Deep Learning for Computer Vision

Homework #3

經濟四 李品樺 B07303024

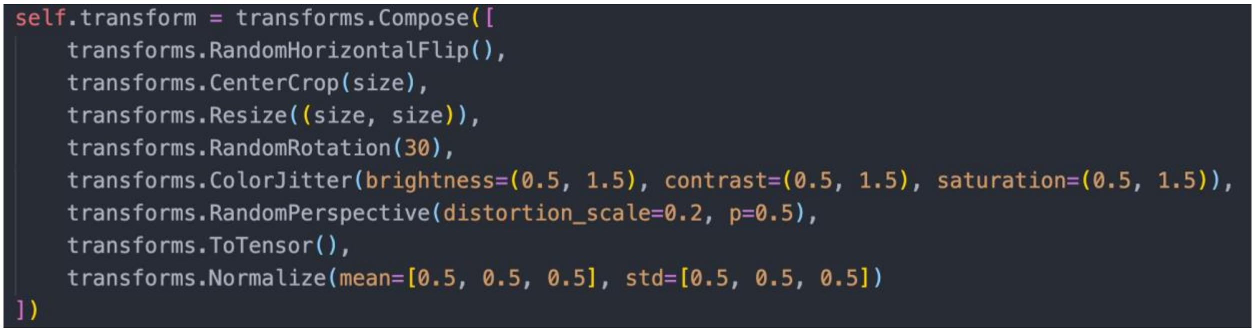
Collaborator: B07901039 劉知穎

Problem 1: Image Classification with ViT

1. Report accuracy of your model on the validation set.

(a) **Discuss and analyze the results with different settings**

Since we are using a pretrained model, the learning rate should be chosen carefully. Using the pretrained ViT model, B_16_imagenet1k, the predicted result is able to reach about 89% in 5 epochs using Adam as optimizer with learning rate 1e-6. If we use the same model architecture but with learning rate1e-4, then the accuracy will increase fast at first but drop back to only 3% eventually. Due to GPU memory limits, we can only use batch size smaller than 32. Adding the following data augmentations and resizing the input images to size (256, 256) can improve the predicted accuracy for 2% to 3%. Moreover, we apply gradient accumulation in order to imitate larger batch size. Setting the iterations of accumulating gradients as 15, and we can reach an accuracy over 94%.



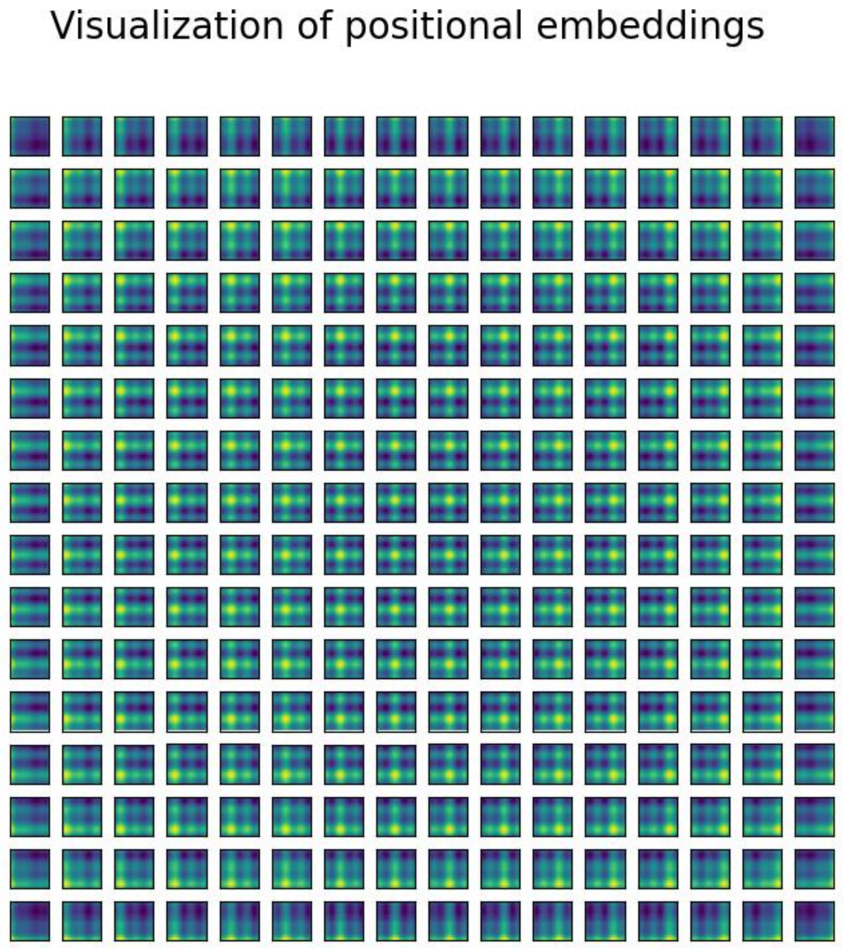
```
self.transform = transforms.Compose([
    transforms.RandomHorizontalFlip(),
    transforms.CenterCrop(size),
    transforms.Resize((size, size)),
    transforms.RandomRotation(30),
    transforms.ColorJitter(brightness=(0.5, 1.5), contrast=(0.5, 1.5), saturation=(0.5, 1.5)),
    transforms.RandomPerspective(distortion_scale=0.2, p=0.5),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5])
])
```

