Problem 1: Grading - Report (50%)

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%)

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

Ans:

a. 我一開始的時候並沒有去查看 paper 上面的內容，所以就直接把 model 丟進來 train，結果發現都 train 的不是很好，那時候我用的 optimizer 是 Adam，learning rate 設為 0.01。最後我去看了 paper 上的數據後，我發現他是用 SGD 來當作 optimizer，而且用了 warm-up 的方法來訓練，但最後我是參考 paper 上面的作法，使用 SGD 且 learning rate 為 0.0003、momentum 為 0.9，並且使用 pretrain 的 model(“vit_small_patch16_384”)這樣得到的效果就非常好了，訓練兩個 epoch 就可以再 training set 上得到 96%的準確度，而至於訓練到 validation set 到過了 strong baseline 則花了 6 個 epoch.

b. **Accracy : 94.67%**

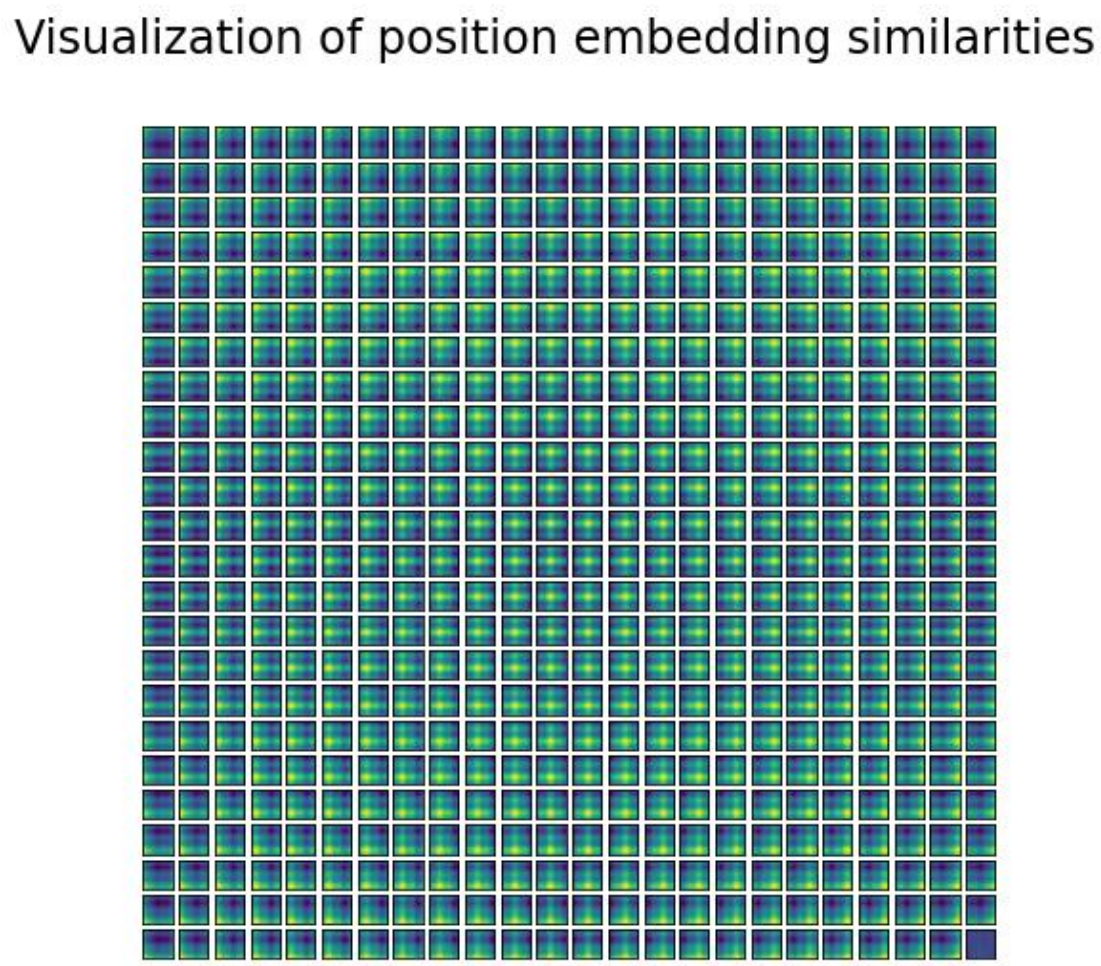
2. Visualize position embeddings (20%)

