## Problem1

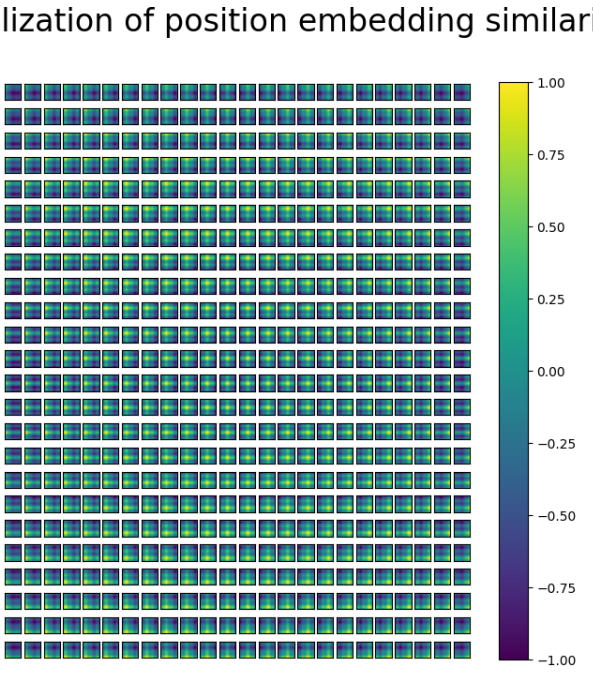
- **Report the accuracy on validation set and discuss the result with different settings.**

  Accuracy = 0.95

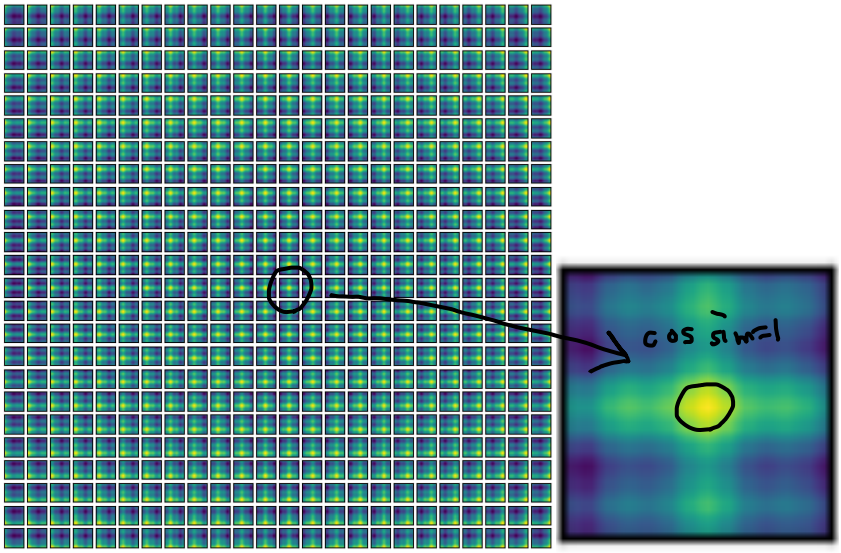
  The final implementation details are below:
  Pretrain model = vit_base_patch16_384 (from timm)
  Patch_size = 16
  Image_size = 384
  Batch_size = 16
  Learning rate = 1e-5
  Optimizer = Adam(betas=(0.9, 0,98))
  Epoch = 20

  In the training process I found that image size = 384 is better than image size = 224, and as for the patch size, 16 is better than 32. I think this may because a larger image size and a smaller patch size can store the features of original image more clearly.
  Another hyperparameter that has a significant impact on the model is learning rate. At the beginning, I set the learning rate between 1e-4 and 1e-3, and the training result did not even exceed the sample baseline. After I lowered the learning rate to around 1e-5, the result was visibly improved.

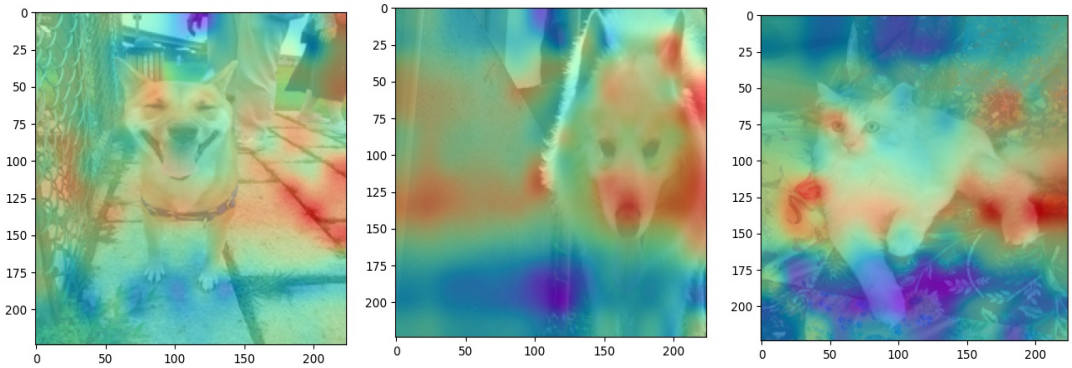
- **Visualize position embeddings of your model and discuss the result**