(b) **Clearly mark out a single final result for TAs to reproduce**

94.467%

2. Visualize position embeddings

(a) Visualize cosine similarities from all positional embeddings

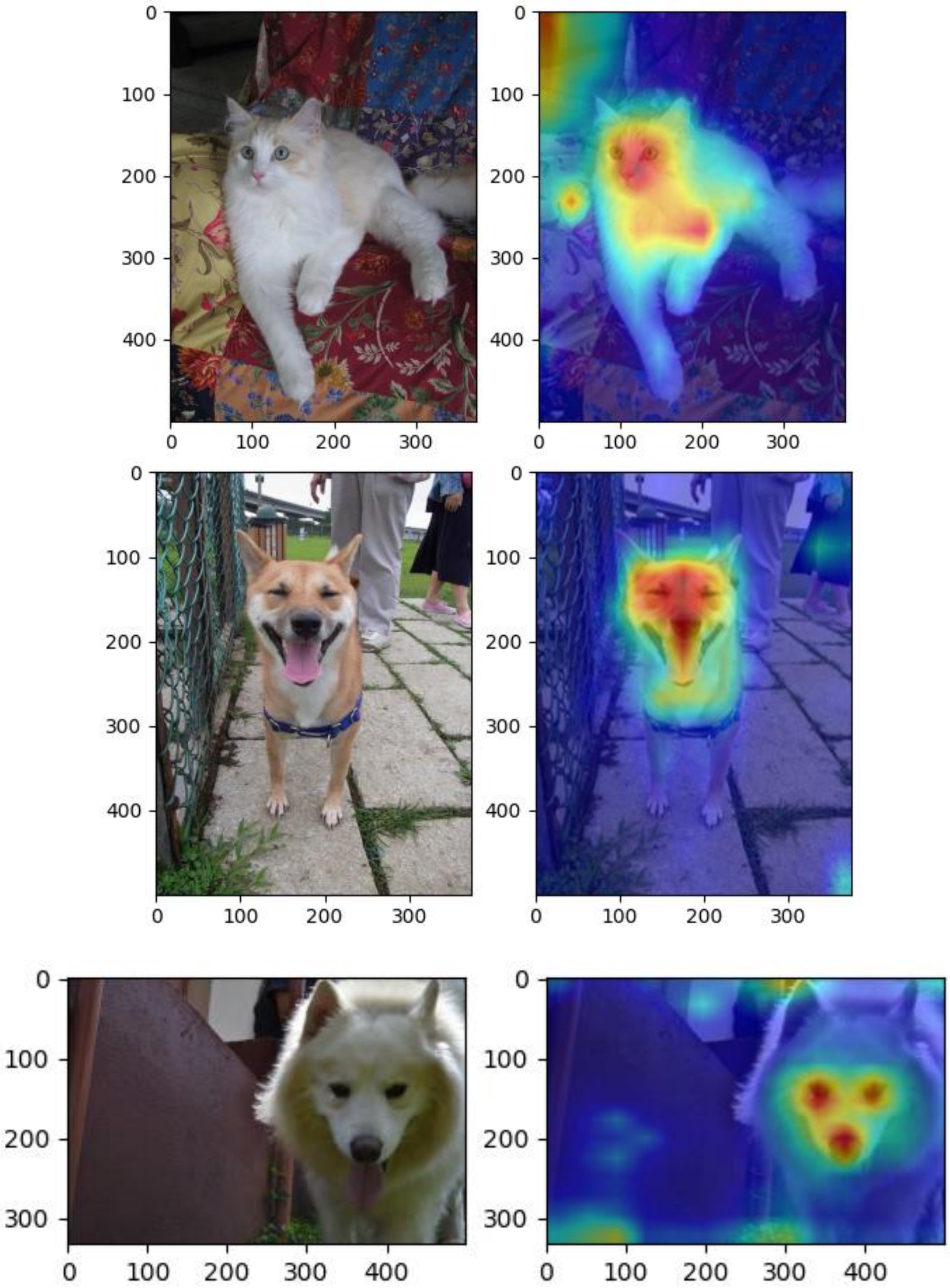

Visualization of positional embeddings

(b) Discuss or analyze the visualization results

In the figure above, color yellow indicates the cosine similarity is 1, and indigo blue means the cosine similarity is -1. From the figure, we can find out the cosine similarities of the positional embeddings preserve the grid structure of the image. For example, the image in the left upper corner has the largest cosine similarity in the left upper corner. This is reasonable since positional embeddings encodes the relative position of the input patches.

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)

(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**



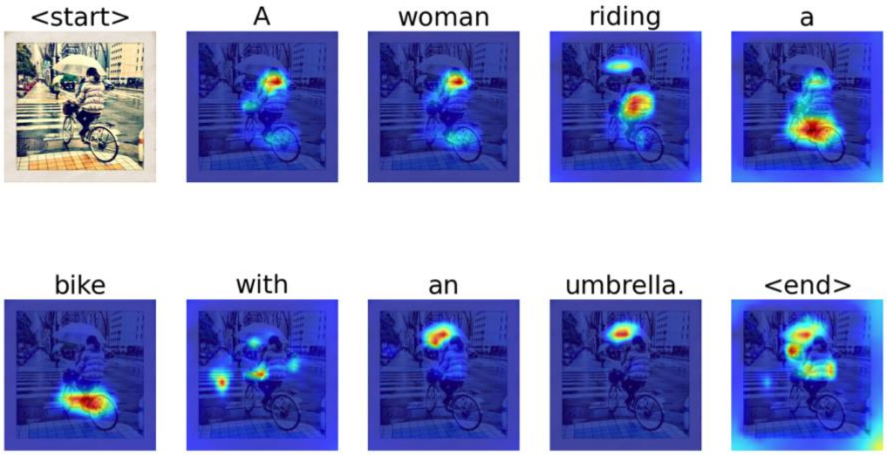
