Deep Learning for Computer Vision homework 3 R09945021 洪怡庭

Problem 1

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%)

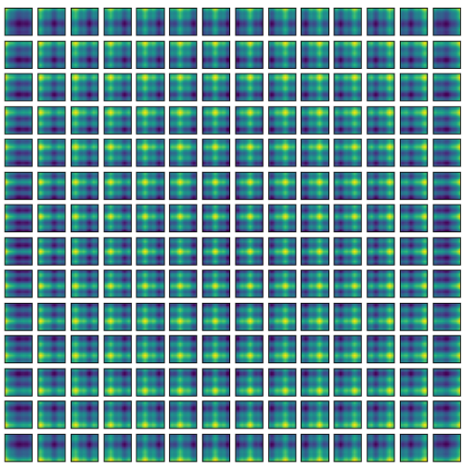
	Architecture	Optimizer	Accuracy
Model 1	Pretrained ViT-Small	Adam	0.926
Model 2	Not Pretrained ViT-Base	Adam	0.075
Model 3	Pretrained ViT-Base	Adam	0.928
Model 4	Pretrained ViT-Base	SAM (base: Adam)	0.942

   i. 只有 Model 2 是未經 pretrain 的，經過 100 epoch 訓練後準確度仍只有 0.075，可以看出 pretrain 的重要性，利用大量的 ImageNet 資料訓練好的參數作為初始參數，能大量加快模型的收斂時間。
   ii. Model 1(ViT-Small)使用 embed_dim=384, num_heads=6 而 Model 3(ViT-Base)使用 embed_dim=768, num_heads=12，雖然 Model 3 參數量較多，但在此任務上準確度上沒有太大的差異。
   iii. Model 4 嘗試使用 SAM 作為 optimizer(base optimizer 仍為 Adam)，使得在 minimize loss function 的同時，也考慮 loss landscape 的平坦程度，訓練結果可見模型較快收斂也有更好的 generalization 能力，比較沒有 overfitting 在 training set 上，在 validation set 上也有較高的準確度。

   b. Clearly mark out a single final result for TAs to reproduce (2%)
   Model 4 (Pretrained ViT-Base, SAM) : Accuracy = 0.942

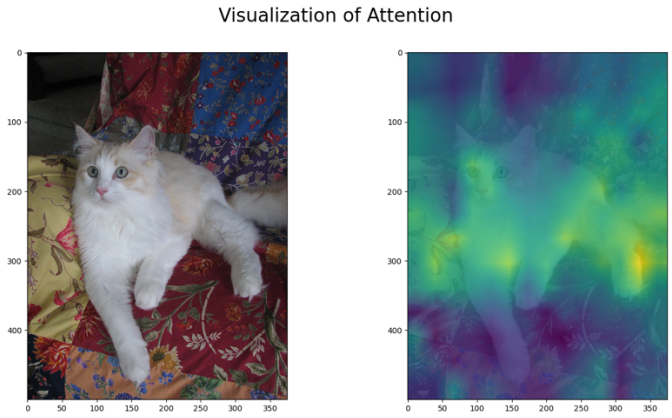
2. Visualize position embeddings (20%)
   a. Visualize cosine similarities from all positional embeddings (15%)

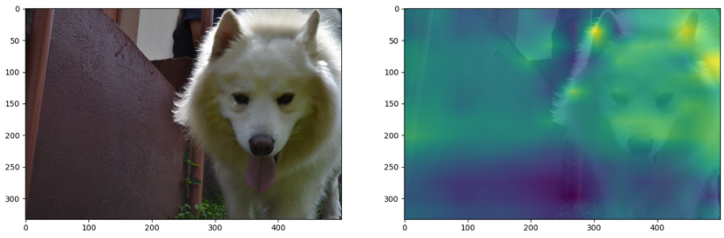
Visualization of position embedding similarities

