DLCV HW3

學號 r08945050

姓名 蔡宇晴

Problem 1: Grading - Report

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

a. Discuss and analyze the results with different settings (e.g. pretrain or not, model architecture, learning rate, etc.) (8%)

我使用timm 中"vit_base_patch16_224"模型，並使用pretrain model 後再進行fine tune。接著使用Sharpness-Aware Minimization (SAM)這個optimizer進行訓練。一開始的預設lr 為0.1，模型的loss直接發散。後來我調整了learning rate 至 0.00001結果則會收斂再acc = 0.946左右。

```
optimizer = SAM(model.parameters(), base_optimizer, lr=0.00001, momentum=0.9)

epochs = 200
```

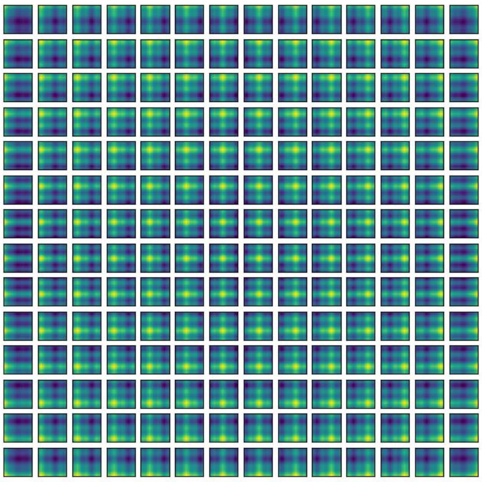
b. Clearly mark out a single final result for TAs to reproduce (2%)

```
acc is 0.9453333020210266
```

2. Visualize position embeddings (20%)

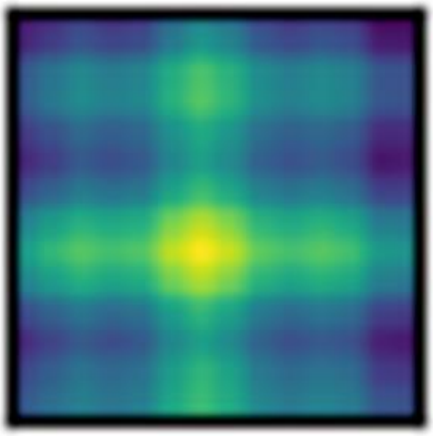
a. Visualize cosine similarities from all positional embeddings (15%)

Visualization of position embedding similarities



b. Discuss or analyze the visualization results (5%)

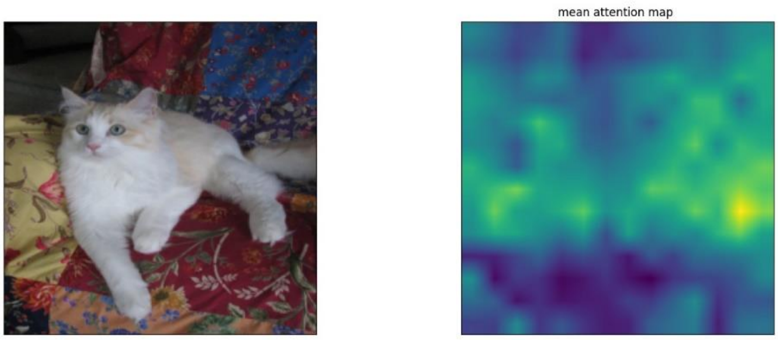
上圖中每個grid都會有一個較黃的部分(weight 高)，而每個gird 從左上至右下的過程，其黃色區域則是隨著由左上至右下的變動。其原因是由於向量與自己的內積會最大，因此左上角的patch會使得左上角的grid的左上角weight較高。



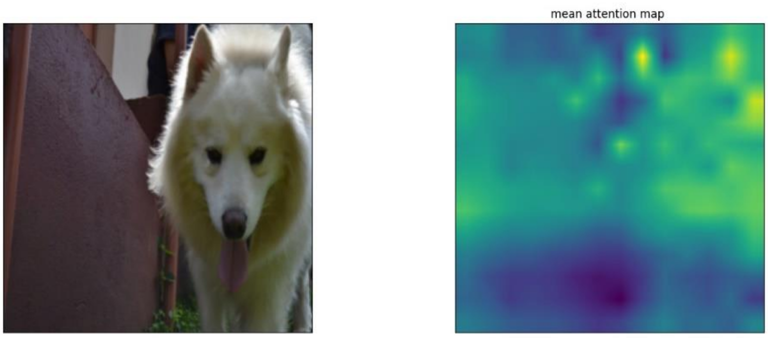
而中間patch則會使grid中間較亮如右圖。

p1_data/val/29_4718.jpg, p1_data/val/31_4838.jpg) (20%)