In the visualization, there are 24*24 = 576 patches. One cell shows cosine similarity between itself and all the other embeddings. The position embedding vectors show the distance within the image thus neighboring ones have high similarity. As we can see the patch in the middle of all patches, the cosine similarity in middle equals to 1(the most green place, or yellow, not sure), which means the position of this patch in the original image is in the center point. On the other hand, the farther from the center point, the cosine similarity is lower(the most purple place).



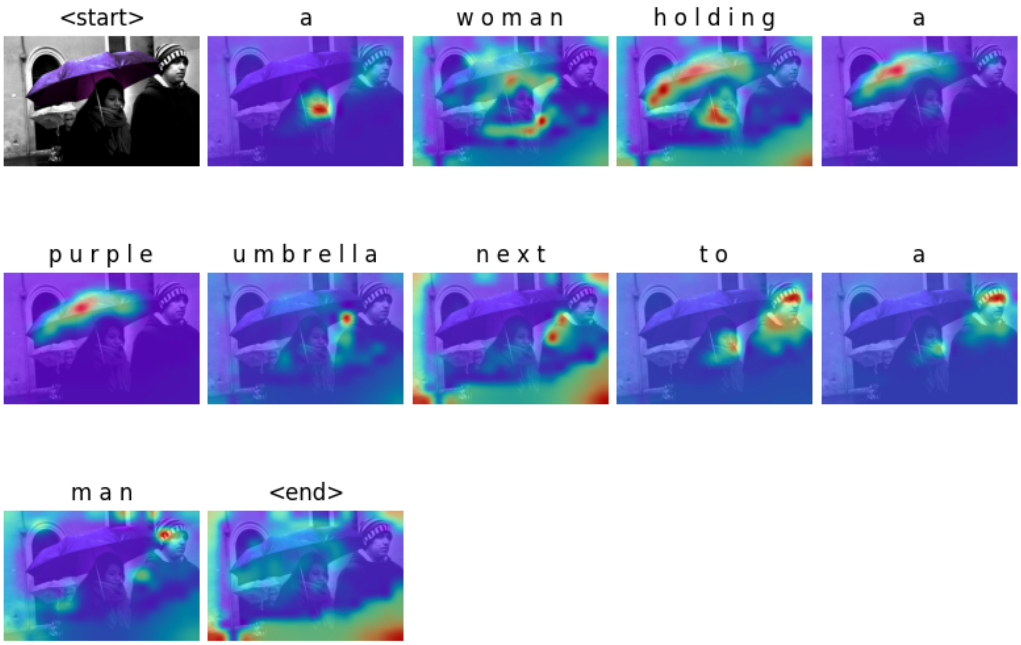

- **Visualize attention map of 3 images and discuss the result**



The results show that the attention is generally on the object that we want to focus, but they are still not very accurate.

**Problem2**

- **Analyze the predicted caption and the attention maps for each word.**




In this result, the caption is reasonable. But there are still some deviation which the attended region does not reflect the corresponding word in the caption.
At the word "**holding**", it accurately pay attention on the woman's hand and the umbrella. On the contrary, at the word "**umbrella**", the attended region does not on the umbrella.

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

In this problem, I learned how the attention mechanism works and how to visualize the attention weight. The most difficult process to me is to understand the model architecture. Since the pretrained model architecture is complicated, it is hard to find out where the attention mechanism works.

##################################
Discuss with:
R09521601, R09521603

Reference:
https://arxiv.org/pdf/1706.03762.pdf
https://github.com/rwightman/pytorch-image-models/
https://colab.research.google.com/github/hirotomusiker/schwert_colab_data_storage/blob/master/notebook/Vision_Transformer_Tutorial.ipynb#scrollTo=nI6rRunEO6bI
https://github.com/saahiluppal/catr
https://blog.csdn.net/qq_37541097/article/details/117691873
##################################