

Problem1-Image Classification with Vision Transformer

- [Problem 1-1] Report accuracy of your model on the validation set.

- a. Discuss and analyze the results with different settings.

Ans:在本次作業的 ViT 模型是使用 github PyTorch-Pretrained-ViT[1]，使用的版本為 pretrained model B_16，並修改 ViT 最後一層 fc 的輸出維度。在模型訓練參數中，optimizer 為 SGD，learning rate 設定 0.02，epoch 為 100。關於訓練模型實驗分析，在作 fine-tuning 時會有兩個方向：(1)固定住模型 feature extractor 的參數，只更新最後一層 fc 層參數；(2)更新整個模型參數。實驗結果發現使用第一個方法，Adam 和 SGD 表現差不多(但須根據不同 optimizer 調整 learning rate)；而第二個方法使用 SGD 可以更快達到較高正確率。

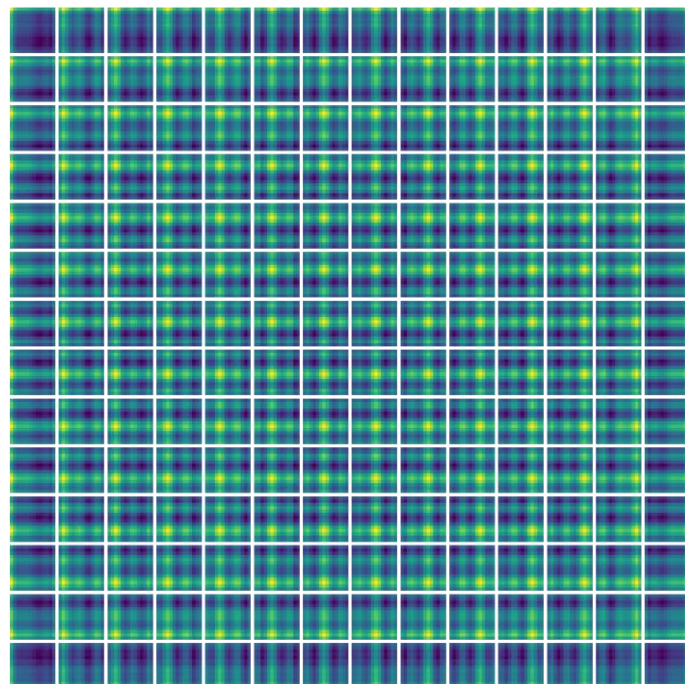
- b. Clearly mark out a single final result.

Ans:訓練好的模型在 validation set 上的 accuracy 表現為 **94.13%**。

- [Problem 1-2] Visualize position embeddings of your model.

- a. Visualize cosine similarities from all positional embeddings.

Ans:載入訓練好的 ViT 模型，取模型中 positional embedding 的部分，並對每一個 positional embedding 計算 cosine similarity，下圖為 cosine similarity 結果圖，越偏向黃色代表 cosine similarity 越高，其橫軸為 input patch column，縱軸為 input patch row。




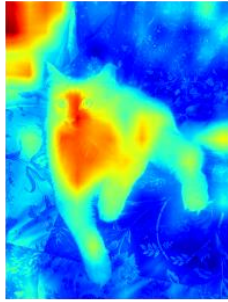

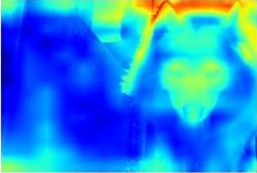

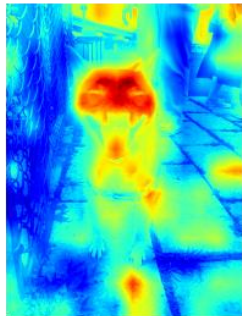
b. Discuss or analyze the visualization result.

Ans:根據 visualization 結果顯示，對於該 patch i ，與鄰近的 patch 會有較高的 cosine similarity。除此之外，該 patch i 也對相同 row 的 patch 和相同 col 的 patch 有較高的 cosine similarity，可以推估模型對於相同 row 或是相同 col 認為有一定的關聯。

■ [Problem 1-3] Visualize attention map of 3 images.