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

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

Ans:

a.



b. 在向量空間中，我們計算距離的方法有很多種，像是常用的歐基里德距離或是曼哈頓距離，而在本題中我們計算的是 cosine similarity，代表著我們在乎的是向量之間的方向而不是長度，可以來衡量向量之間的相似度，當算出來的 cosine similarity 越大，表示說兩向量之間的夾角越小，反之亦然。
而在 visualize position embedding 的 cosine similarity 之後，可以看到圖上把他切成 24*24 個 patches，中間區塊的亮面積代表著向量之間的相似度比較高，而相較於邊緣區域黯淡的地方則是相似度比較低的地方。

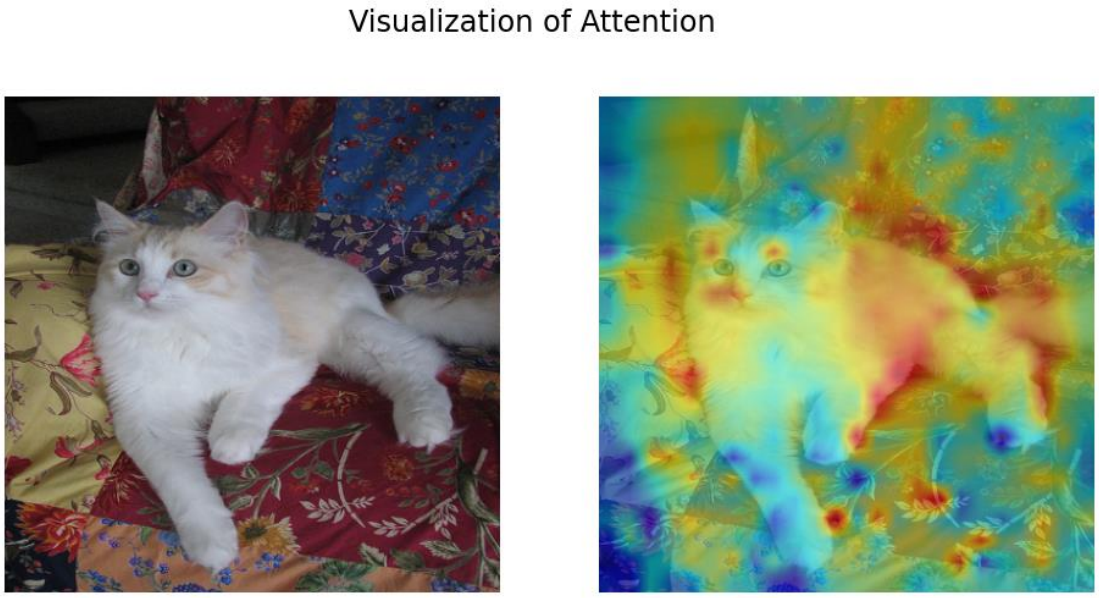
3. **Visualize attention map of 3 images (p1_data/val/26_5064.jpg, p1_data/val/29_4718.jpg, p1_data/val/31_4838.jpg) (20%)**

