Deep Learning for Computer Vision Homework 3

R10522606 曾柏翔

Promble 1

1. Report accuracy of your model on the validation set.

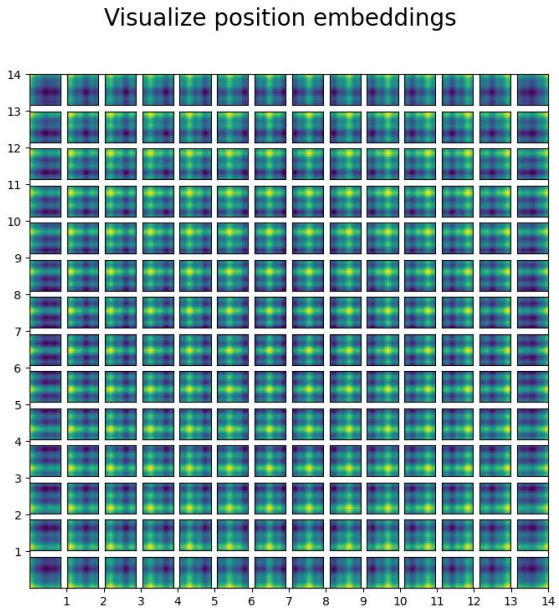
Accuracy : 0.942

在沒有 pretrain 的狀況下，由於訓練的資料量是不夠的，因此準確度很快就達到極限，因此後續的嘗試都使用了 pretrain。

首先使用了 ViT-Base，但精確度仍卡在 90%上下，因此調整成 ViT-large，雖然說有略微進步，但由於 Training 時間與記憶體消耗太龐大，所以認為效益不高。

因此最終仍調整回 ViT-Base 並將 learning rate 從 0.01 調整為 0.002，以及加上了 ColorJitter、RandomRotation、RandomHorizontalFlip 來增強訓練，最終成功通過了 Strong Baseline。

2. Visualize position embeddings

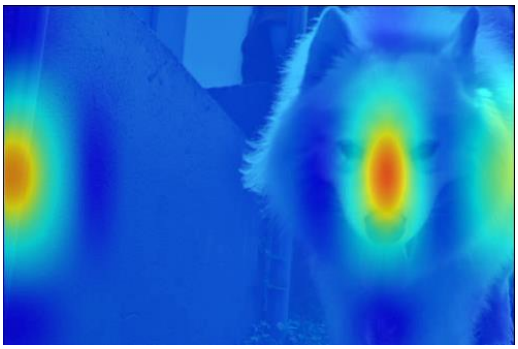


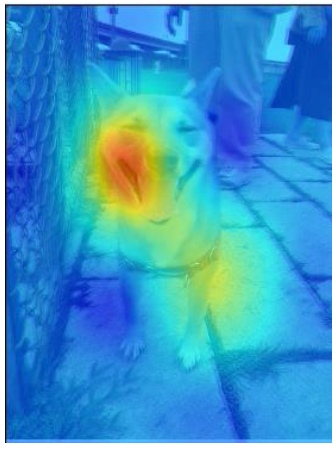
可注意到 ViT 的輸入是採用 patchs 而不是相片常用的 pixels，主要是為了讓模型更輕易的學習位置信息。由於 self-attention 是 permutation-invariant，因此學習位置信息是為了讓模型不需要花額外的成本去學習拼裝。

