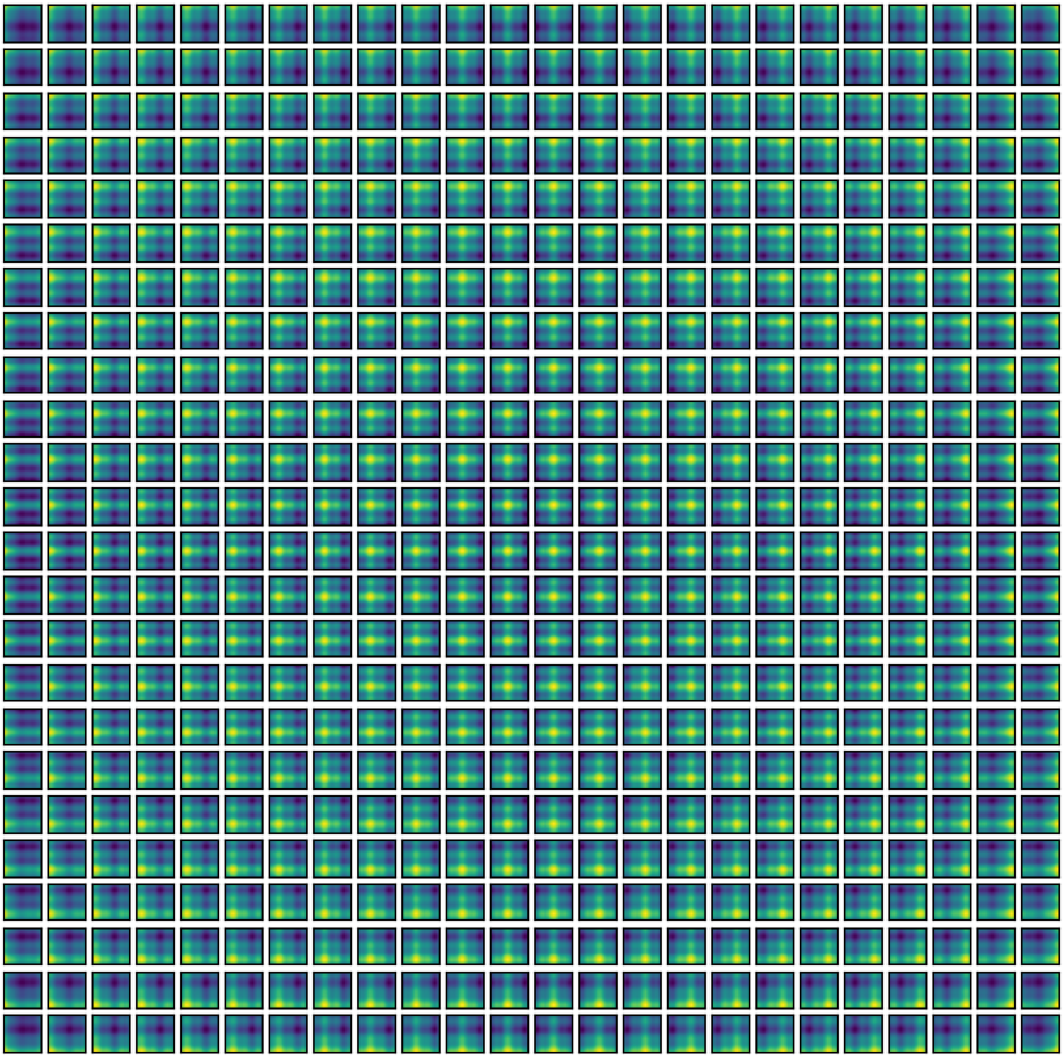
DLCV Fall 2021 HW3

Problem 1

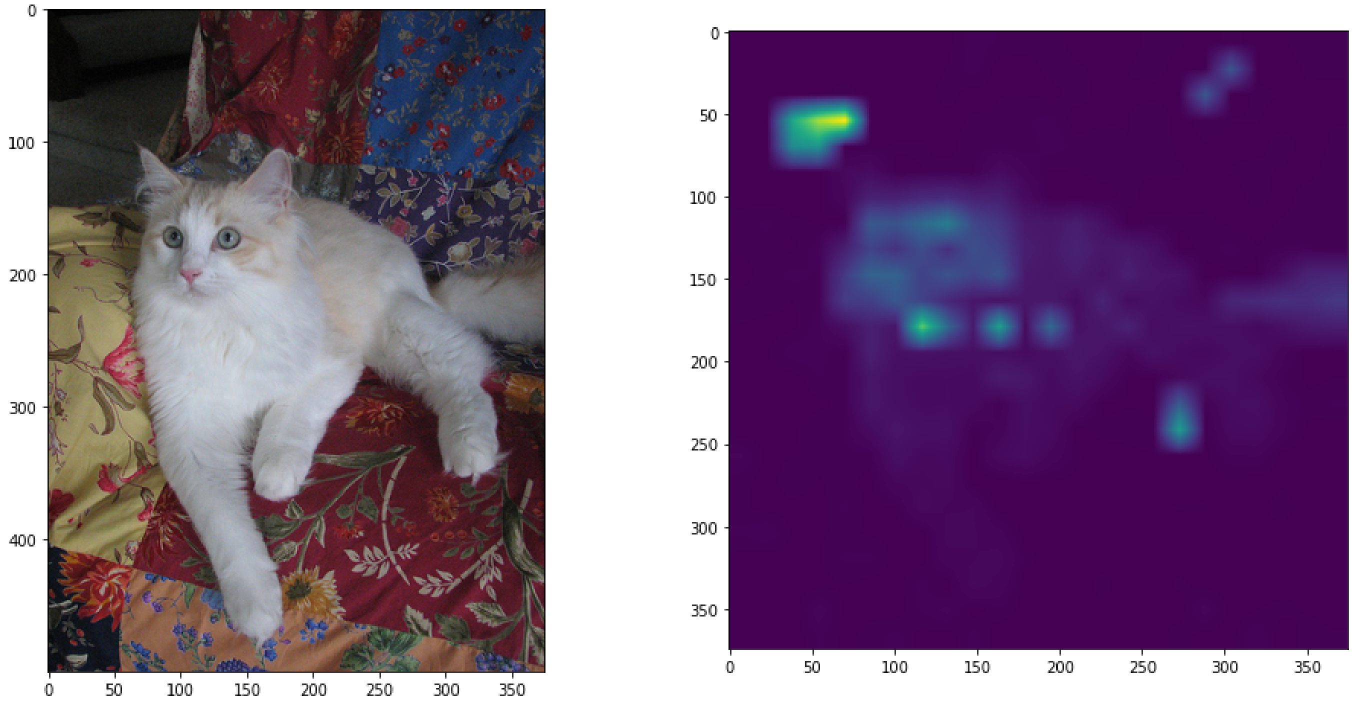
```
1. Report accuracy of your model on the validation set. (TA will reproduce
your results, error ±0.5%) (10%)

    a. Discussion
        Obviously, the task requires the pretrained module. While it's easy
to reach the simple baseline with arbitary input size of module, it turned
out hard to reach over the harder one. So I tried the module
'vit_base_patch16_384'. Due to the limitation of GPU, the max of batch size
is 4, so I accumulate the gradient as batch size = 512. I also tried lots
of transform and normalization. The best result is shown in training code,
and the most important result is adding normalization, which help me passed
the score finally. I use adam optimizer with config lr=1e-4,
weight_decay=1e-1 and CosineAnnealingLR scheduler.

    b. Accuracy: 0.942 (0.9419999718666077)

2. Visualize position embeddings (20%)
    a. Visualization
```