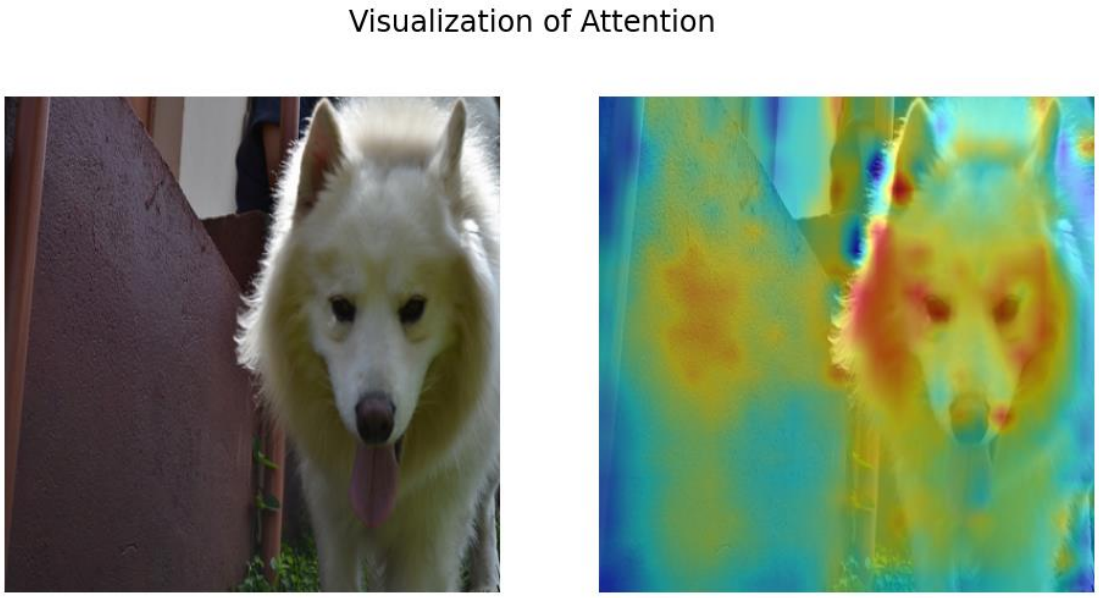
   a. Visualize the attention map between the [class] token (as query vector) and all patches (as key vectors) from the LAST multi-head attention layer. Note that you have to average the attention weights across all heads (15%)

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

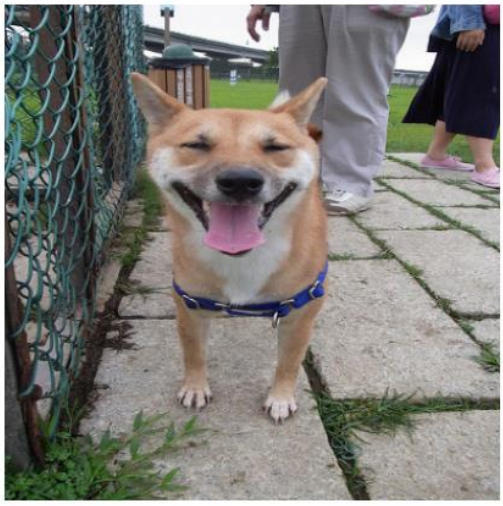
Ans:

a.





Visualization of Attention



b. 我通過 output 出來的 csv 檔，發現這三張圖片預測出來的 class 都有分類正確，而我們從圖上也可以看到，在判斷貓和薩摩耶的時候可以很明顯看出 transformer 的注意力都有集中在動物的身體上，而在判斷柴犬的時候，雖然注意力比較沒有集中，但仍舊是有在柴犬的身上有注意力的集中點，也應證了為甚麼最後判斷的結果也會是正確的。

Problem 2: Grading - Results (20%)

1. For the five test images, please visualize the predicted caption and the corresponding series of attention maps in a single PNG output. (10%)