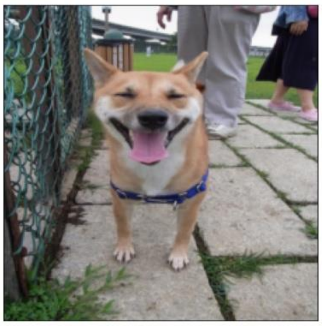
a. Visualize the attention map between the [class] token (as query vector) and all patches (as key vectors) from the LAST multi-head attention layer. Note that you have to average the attention weights across all heads (15%)

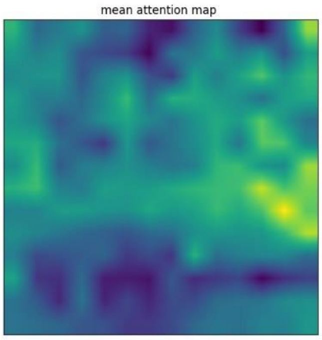


(26_5064.png)



(29_4718.png)

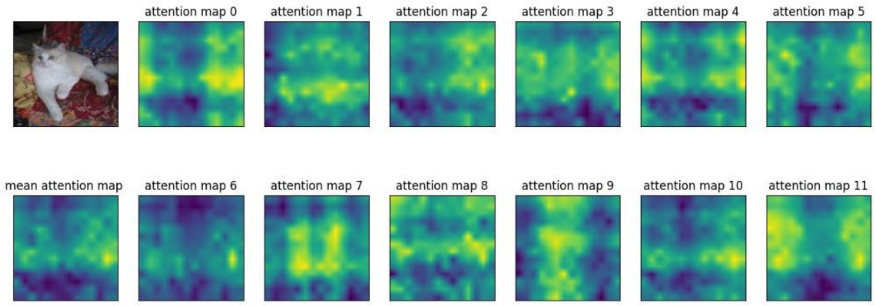


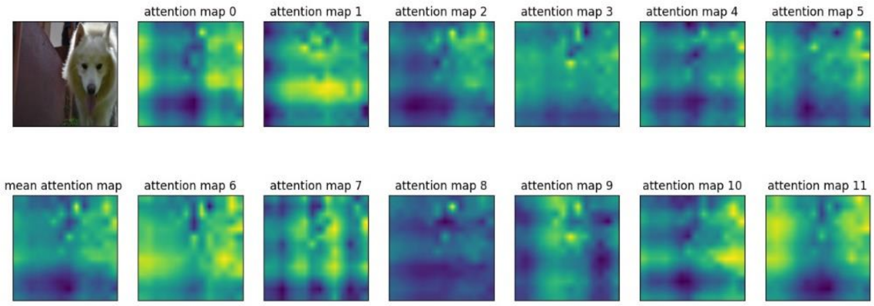


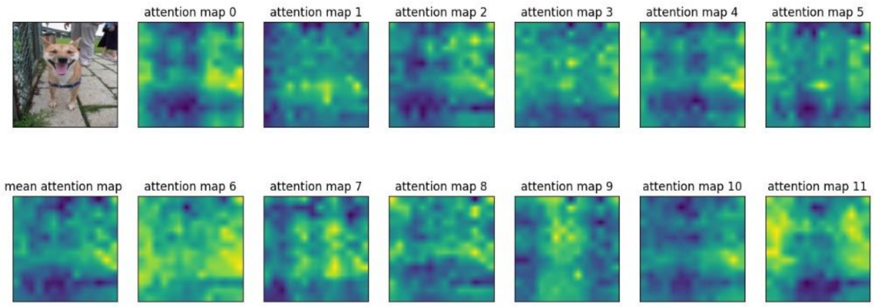
(31_4838.png)

b. Discuss or analyze the visualization results (5%)

以下是每個head對應的attention matrix 總共有12個，且包含一張mean attention map (左下)。其中，(26_5064.png)、(29_4718.png)的mean attention map 其結果有對應至欲分析的物體，而(31_4838.png)則看不出weight集中在物體上的感覺。然而從12張attention map 可以看出，每個head都會有對應所分配較高weight的位置，因此即便mean attention map 沒有對應物體，也不能代表分類器不能通過其他head找出其他head來決定正確的物體類別。在(29_4718.png)中可以看出mean attention map 在狗耳朵附近有較高的權重。