Visualization of position embedding similarities

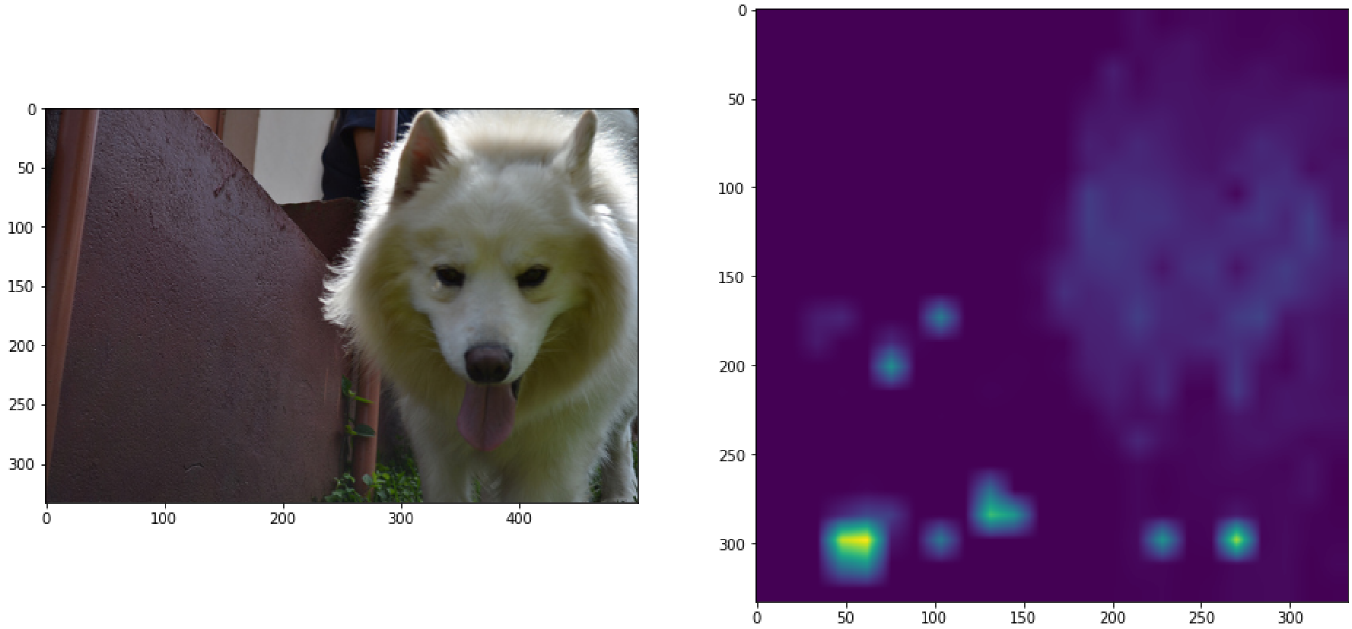


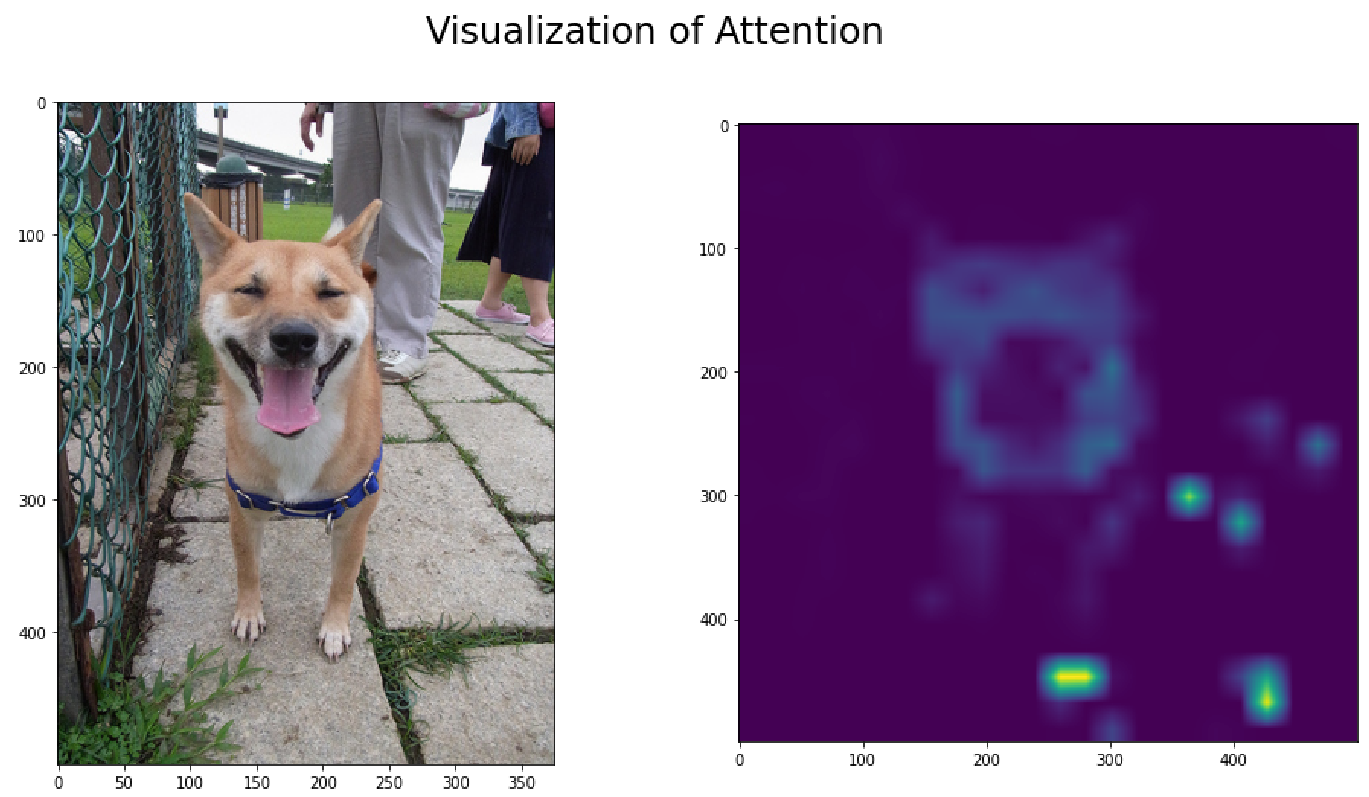
```
    b. Discussion
        The result follows the expectation, which is that the light portion
of each patch is almost the same as its position in original image.

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. Visualizaiton
```

