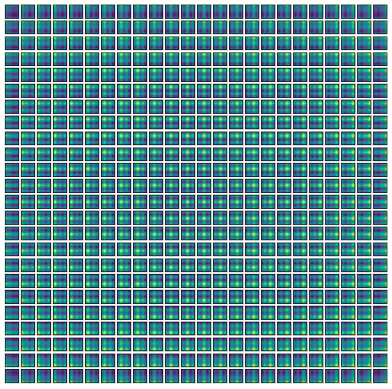
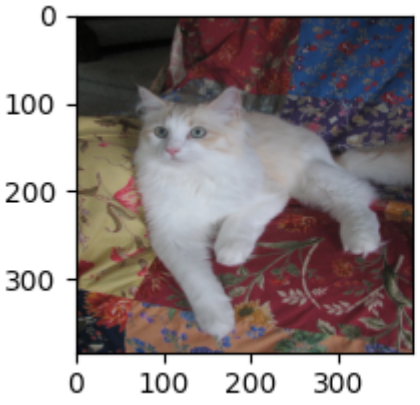
NTU DLCV 2021 Fall hw3 r09942091 謝承軒

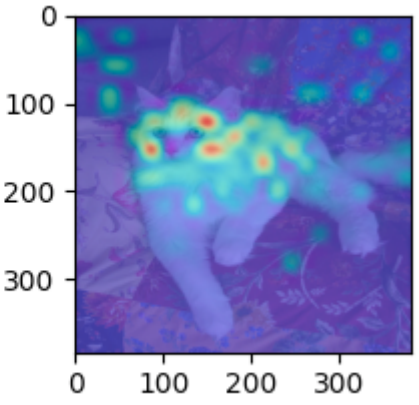
Problem 1.

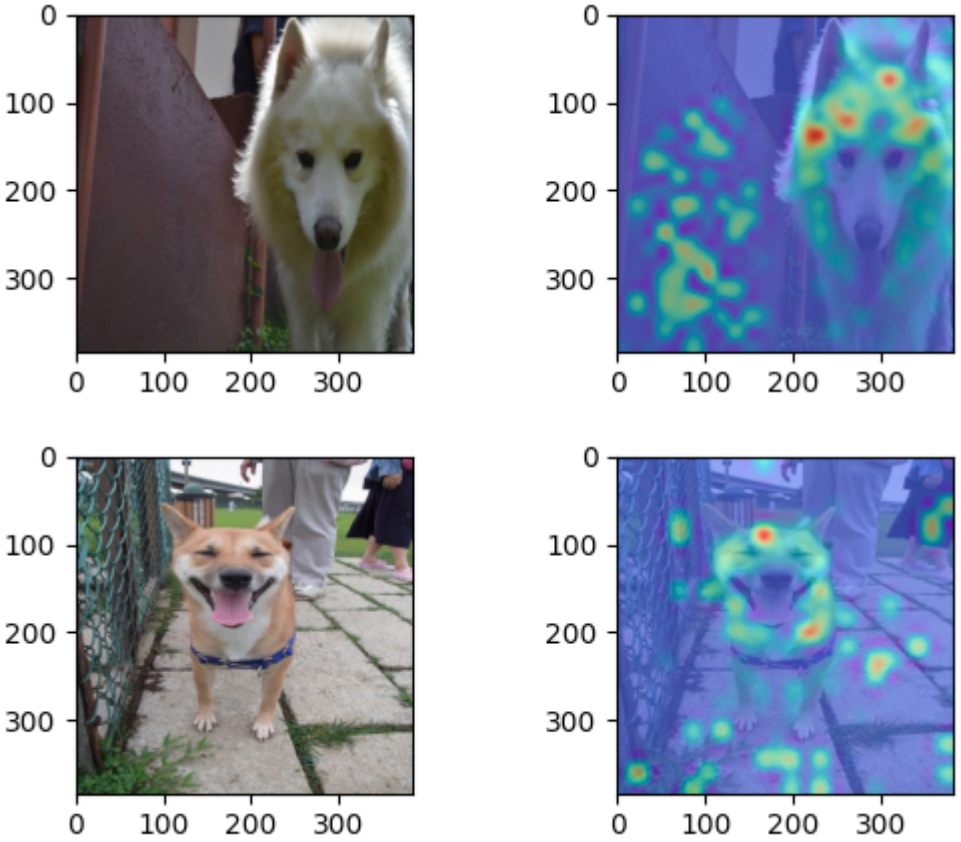
1. accuracy: 94.4%
   I have used the “B_16_imagenet1k” in pytorch_pretrained_vit
   During my training, I fine tuned the model using Adam optimizer with lr=0.0001 and lr scheduler, and fine tuned the model second time using SGD optimizer with lr=0.01.
   And the augmentation I used was
   a. RandomResizedCrop((0.7, 1.0))
   b. ColorJitter(brightness=0.3)
   c. RandomHorizonFlip(p=0.5)
   d. RandomRotation(15)

2. 

