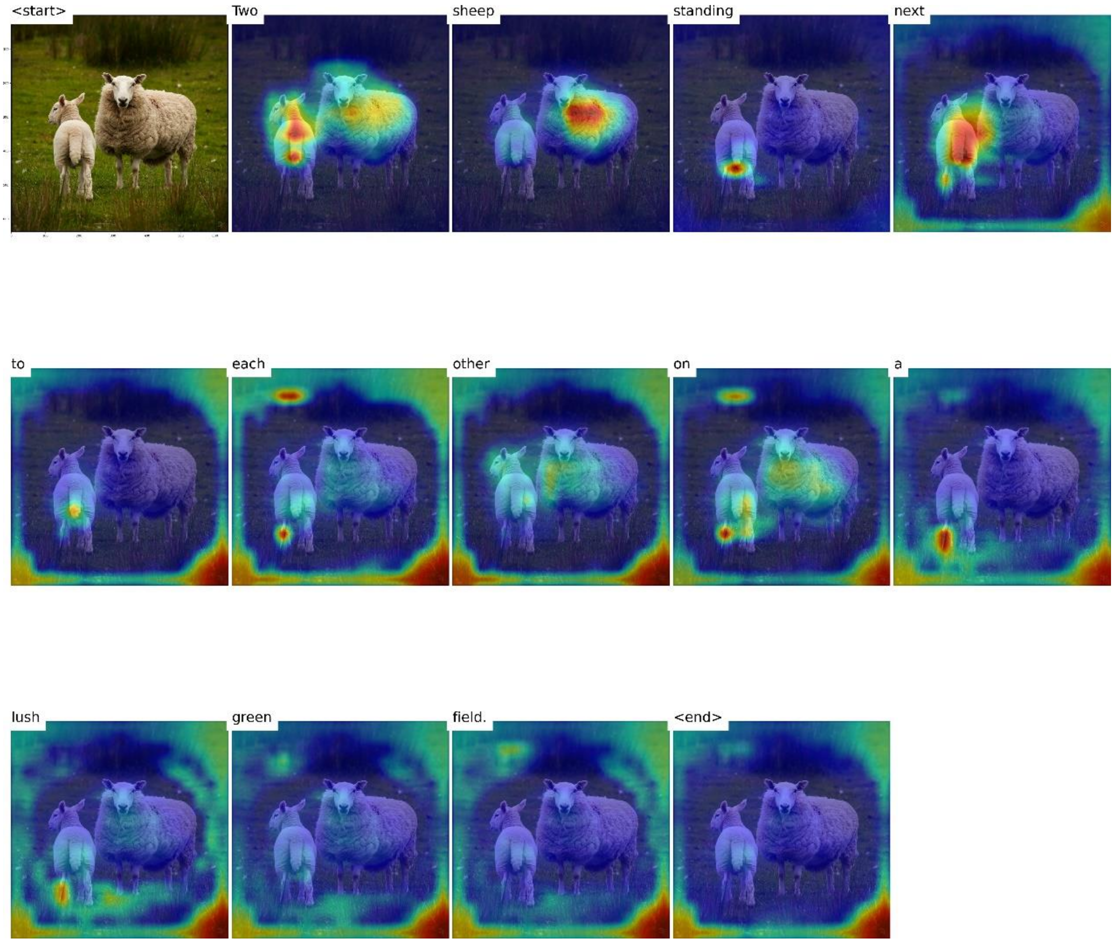
0
50
100
150
200
250
300
0 100 200 300 400
	

 0 100 200 300 400 0 50 100 150 200 250 300 350	

Analysis: In this task, I used the “vit_base_patch16_224” pretrained model from timm package. All the pictures are resized to (256, 256) then cropped into (224, 224). I found that if the target(such as dogs) has obviously different in luminance or color saturation, the result will be better. We can compare the second set and the third set, the third set show a better result. Besides, if the background is too colorful such as the first set, it will also affect the result of attention map.

P2、Image captioning

1. Visualization



Analysis: According to the five outputs of visualizing the cross-attention between images and generated captions, I found that it is harder to separate two captions which are conjunction or adjective used to describe background. Besides, it is easier to recognize objects, colors, and quantity of something.

Reference:

https://github.com/hila-chefer/Transformer-Explainability.git

https://colab.research.google.com/github/hirotomusiker/schwert_colab_data_storage/blob/master/notebook/Vision_Transformer_Tutorial.ipynb

Collaborators: B08902134 曾揚哲