(b) **Discuss or analyze the visualization results**

In the figures above, the area of red means the attention result is more important (the weights are larger), and the area of blue is less important (the weights are smaller). From the figures, we can find out that the areas around the faces of the pets are more important. For 26_5064.jpg, it is possible that cats are more flexible, thus the pose of the cat is not so important. For 29_4718.jpg, face and the neck of the dog are important to the model. For 31_4838.jpg, the most important part is the eyes and nose of the Samoyed dog, instead of the entire dog. Although the visualization of the attention looks quite intuitively, there are still some minor mistakes of the visualization result. For example, the top right corner of 26_5064.jpg is the background but the weights are large.

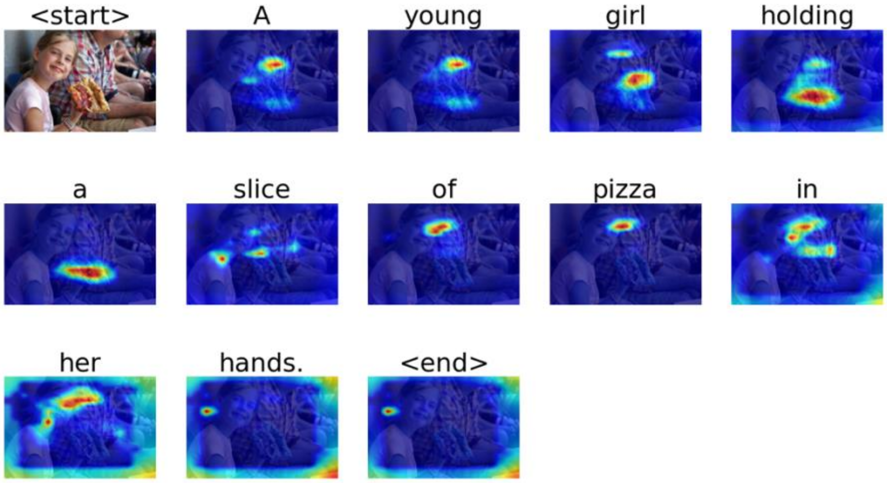
Problem 2: Visualization in Image Captioning