而從圖中可以注意到相近的 patchs 的 positional embedding 比較相近。

3. Visualize attention map of 3 images (26_5064.jpg, 29_4718.jpg, 31_4838.jpg)





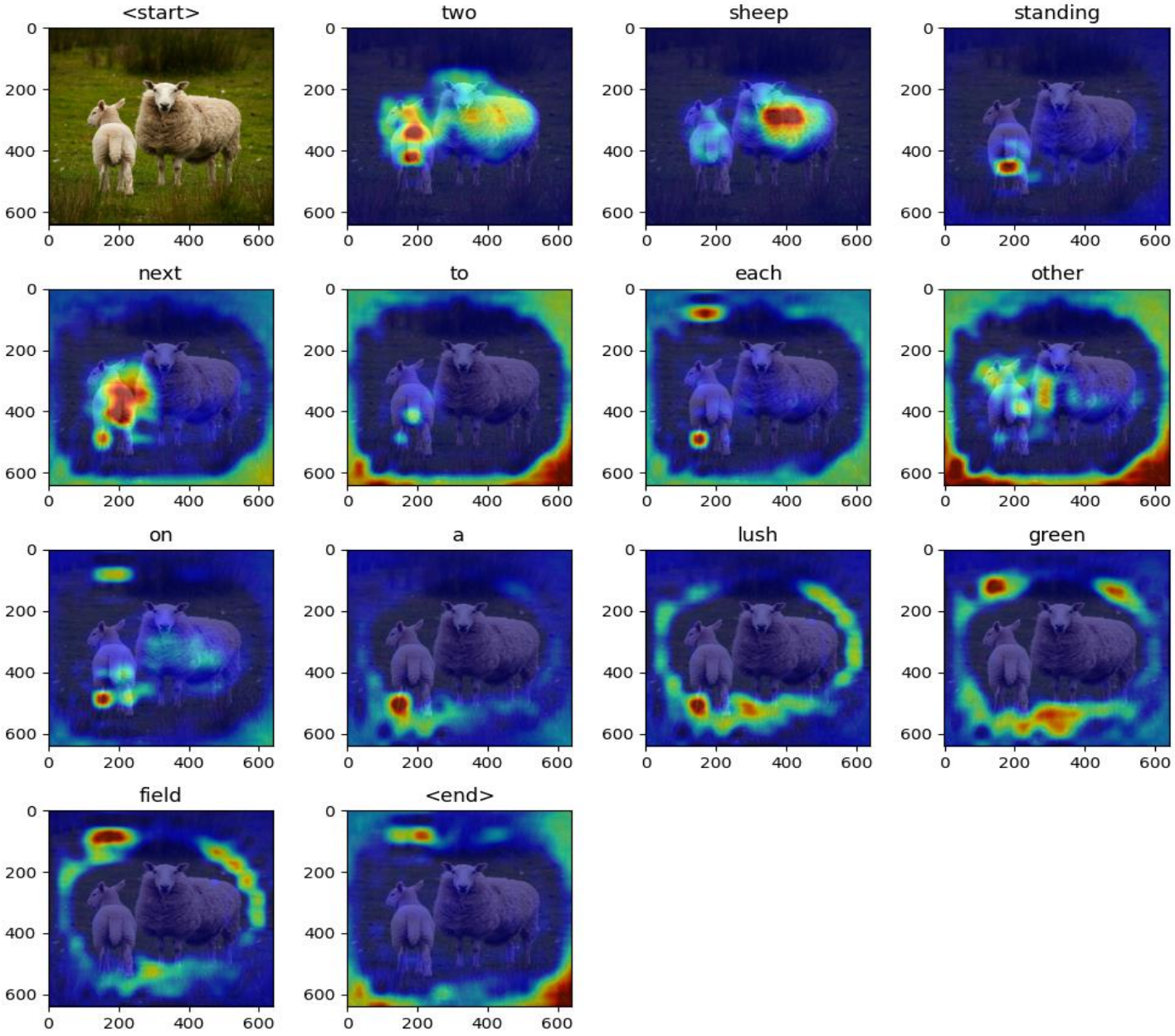
由上圖可以觀察到 model 在 classification 時，將注意力放在何處，先觀察第一張貓咪的圖片，可以看到在胸口與身體有較明顯的注意力，因此我回去看了看訓練資料，發現 class_26 的圖片是很多不同品種的貓，但都具有胸口是白色毛的特徵。

再來是薩摩耶，很直接的注意力放在了狗的頭上，但這邊可以注意到的是在左側也有一區，根據訓練資料推測，薩摩耶為了散熱，所以多半照片都是吐出舌頭的狀態，因此造成 model 誤把圖中咖啡色管子當成舌頭。

最後是柴犬，也是很直接的抓到了狗的臉部以及項圈的部分，這也可能是因為訓練資料的柴犬多半為有項圈的照片，以及柴犬脖子及胸口毛色為淺色，較易凸顯出項圈的特徵。

Promble 2

1. Choose one test image and show its visualization result in your report.



先來檢視所呈現的語句是否合理：two sheep standing next to each other on a lush green field.，可以說是非常的正確，其中使用到了 lush green 這個詞，其實令我蠻驚訝的。

再來看看個別圖片所對應到的單字狀況，前兩張清楚的抓到了羊的數量與位置，再來是 stand 也將注意力放在了腳的部分，而最後的幾張也清楚地抓到環繞在四周的草地。其中還有一個值得注意的是，針對介詞、代詞，圖片遮罩得部分通常都是出現在最外圍或是沒有明顯物體的位置。

在這題之中，首先遇到困難是弄清楚整個觀念，例如:模型如何抓取資訊、Multi-Head Attention 如何運作、如何轉換成文字等等。再來是看懂別人的 code 是如何撰寫，需要一層一層去理解，才能成功修改出所需要的樣子。

在做這題的時候，我有丟了很多我自己或是網路上的照片去試試看，講真的結果令我蠻意外的，應該是說模型判斷的精確度與用詞的精準性很驚人，雖然我故意丟長髮男生，模型卻給出了 woman。

最後是這一題讓我更了解到了 Image Captioning 的概念，先前就有稍微認識到 Mask R-CNN，但其實有點一知半解，但透過這題的引導，我有更完整的理解，並且更加意識到 Visualization 的重要性，也就如同前幾份作業的 T-sne，能夠讓我更直觀的明白其中運作的概念。

Reference

[1] ViT
https://huggingface.co/docs/transformers/model_doc/vit

[2] Visualization in Image Captioning
https://github.com/saahiluppal/catr

[3] Visualize Attention Map
https://reurl.cc/2odzD6

[4] Visualization in Image Captioning
https://github.com/saahiluppal/catr

[5] Self-Attention
https://blog.csdn.net/weixin_41811314/article/details/106804906