Visualization of Attention



Visualization of Attention

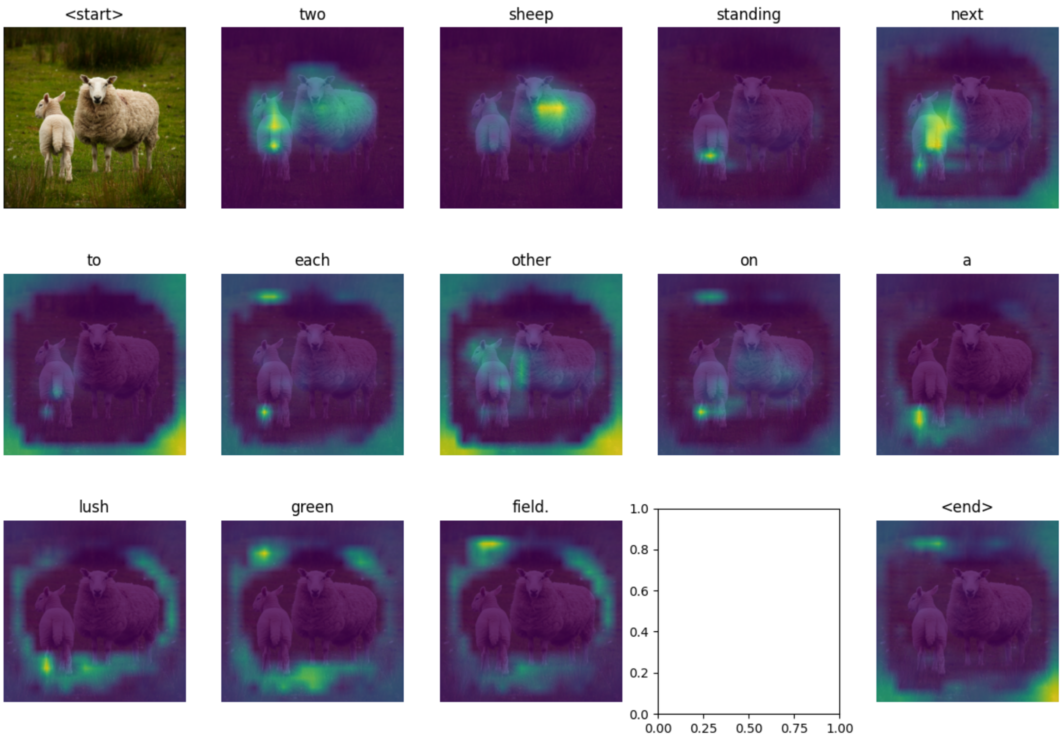




```
    b. Discussion
        The highlight parts of attention weights are mostly same as the
important object in each picture. The accuracy is amazing, and the contours
almost draw the animals !
```

Problem2

```
1. Visualization for 5 picture
```



```
2. Discussion

    a. I think the work is well done on this task. In the first one, the
highlight of two sheep stands for "two"; The sheep in right hand side,
which faced directly to camera, so it's suitable for module to recognize;
The area betweeen two sheep implies "next", which is interesting to me that
machine can learn preposition. "green", "field" also correctly point out.
The words who possess less meaning are difficult for machine to learn, but
it's fine since those can't be shown practically in daliy life too.

    b. One of the  difficulties is finding the correct layer to be hooked,
since the model is complicated, so the correct ways to point the layer
should be tried. The other is how to correctly plot layers of images.
```