

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

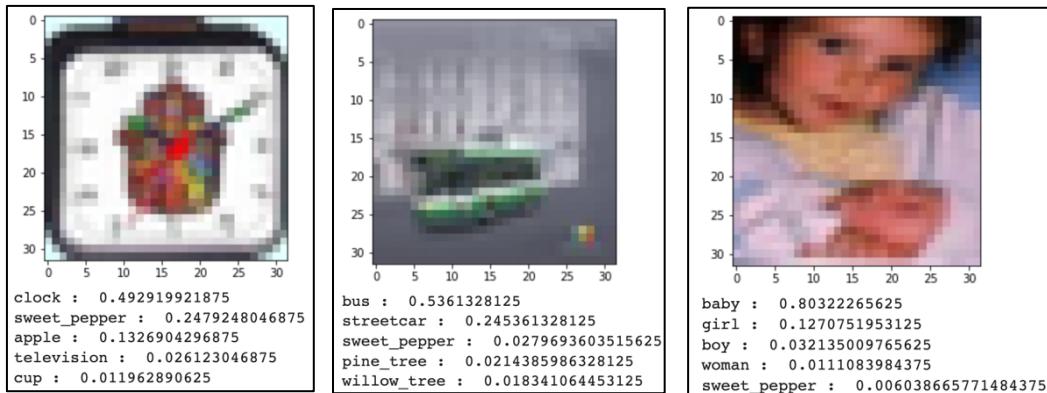
This is a {object} image. : 0.6844

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。

3. 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 級，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。

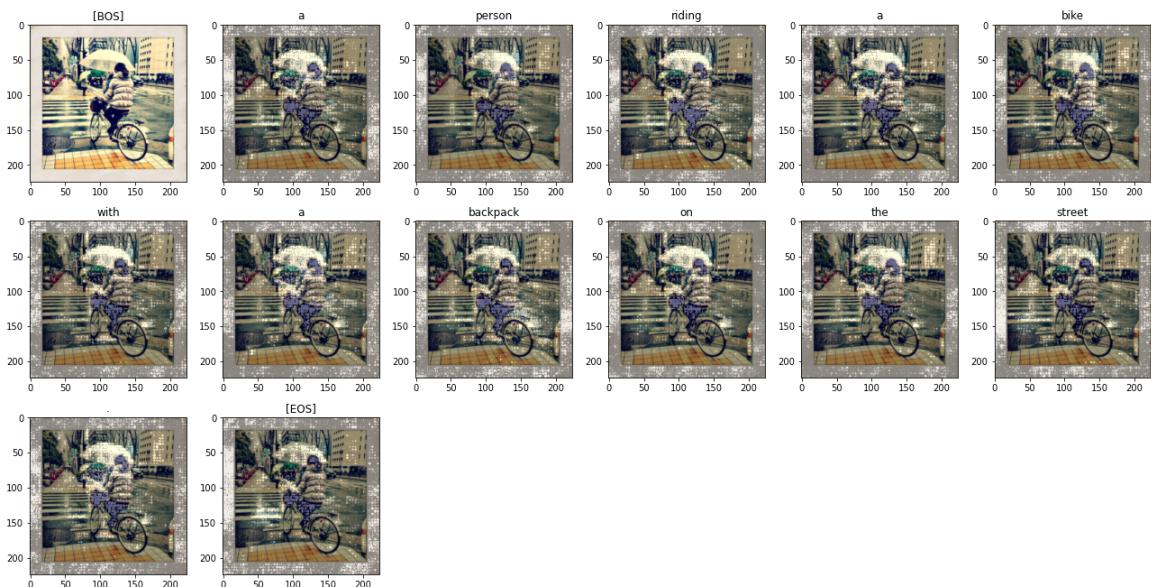
CIDEr: 0.8745 CLIP: 0.7182

2. 以下三種 model 若有 setting 沒有特別說明，均是按照 best setting 去做。
 - I. 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
 - II. Backbone 使用 swinv2_base_window8_256，feed forward layer 維度 1024，learning rate = 1e-5。
CIDEr: 0.7840 CLIP: 0.6991
 - III. 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。