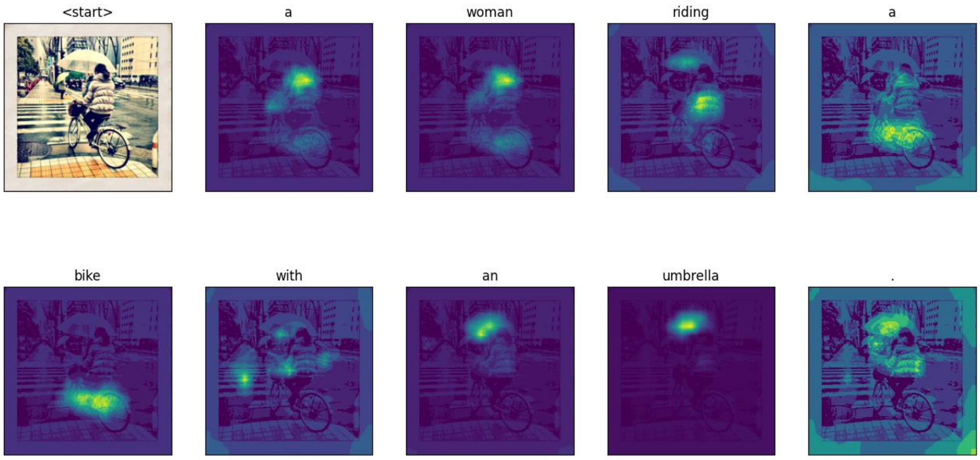


Problem 2: Visualization in Image Captioning

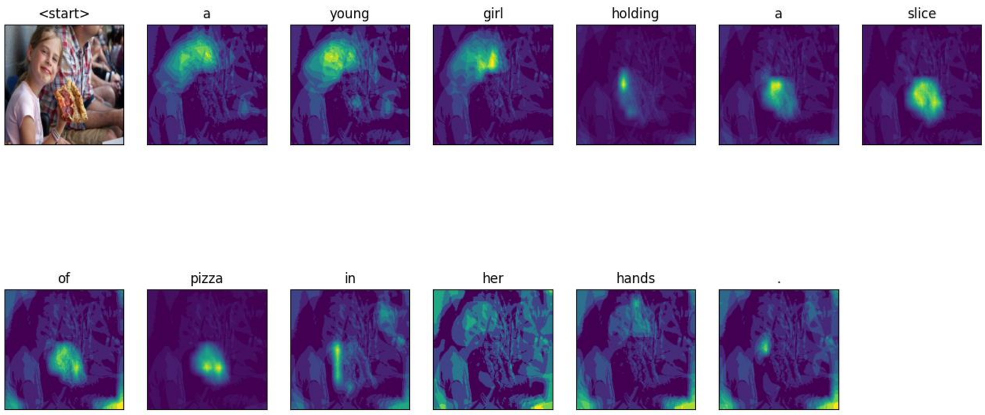
1. For the five test images, please visualize the predicted caption and the corresponding series of attention maps in a single PNG output. TA

will reproduce your visualization results with your bash script. (10%)

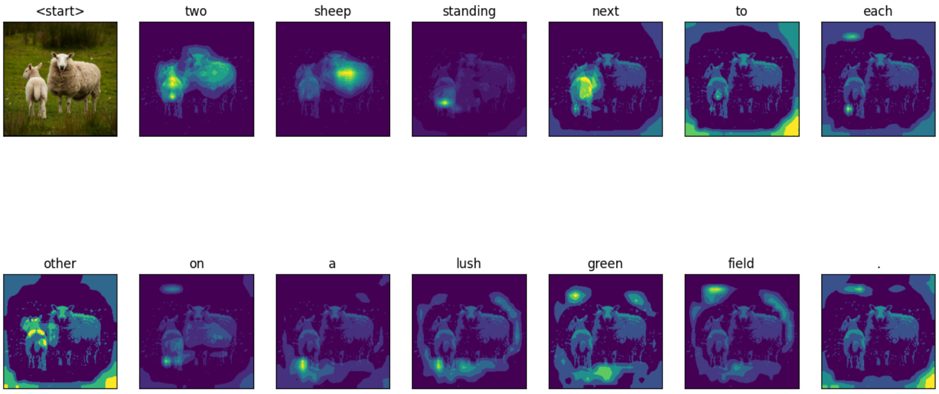
a. Save the five visualization results (PNG images) in the specified folder directory.



