HW3 Report

Problem 1:

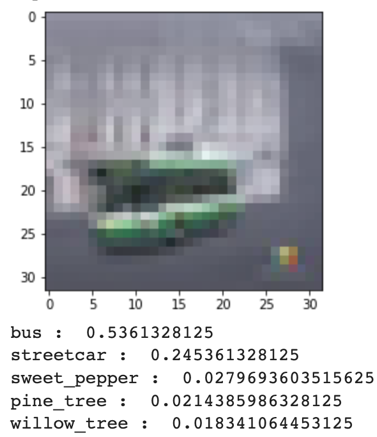
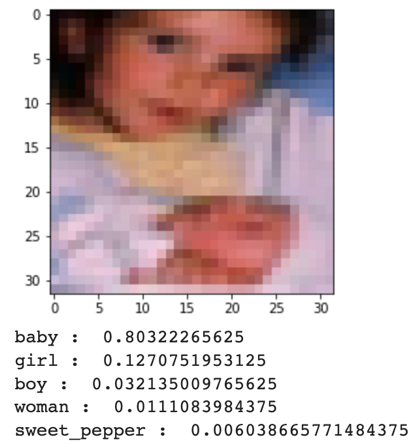
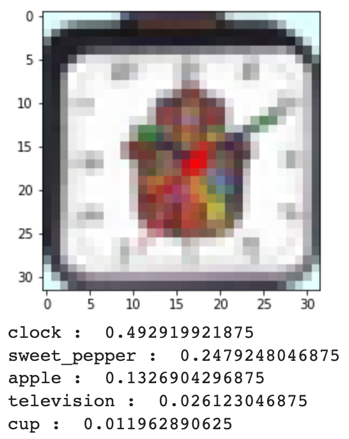
1. CLIP 是以文字當作中介去辨識圖片。對於傳統的VGG、ResNet而言，沒看過的類別、圖片是真的沒有看過，因此在Zero Shot做不好。但是CLIP會試著把圖片或類別轉換成文字，而文字他就看得懂了，因為pretrained model有看過這些文字。所以兩者最大的差別就是，前者真的是在看陌生的東西，後者則是把陌生的東西轉換成熟悉的東西再拿來看。
2. This is a photo of {object} : 0.6084

This is a {object} image. : 0.6844

No {object}, no score. :0.5628

首先看到第三個prompt text，他的語句邏輯是比上面兩句複雜一點的，用到了雙重否定來表達肯定，所以可能對於model來說比較難以理解，而且model也不一定知道score要高才是好，所以即使no score可能對他來說也沒差。

再來是前兩個比較，第一個沒有句號，第二個有，但是我實驗了幾句prompt text之後，發現句號的影響很不穩定，有時變準有時變不準。所以我猜測影響關鍵可能是” a ”。像是在problem 2中，我自己生成的句子幾乎都是以” a “開頭，然後只要是名詞前面都常常掛一個” a “，所以” a “對於這種語言模型來說可能很重要。我實驗用This is a photo of a {object}當作prompt text，結果準確率就來到0.6876了。

1. 三張圖片：

Problem 2

1. Best setting:

用timm 的swin\_large\_patch4\_window7\_224當作backbone進行image encoding（有使用pretrained weight），後面再接上8-layer 8-head decoder做cross attention還有輸出。其中attention的feed forward layer中間維度設為2048維，並且attention有0.1 dropout rate。另外在back prop結束後，會把gradient做clipping 至(-1, 1)。使用torchvision.transforms的ColorJitter(brightness = 0.4, saturation = 0.4, contrast = 0.4, hue = 0)做data augmentation，並將圖片以ImageNet的標準標準化。Training時，依序使用1e-5、3e-6、1e-6三種learning rate ，每當cider和clip score停止上升，就換到小一點的數值繼續訓練。生成captions的時候，使用greedy的方式生成各字。

CIDEr: 0.8745 CLIP: 0.7182

1. 以下三種model若有setting沒有特別說明，均是按照best setting 去做。
2. Backbone使用swin\_base\_patch4\_window7\_224，feed forward layer維度1024，沒有使用data augmentation，圖片標準化使用mean = 0.5 std = 0.5。