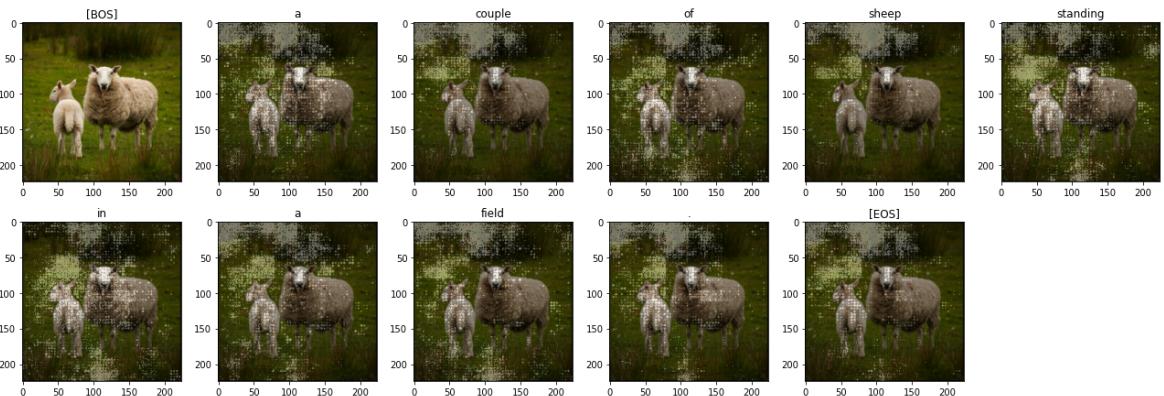
1. bike.jpg



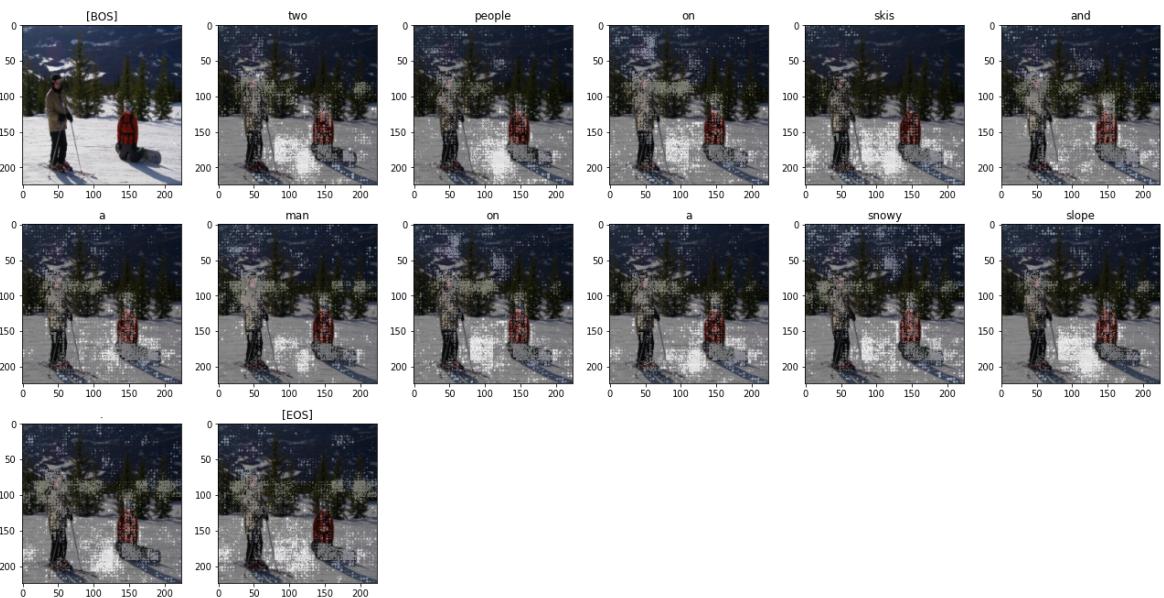
girl.jpg



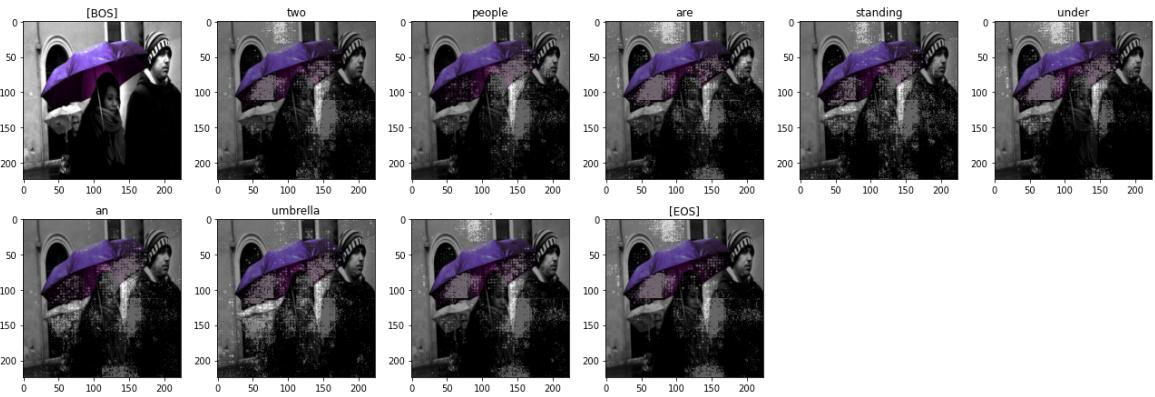
sheep.jpg



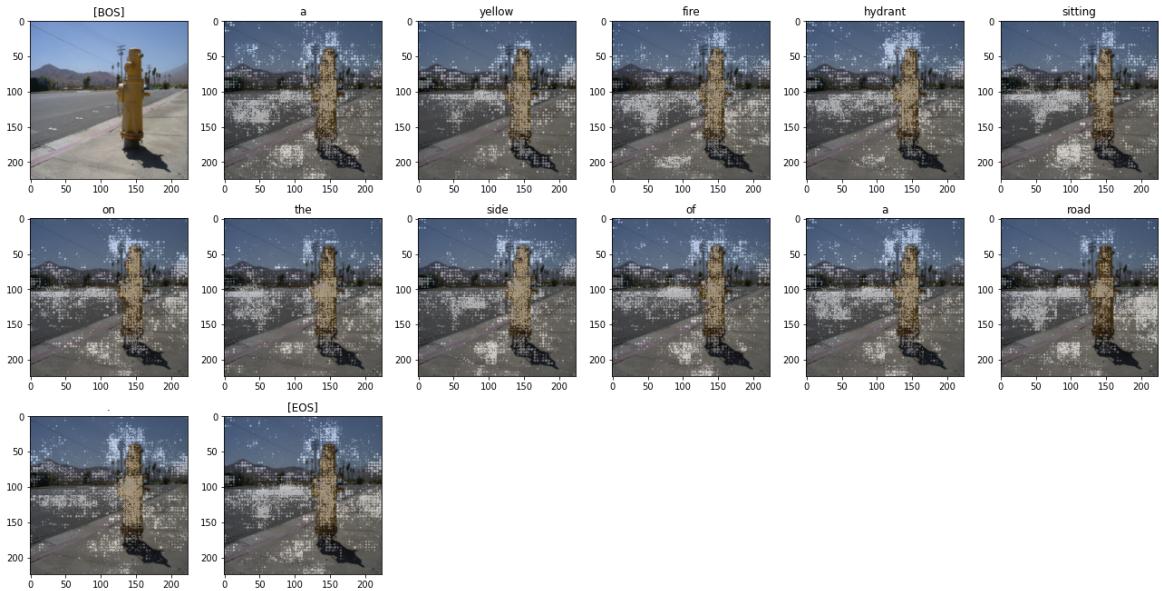