   From the result of cosine similarity of positional embedding, it is observed that even when turning the image to paches, the model has recovered the grid structure of the original images.
   I think it does not look good in the surroundings because I do the RandomResized augmentation.

3. 

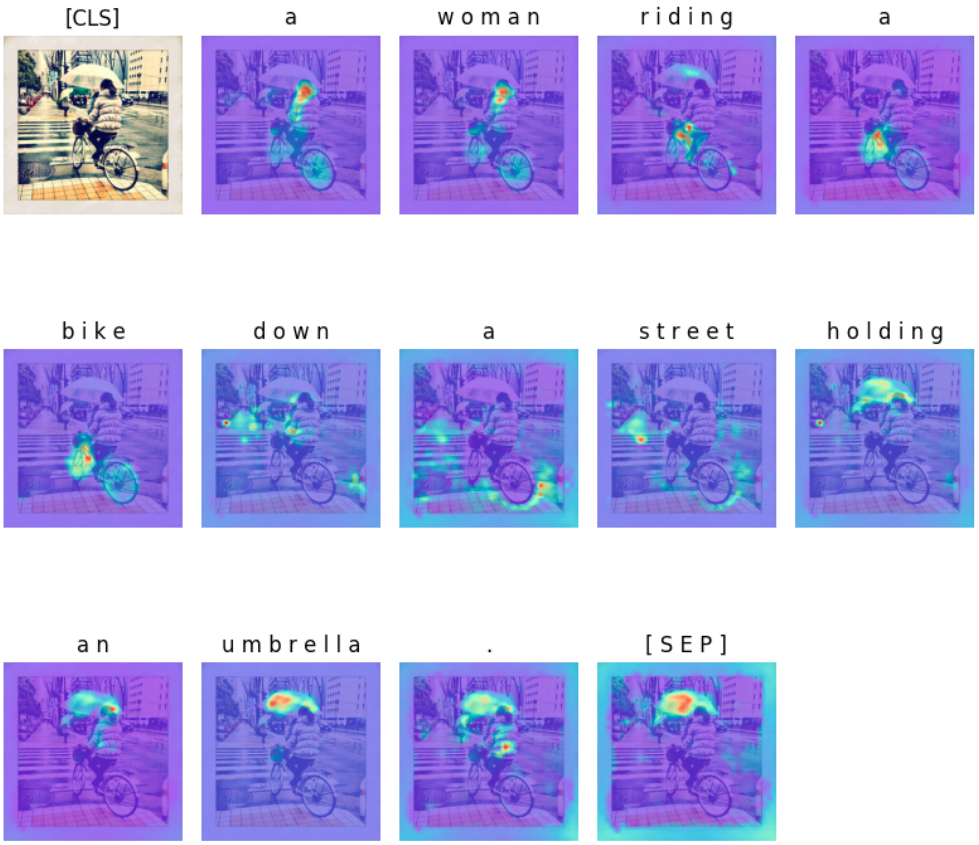




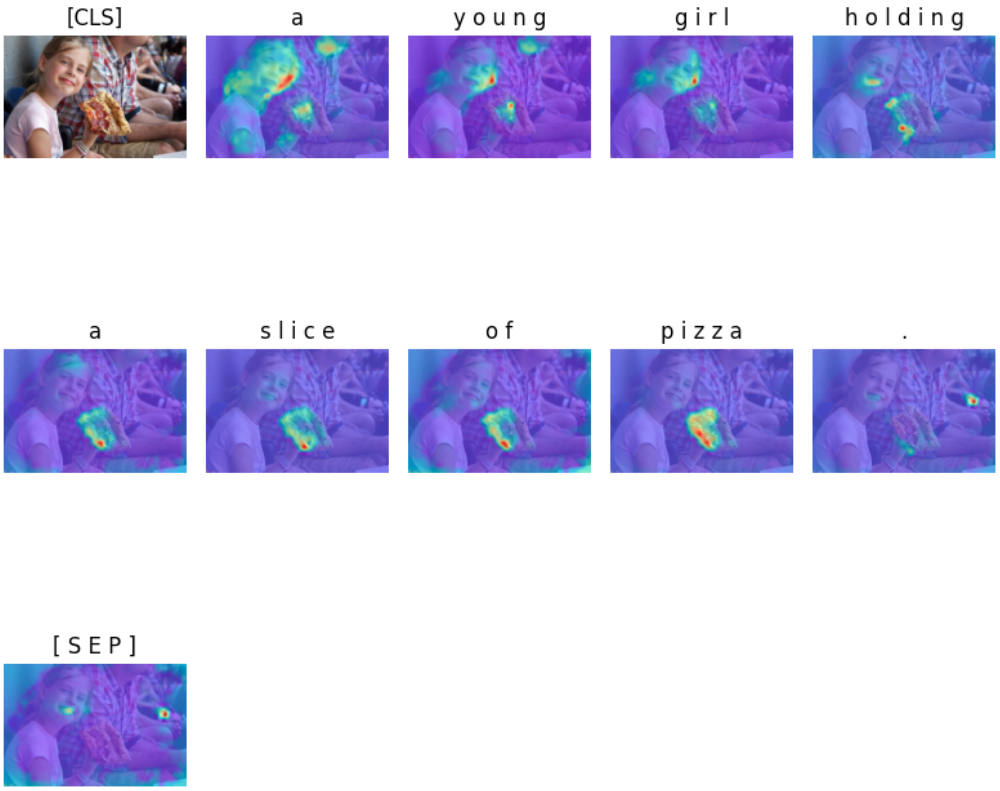
In my attention map, it seems to focus on the outline of animals more, instead of their faces. I think it is because this task is to find different species of cat and dogs, and most cats/dogs have similar faces, so the outline can decide the species.

Problem 2.

1. bike:



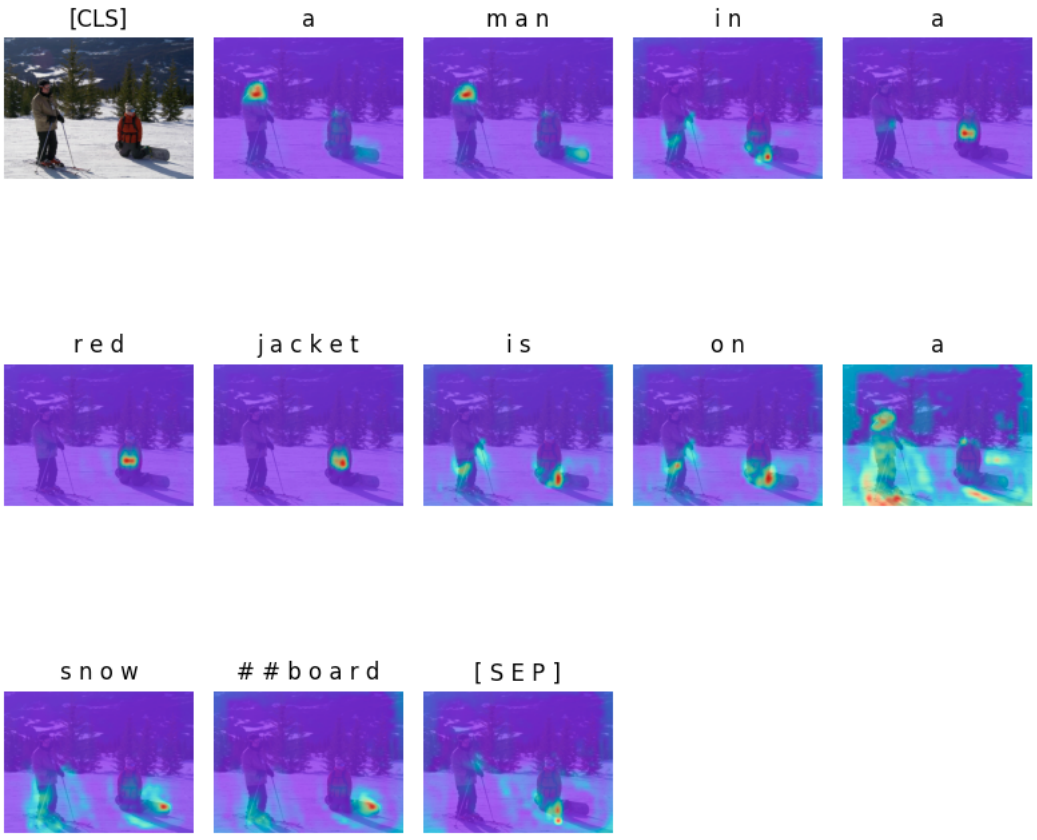
giel



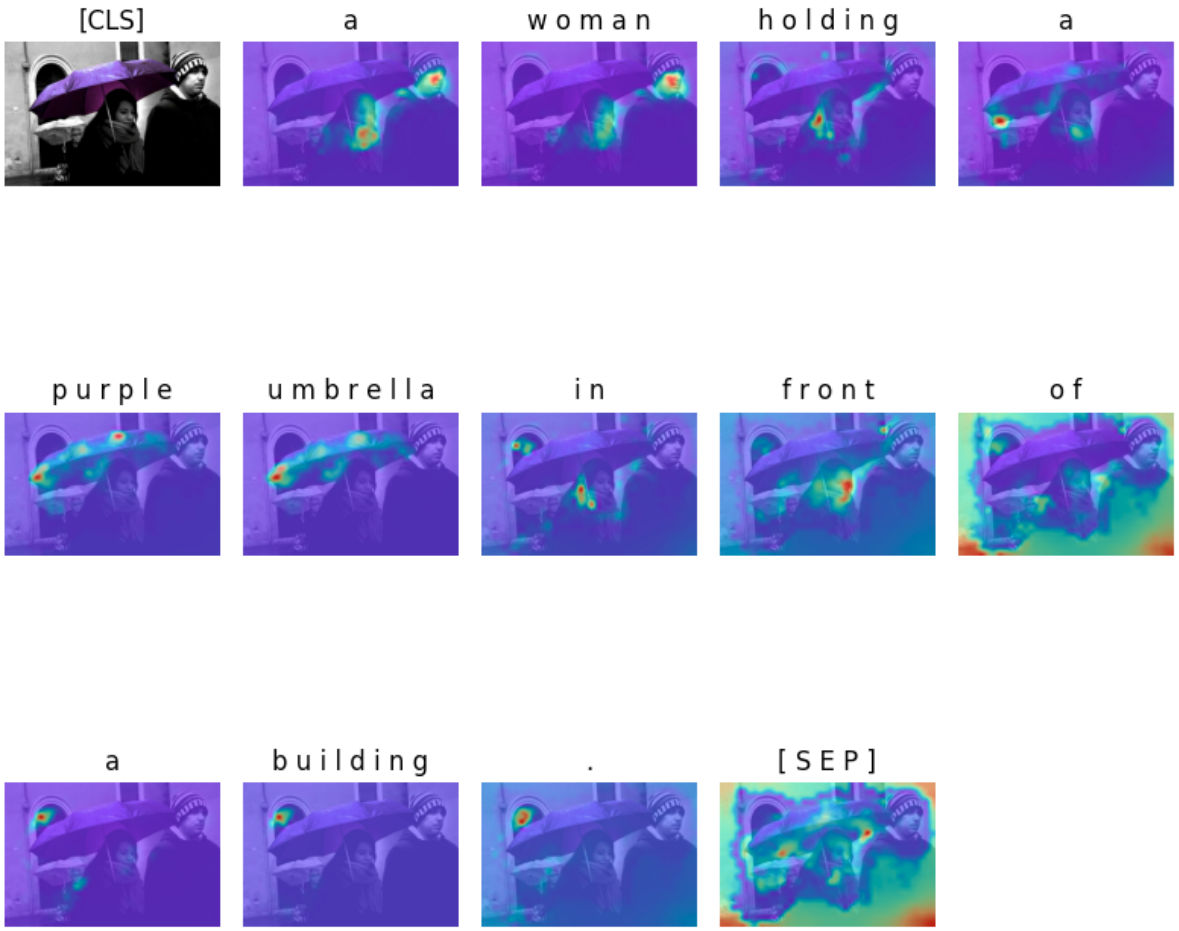
sheep


[CLS]
a
sheep
standing
in
a
field
of
grass
.
[SEP]

ski


[CLS]
a
man
in
a
red
jacket
is
on
a
snow
##board
[SEP]

umbrella



take the”umbrella” as a sample, it is not reasonable for words “**a, woman**”, and not sure for the words “**a, building**”.
In my other samples, it can be seen that the model is not good at predicting more than one subject, e.g. sheep,ski, umbrella.
However, most of the attention maps are reasonable, such as the “purple umbrella” and they seems perfect.

2. The most difficult parts are learning the whole code, especially finding out the multi-head attention and understanding it. Also, it’s my first time to do the heatmap so I took some time to figure it out.
   But it is pretty good to understand it because I am more clearly about the working of transformers and ViT.