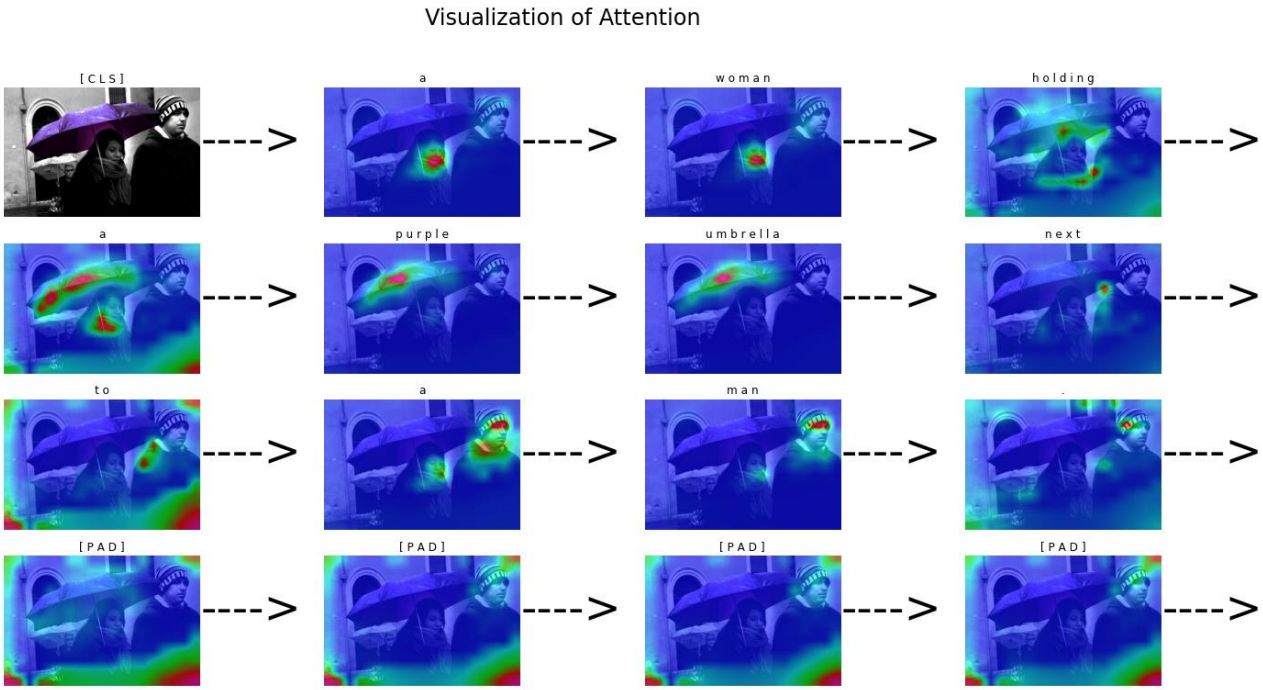
a. Save the five visualization results (PNG images) in the specified folder directory.

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?

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



Ans:

a. 我選擇 output 的圖片為 dataset 裡面 umbrella.jpg 這筆資料來做，可以看到說在經過 model 之後他所 output 出來的字句為: a woman holding a purple umbrella next to a man. 而我將輸出固定為 4*4 張子圖片來表示整個 series，透過觀察可以發現在字輸出為 woman、purple、umbrella、man 的時候，可以看到 decoder 是有把注意力真的集中在該對應的物體上，而進一步觀察我發現一件有趣的事情，就是有三個 a 這個單字出現，他卻能夠對應到不同的物體上，可以明顯看出在 a woman 的時候，a 會集中注意力在 woman 上面，而在 a purple umbrella 的時候，這時的 a 又會集中在傘上面，所以可以應證說我們字句的前後順序是彼此有相關聯的，我們的這個 model 也有很好

的學到這件事情。

b. 在這題當中我花比較多的時間在於說理解它整個的架構是怎麼寫的以及要從原本的 model 裡面拿哪些東西出來，最後我把 multiheadattn 的最後一層 layer 的 weight 取出來，以及 backbone 出來的 pos 來當作圖片要 resize 成的尺寸，在做完此題之後我覺得這個東西還蠻酷的。

Reference:

https://arxiv.org/abs/2010.11929

https://colab.research.google.com/github/hirotomusiker/schwert_colab_data_storage/blob/master/notebook/Vision_Transformer_Tutorial.ipynb#scrollTo=koV0ey9Y1f_4

https://github.com/saahiluppal/catr

https://stackoverflow.com/questions/56275515/visualizing-a-heatmap-matrix-on-to-an-image-in-opencv