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.

Ans:以下為三張結果圖，每張結果圖左邊為原始圖，右邊為 attention map。

檔名	結果圖
p1_data/val/26_5064.jpg	<div>Image</div>  <div>Visualization results</div> 
p1_data/val/29_4718.jpg	<div>Image</div>  <div>Visualization results</div> 
p1_data/val/31_4838.jpg	<div>Image</div>  <div>Visualization results</div> 

b. Discuss or analyze the visualization results.

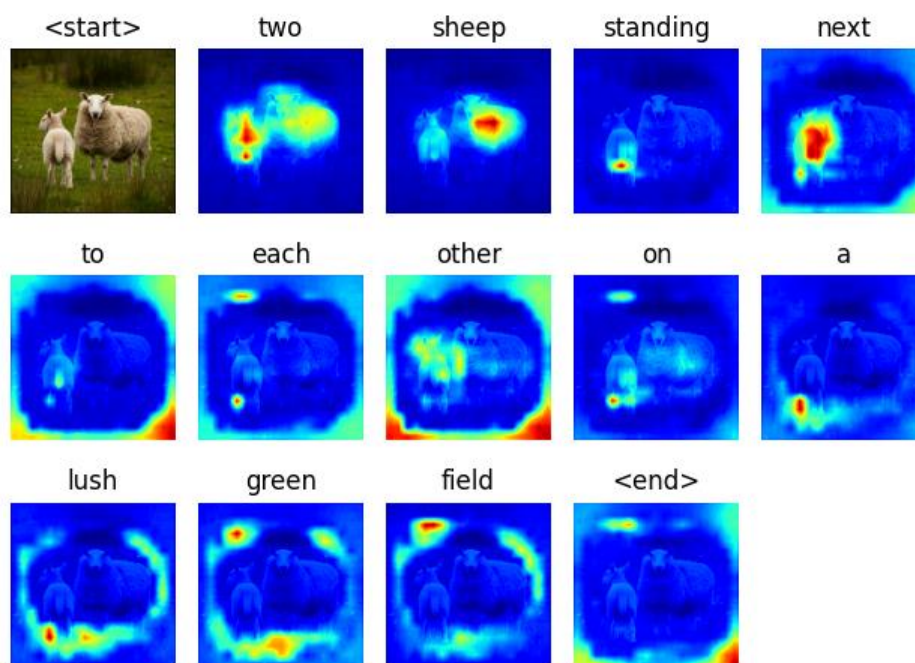
Ans:根據以上 3 張結果圖發現，模型有較高的關注在動物臉部和身體上，尤其在 26_5064.jpg 和 31_4838.jpg 較明顯。而在 29_4718.jpg 中較少關注在動物上，而是動物的頭部邊緣輪廓。另外，在 26_5064.jpg 除了動物之外，還關注了圖片左上角的區域。

Problem2-Visualization of Attention in Image Captioning

- [Problem 2-2] Choose one test image and show its visualization result.

- a. Analyze the predicted caption and the attention maps for each words.

Ans:使用 github Image Captioning with Transformers[2]的 v3 模型進行 image captioning 預測，由下圖可視化可以發現三件事情：(1)模型是可以了解語意抽象資訊，並反映在影像上，像是 two 在可視化上主要關注在兩隻羊上；(2)模型可以透過上一個字的線索，預測現在文字在圖像語意的位置，像是 two sheep、a lush green field；(3)有些抽象的文字如 to、each、other，模型在影像上的關注上是沒有意義的。



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

Ans: 在本次的 image captioning 的模型，inference 時模型是以影像和對應的字幕作為輸入，先分別作影像的 feature extractor 和 word embedding 後，再送入 transformer encoder 和 decoder 作 image captioning 預測。而獨特的地方在於 decoder 的 multi-head attention，作了 image patch high-level features 和 word high-level features 的 attention。另外，在 inference 時須根據模型上一個 output 作輸入來預測下一個字。

Reference

- [1] PyTorch-Pretrained-ViT: <https://github.com/lukemelas/PyTorch-Pretrained-ViT>
- [2] Image Captioning with Transformers : <https://github.com/saahiluppal/catr>