

# Deep Learning for Computer Vision - Homework 3

資工碩一 R10922005 李澤謙

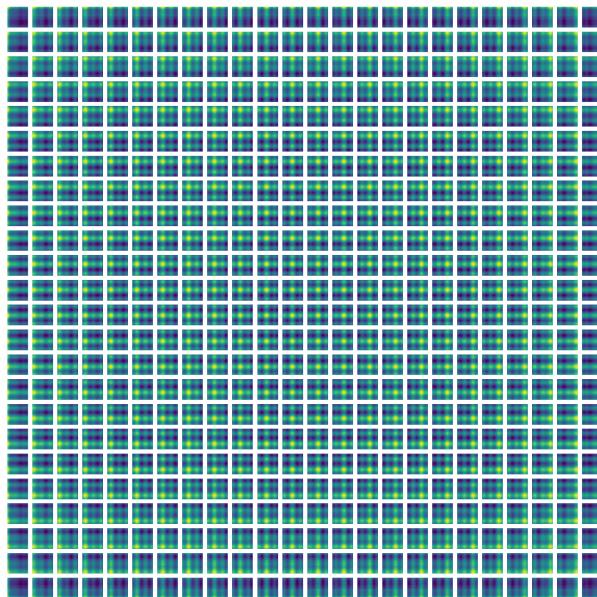
December 14, 2021

## Problem 1: Image Classification with Vision Transformer

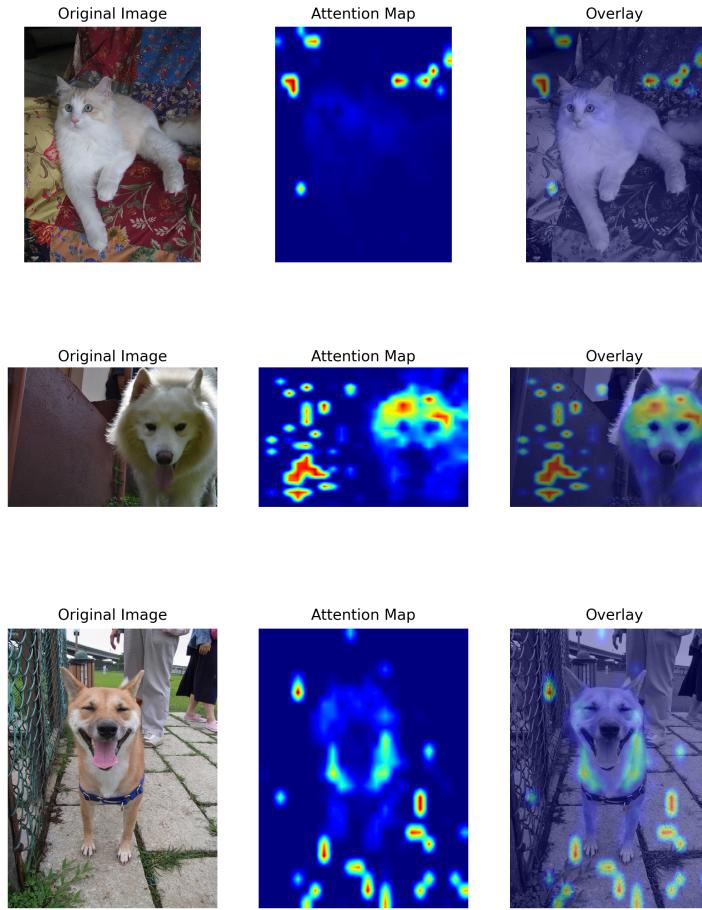
我於本次作業中使用了套件 pytorch\_pretrained\_vit 所提供的 pretrained model (B\_16\_imagenet1k)，並將其 output layer 中的 neuron 數量改為 37 個，以此作為本次作業所使用的 model。而在 data preprocessing 方面，我將 training dataset 中所有的圖片皆縮放至  $384 \times 384$  的大小，並將其 pixel 的值 normalize 到 -1 和 1 之間，以此進行 data preprocessing，此外，我也使用了 torchvision 的 RandomHorizontalFlip、RandomAffine、ColorJitter 進行 data augmentation。最後，我使用了 SGD 作為 optimizer，並使用 cross entropy 作為 loss function，以此訓練 ViT，其中，batch size 為 4，learning rate 為 0.0005，finetune 了 30 個 epoch，以此得到最後的 model。以下為我訓練出來的 ViT 所得到的 accuracy：