ski.jpg



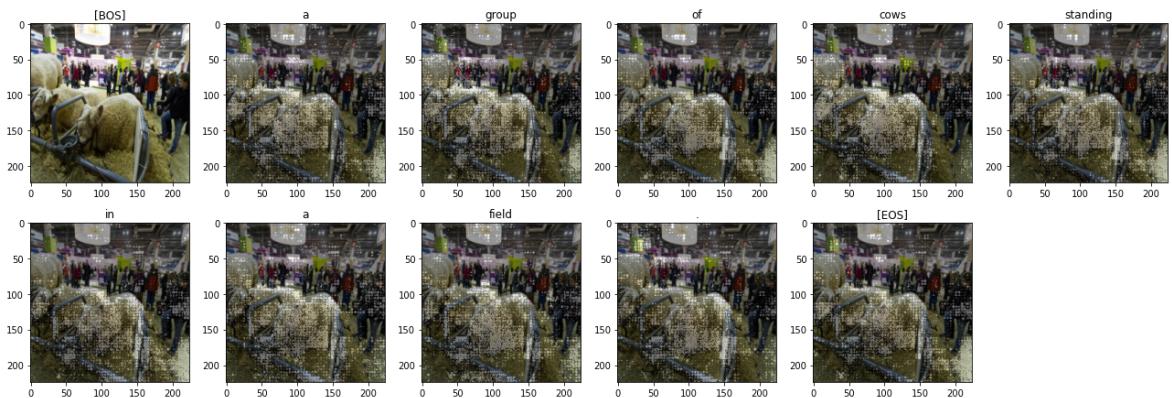
umbrella.jpg



2. 000000392315.jpg, CLIP score = 1.0



000000141426.jpg, CLIP score = 0.4294



3. 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。