1. **For the five test images, please visualize the predicted caption and the corresponding series of attention maps in a single PNG output.**

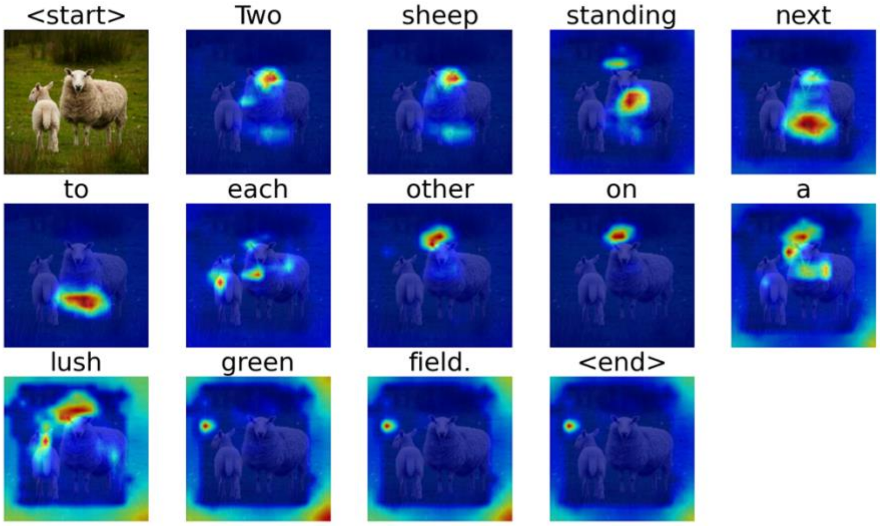
- **bike.png**



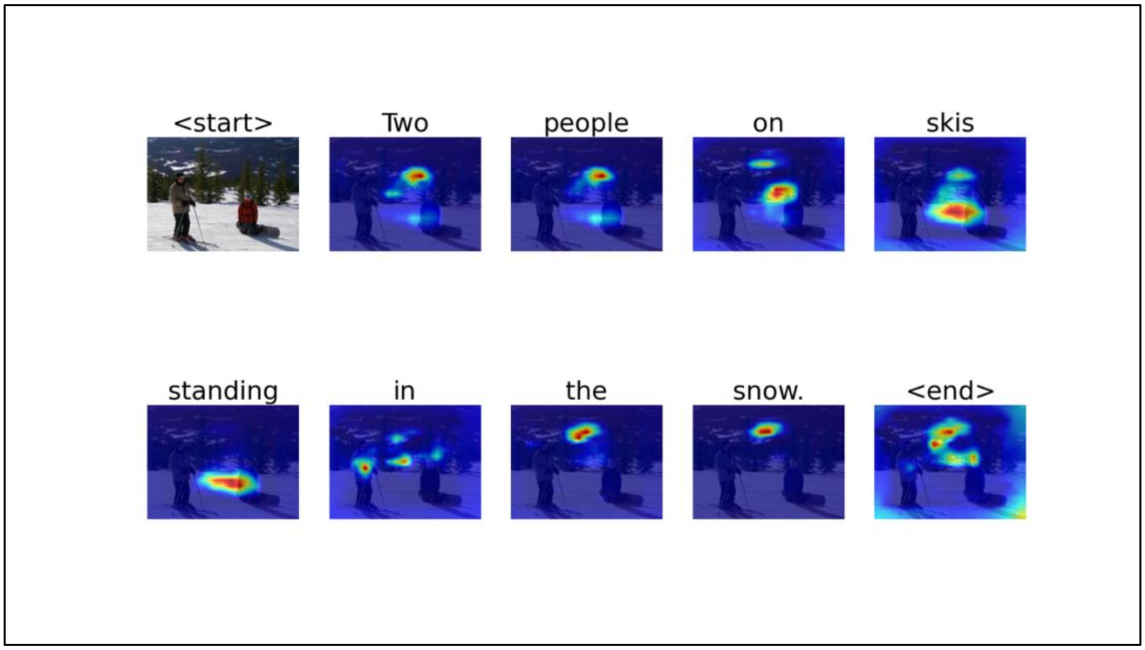
- **girl.png**



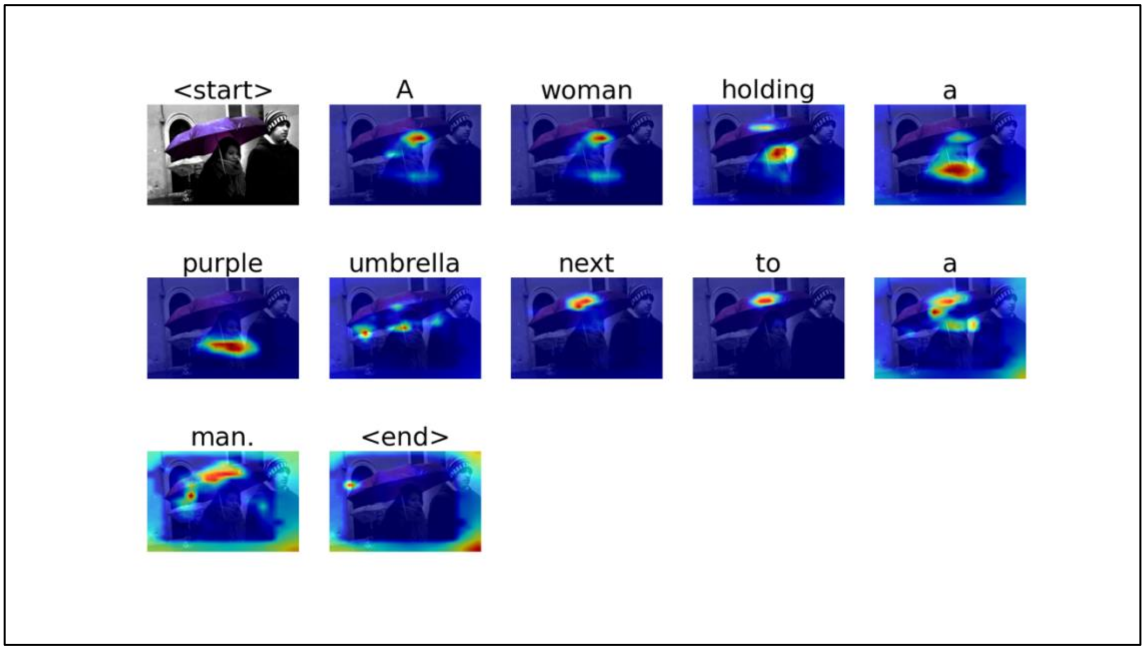
- **sheep.png**



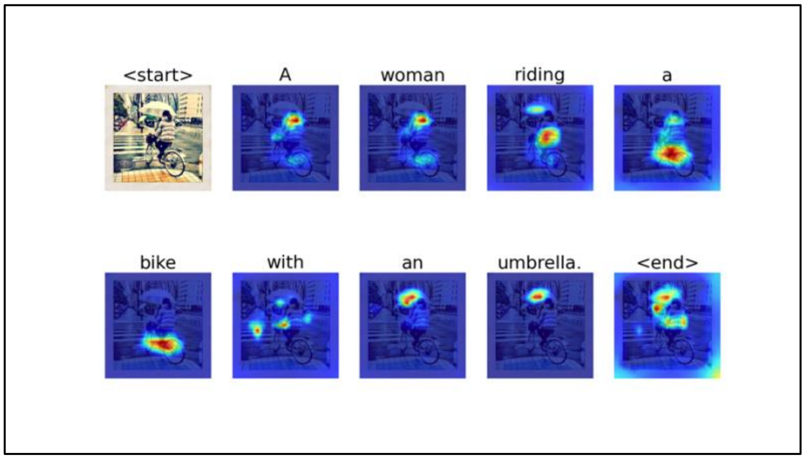
- **ski.png**



- **umbrella.png**



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?

In the image **bike.png** above, color red means the weights of the attention results of those pixels are higher, on the other hand, blue means the attention results are lower. Overall, the attended maps are reasonable with respect to the predicted captions. Furthermore, we can observe the images of the articles and the corresponding nouns are alike. For example, the images of "A" and "woman" are similar; "a" and "bike" are similar; and the images of "an" and "umbrella" are similar. "riding" attended to the body of the woman instead of the bike. The preposition "with" attended to several areas around the woman. For the <end> token, the attended map not only focus on the main object in the image but also the surrounding area of the image.

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

From doing this problem, I have learned to trace the output of decoder, encoder, feature extractor so as to further register forward hook, and find out the attended weights. The complicated structures of the pipeline and models are hard to follow at first, however, printing out the shape of output is helpful for solving this problem. Moreover, the output result of the attended image is in wrong color at first, I find out that we need to normalize the pixel values back to 0-255 before applying attention map.