DLCV HW3 Report

B07902054 資工四 林子權

Problem 1

Report accuracy of your model on the validation set

我在一些可能會影響到accuracy的因素上做了測試，結果如下

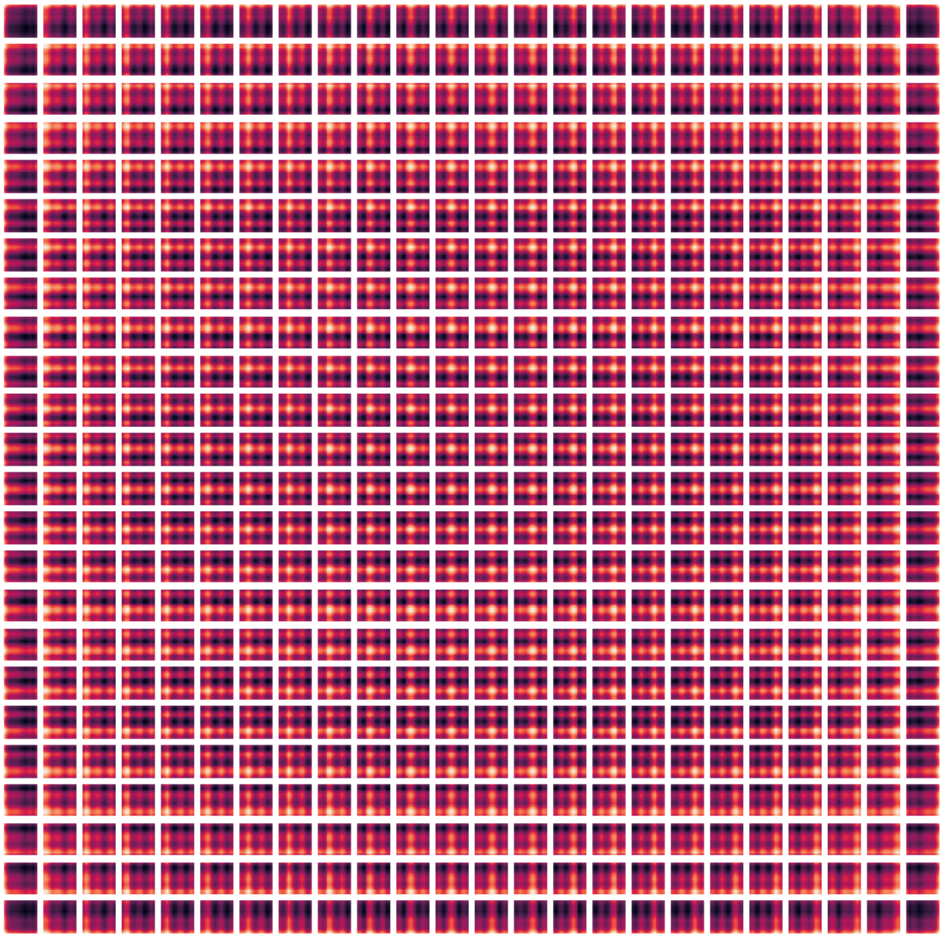
- Strong augmentation / Weak augmentat: Strong augmentation在一開始會收斂的比較慢，但會收斂在一個比較低的loss。在使用同樣的hyperparameters的情況下，最終的accuracy約為 94.8% / 93.9，使用strong augmentation的結果會比較好。
- Learning rate = 1e-5 / 1e-3 / 1e-1: 不同的learning rate會使得收斂的趨勢大不相同。測試過後發現當learning rate = 1e-5時可以收斂的比較穩定，loss呈現穩定的下降。
- Pretrained or not: 顯然地，使用pretrained model的結果絕對會好非常多。Train from scratch的情況下訓練了大約3個epoch，validation accuracy沒有突破4%。對比來說，用fine-tuned的方式訓練了第1個epoch結束，validation accuracy就已經79%左右了。

最終我使用了pretrained的ViT去做fine-tuning，learning rate = 1e-5，並且使用了相對強力的augmentation去防止overfitting在training data上。最終在validation set上得到的結果為約 **94.8%** 的正確率。

Visualize position embeddings

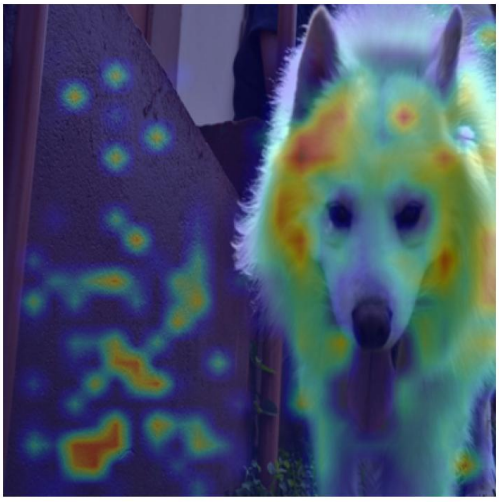
在下圖中，每一個小方形代表一個patch和其他patches之間的position embedding的correlation。亮度越高代表correlation越大。
觀察後可以發現，在四個角落的patch，和圖上所有的patches的correlation都很低，應該是因為image裡的四個角落能包含有用的信息本來就比較少。再來可以看到，每個patch的圖上亮度最高的地方，都和那個patch在圖上所在的位置分佈是基本一致的，這個現象是蠻好的，因為

position embedding的目的就是要embedd位置的資訊，跟相鄰的位置correlation高一些才是正常的。



Visualiza attention map

貓的那張圖可能attention的結果比較沒那麼好，model只focus在貓胸口的一搓毛髮上，其他則散落在背景。推測是因為這張圖的背景蠻花的，有可能會導致模型錯誤辨認成模種毛髮的特徵。其他兩張狗的圖，可以看到attention最高的區域幾乎都在狗的臉上，是蠻合理的結果。



Problem 2

Choose one test image and show its visualization result

從下圖可以看出來，每個word對應的attention都還蠻合理的。
答案是: A woman riding a bike with an umbrella on the street.

- A: 因為是woman的冠詞，所以focus在女人的身上
- woman: focus在女人身上
- riding: attention比較多集中在女人的上半身靠下半部，接近踩腳踏車的地方
- a: 因為是bike的冠詞，所以focus在腳踏車上
- bike: focus在腳踏車上
- with: 這個focus在女人還有街道上。在女人身上還可以理解，但在街道上感覺就比較難解釋。但其實可能也是with這個字本來就很難被visualize
- an: 因為是umbrella的冠詞，所以focus在雨傘上

- umbrella: focus在雨傘上