Train	Validation
0.97609	0.94400

下圖為我訓練出來的 ViT 其 positional embedding 之間的 cosine similarity：



由上圖可以看出，彼此相鄰或者是位在同一個 row 或 column 的 patch 之間的 positional embedding 會越為相似，由此可以說明 ViT 確實有學到 patch 的位置資訊。而下圖為我將 p1\_data/val/26\_5064.jpg、p1\_data/val/29\_4718.jpg、p1\_data/val/31\_4838.jpg 輸入 ViT 之後，將最後一層 multi-head attention layer 中以 CLS token 作為 query vector 而其它 patch 作為 key vector 所得到的 average attention map：

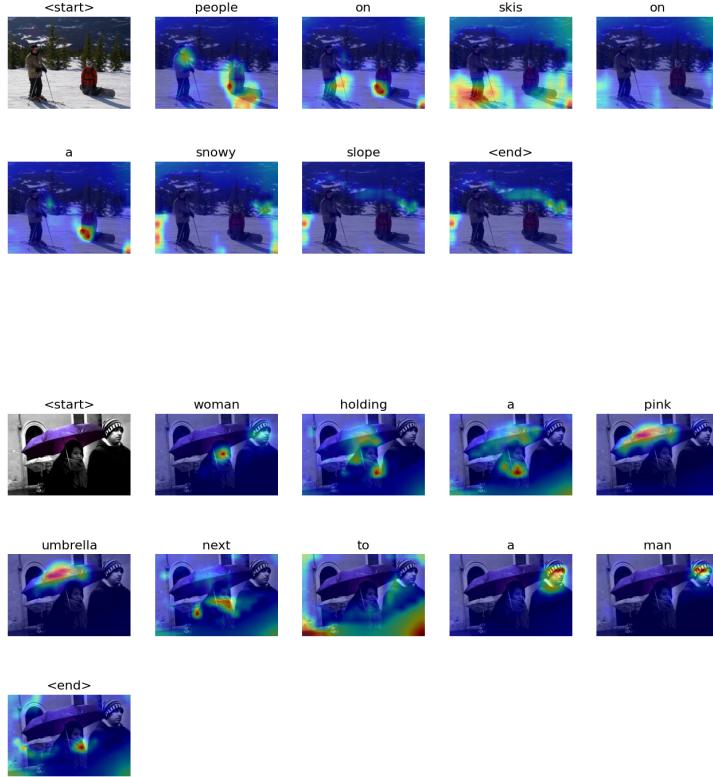


由上圖可以看出，圖片中動物部分的 attention weight 有比圖片中的其餘部分稍微高出一點，由此可以說明 ViT 確實是以圖片中動物的部分來進行 classification，但由上圖也可以看出，圖片中的背景部分經常會有小區域出現極高的 attention weight，推斷其可能為 noise 所造成，其可能會使得 ViT 判斷失誤，也許將輸入圖片進行 denoise 之後可以再進一步地提高 ViT 的 accuracy。

## Problem 2: Visualization in Image Captioning

下圖為我將 p2\_data 中的 5 張圖片輸入CATR後所得到的 caption 和 attention map：





由上圖可以看出，雖然 CATR 有時會辨識錯誤（例如 girl.jpg 中 CATR 將 girl 辨識成了 boy），但不論其辨識正確與否，其在輸出名詞的時候確實會專注於圖片中相對應的物體部分，在輸出形容詞時則會專注於其所形容的物體上，在輸出動詞時則會專注於做出該動作的身體部位，輸出其它詞時則不一定會專注於圖片的任一部位，推測此時可能是根據之前輸出過的詞而不是圖片來決定接下來要輸出的介系詞或冠詞等等。經由本次作業，可以看出 transformer 確實可以在 computer vision 方面的任務得到極佳的成果，而且其 model 的可解釋度極高，convolution 長時間以來稱霸 computer vision 相關的任務，但或許 self-attention 今後不只會在 NLP 領域，其也將一統 computer vision 的領域，在各個任務帶來更大的進步，或許這也意味著 self-attention 機制確實更為接近人腦在學習時的實際機制，其今後的技術發展令人期待。