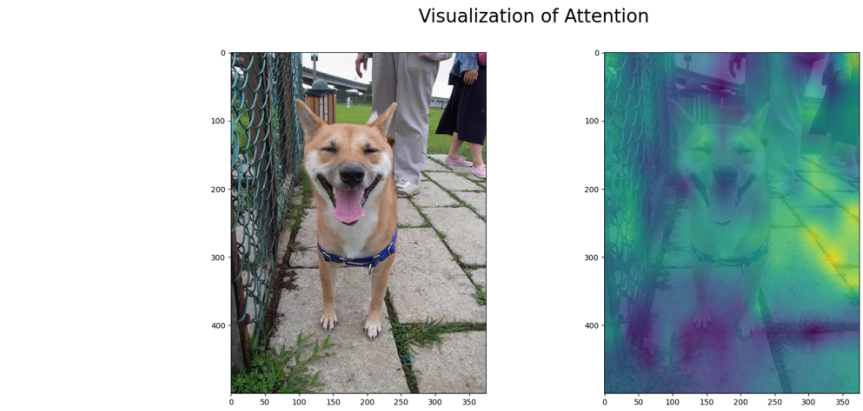


b. Discuss or analyze the visualization results (5%)
   i. 為了使模型能夠根據單詞在句子中的位置做不同判斷，因此在 patch embedding 加上 position embedding。
   ii. 將每個 embedding 和其餘所有 embedding 計算 cosine similarity 後可見模型學習到 patch 之間的位置資訊。
   iii. 每個 embedding 和其他同一列、同一行的 embeddings 有較高的相似度，由圖中也可見相似度最高的點由左上至右下移動，代表模型確實學習到這些 patch 實際在圖中的位置關係。

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%)

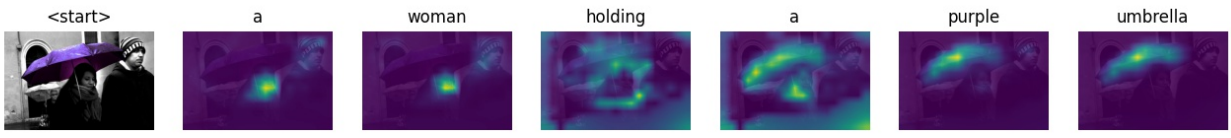
i. 由 class token 和所有 patches 計算 attention map，圖中黃色的部分代表 attention map 中權重較大處，也就是模型認為 class token 和哪些 patches 有較高的相關性。

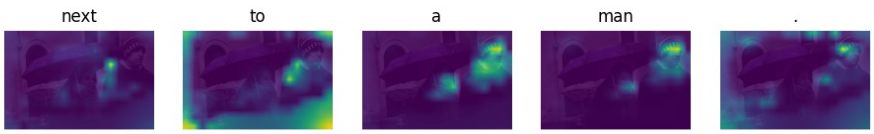
ii. 圖一可見在貓的後腳、尾巴、臉等部位有較高的權重，圖二可以看出在狗的耳朵、臉部有較高權重，都可以看出模型確實有針對圖中重要的部位做出判斷。

iii. 圖三在狗的眼睛部位稍微有較亮的點，但權重最高的部分在圖片的右側地板上，和預期中的結果差異較大，推測有可能是因為測資中有多不同品種的狗，但因此耳朵鼻子這種叫普遍的特徵不完全會是模型考慮的重點，可能也會參考顏色、背景等做出判斷。

Problem 2

1. 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?





i. 產生的 caption 和圖片內容一致，pretrained model 已有很好的表現，attended region 也確實有反映出每個字對應到圖片中的位置。
ii. Attention 圖中 woman、purple、umbrella、next、man 都可以明顯看出對應關係。

b. Discuss what you have learned or what difficulties you have encountered in this problem.
i. 從 attended region 可以看出預測結果受上下文影響，同樣都是”a”但在不同位置的”a”代表的意義不同，”a” woman 注意到 woman 部分，”a” purple umbrella 多注意到 umbrella 部位，“a” man 則注意到 man 的部位。
ii. 預測 caption 時每個字是依序產生的，故第 i 次產生的 attention 只有前 i 個 row 有意義。因此可以使用預測完完整 caption 後的最終 attention map，其第 i 個 row 即代表 caption 中第 i 個單字對應到圖中的位置。
iii. 在 ski.jpg 的 caption 中有一個複數單字”skis”，由於此處的 s 代表複數，在 Bert tokenizer 中後綴 s 也是一個字，因此在預測結果中 ski 與 s 為兩個不同的預測值，各有對應到的 attended region。