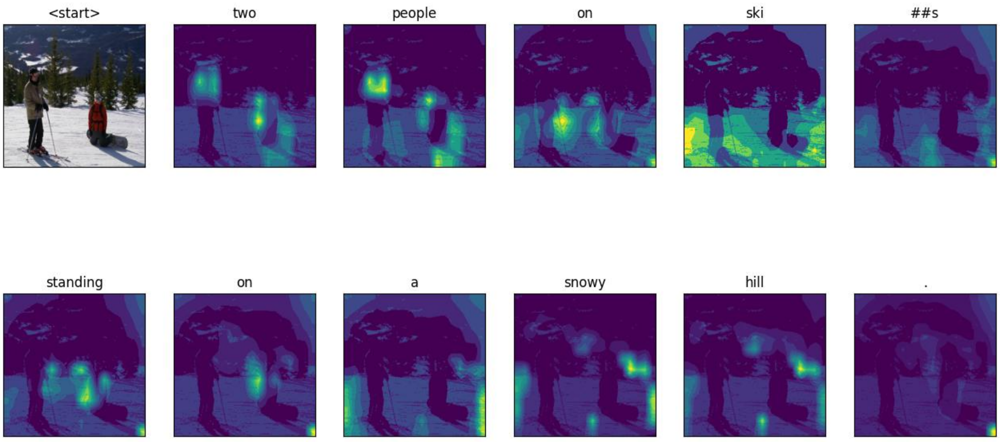
(bike.png)



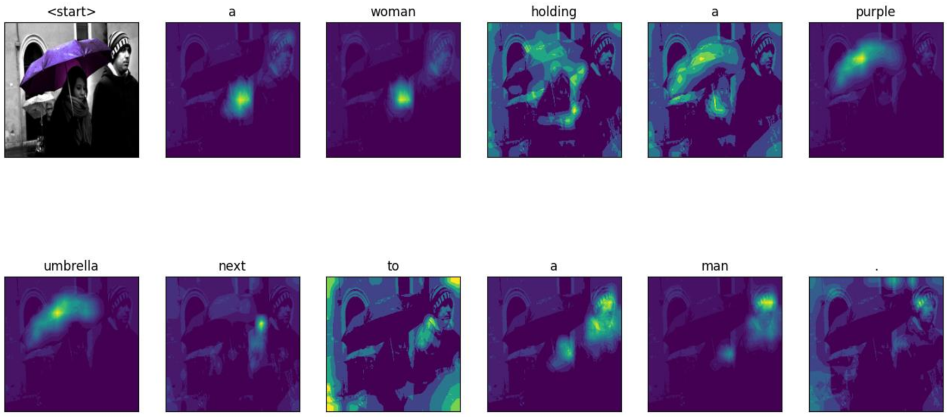
(girl.png)



(sheep.png)



(ski.png)

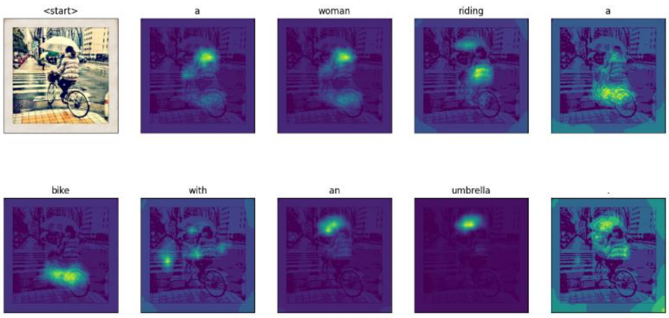


(umbrella.png)

b. Name your output PNG images as follows (same as the input filename):

2. Choose one test image and show its visualization result in your report. (10%)

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?



上圖中，幾乎每個字都有對應至所代表的物體，例如woman、bike、umbrella，然而一些抽象的詞彙我認為並不能好好看出是否有對應至某件物體，例如with。

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

一開始我認為每個attention matrix 的每個row都具有對應的attention map，因此我將對於每行row進行sum amd mean 的動作，然而最後我發現，或許是model為了結構上的簡易or計算成本不影響，因此即便MultiHeadAttention 的輸出具有128*247維度，也不代表每個row都是有意義的attention map，事實上只有當前預測的單字所對應的attention matrix 的row才有意義(包含先前產生的row)。另外我發現由於此decoder 會參考先前產生的詞彙再進行預測，因此即便是相同的詞彙，也不一定會有相同的attention map ，如上圖的兩個”a” ，所代表的意義不同，左邊的”a”代表女人，右邊的”a”代表腳踏車。