CIDEr: 0.8554 CLIP: 0.7183

1. Backbone使用swinv2\_base\_ window8\_256，feed forward layer維度1024，learning rate = 1e-5。

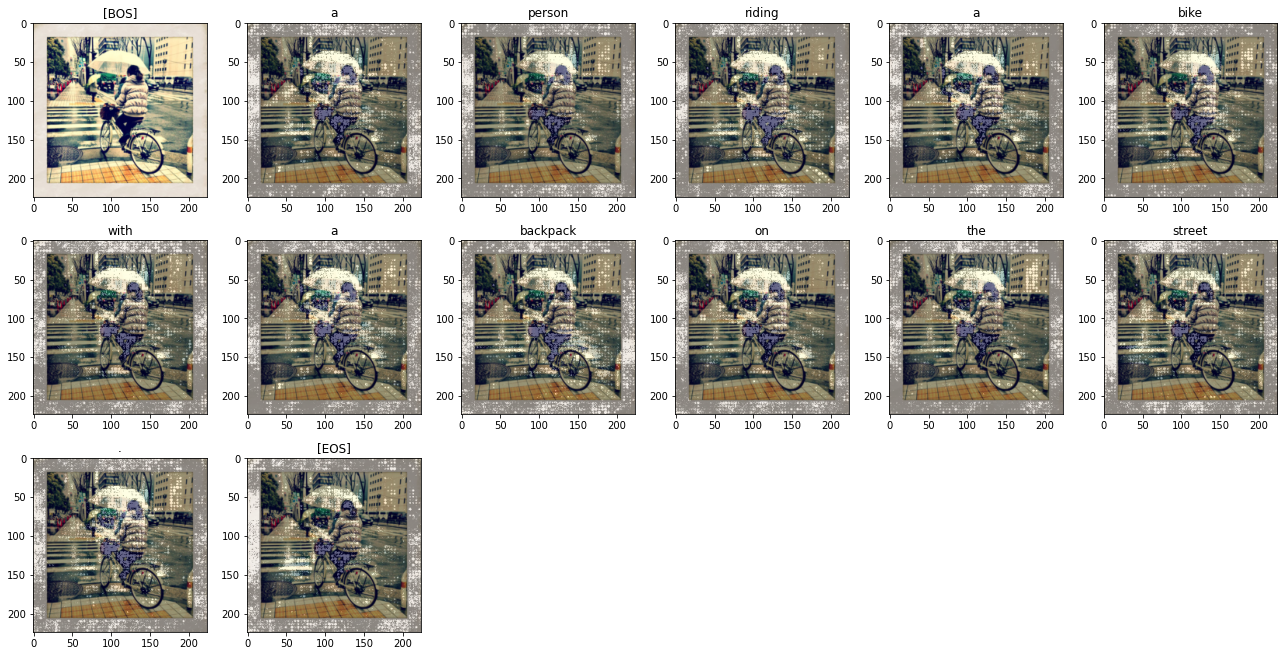
CIDEr: 0.7840 CLIP: 0.6991

1. Backbone使用vit\_base\_patch8 \_224，feed forward layer維度1024，沒有使用data augmentation，圖片標準化使用mean = 0.5 std = 0.5。

CIDEr: 0.8006 CLIP: 0.7063

Problem 3

圖片區域越白表示attention越強。

1. bike.jpg

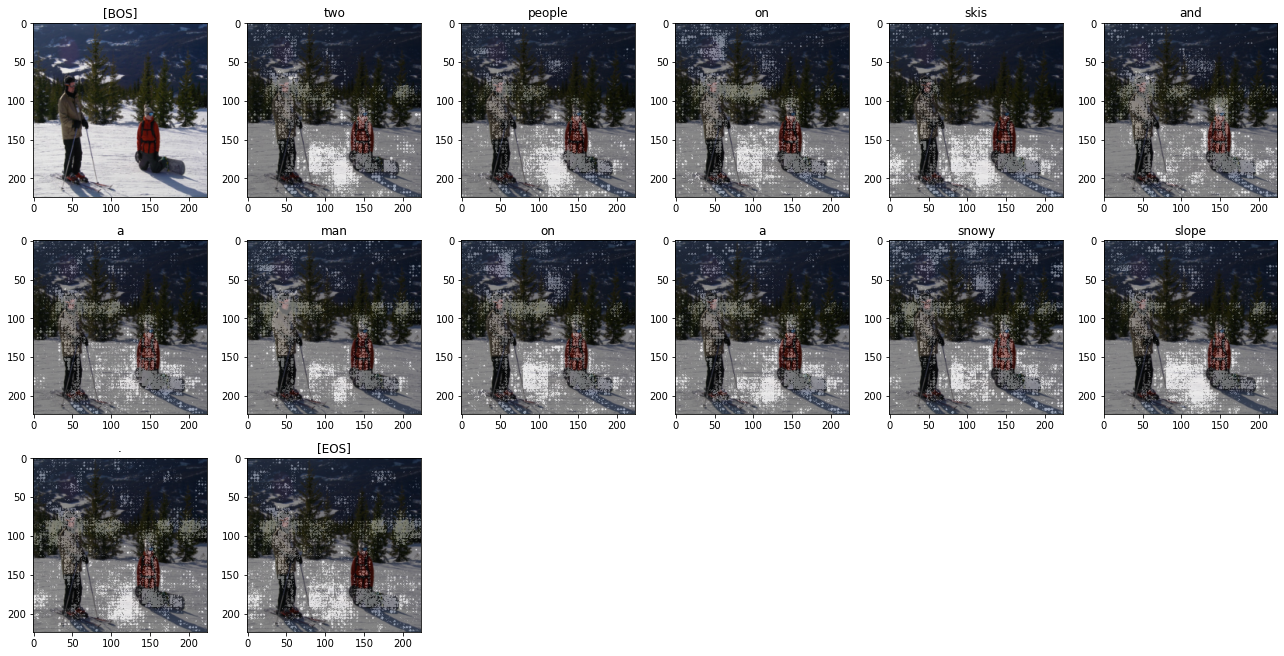
girl.jpg



sheep.jpg



ski.jpg



umbrella.jpg



1. 000000392315.jpg, CLIP score = 1.0



000000141426.jpg, CLIP score = 0.4294



1. CLIP較高的圖產生的文字相當合理，較低的圖是把圖中的生物錯認成牛（我猜是羊），而且完全沒有提到柵欄、人群等字詞。

在CLIP較高的那張圖上面，可以看到只要提到fire hydrant的特質，如顏色、位置等等的時候，fire hydrant上面就會比較白、road的部分會稍微少一點。連接詞的部分大多都是比較意義不明，可能是因為是兩個物體之間的互動，看得要比較全局。

在CLIP較低的圖中，model不斷聚焦在生物身上，也跟他產生的字相符（沒有提及環境的事物），而且可以看到a、group、cows等字他所聚焦的重點都是兩隻生物，而其他字詞僅有一隻生物，顯示即使生成的句子不盡理想，model所關注的焦點也會與生成的字約略一致。