DLCV HW3 Report

Student ID: R10943117
Name: 陳昱仁

Problem 1.

1. **Previous methods (e.g. VGG and ResNet) are good at one task and one task only, and requires significant efforts to adapt to a new task. Please explain why CLIP could achieve competitive zero-shot performance on a great variety of image classification datasets.**

在先前的方法(e.g. VGG and ResNet)，在原 dataset 訓練好後，拿去分辨有同樣 class 但是不同的 dataset，效果就不太好，因為資料的 distribution 在每個 dataset 都不盡相同，因此換一個 dataset 測試成效就不好，然而 clip 的訓練方式解決了這個問題。

OpenAI 蒐集網路上大量圖片以及這些描述圖片的文字來訓練 clip，此訓練不須任何標記 label，因此可訓練大量 data，所以不會像先前方法只能適用在特定 dataset。text 跟 image 會映射到同一個空間，最後得到一個文字跟圖片關聯的模型。這個模型非常強健，有很高的通用化程度，因為沒有針對特定領域特化，所以反而更通用，對於很多領域的 zero-shot test 都有不錯的表現。

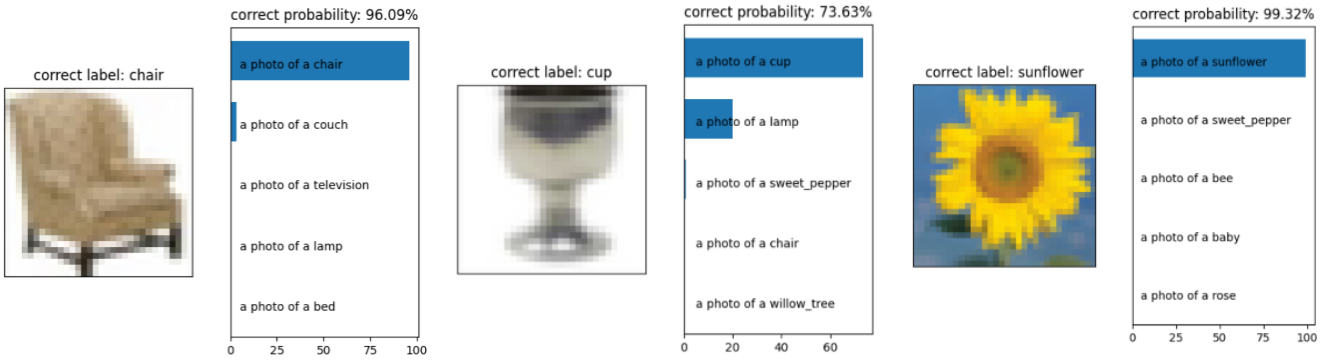
2. **Please compare and discuss the performances of your model with the following three prompt templates.**

Text	Accuracy
This is a photo of {object}	60.84%
This is a {object} image.	68.36%
No {object}, no score.	56.20%

從實驗可以看出，只要 text 中有出現該類類別的名稱，就會有至少 50%的準確度，因 text 中已經有提及該類別，自然有一定的分辨能力，若整個 text 較符合圖片的描述，準確度會越高，所以可以看到前兩種 text 準確度較高，有沒有句點也會影響準度。

前兩個 text 意思相同，但準確度有落差，實驗結果是因為第一種 text 較不符合英文文法，若改成 *This is a photo of a {object}* ，準確度就會和第二個 text 差不多。

3. **Please sample three images from the validation dataset and then visualize the probability of the top-5 similarity scores as following example.**



Problem 2.

1. **Report your best setting and its corresponding CIDEr & CLIPScore on the validation data.**

   Encoder 使用 pretrain 過的 vit clip large，並 freeze encoder，image size 為(224，224)，會經由(mean = 0.5，std = 0.5) normalize，decoder 使用 transformer 的 decoder，nhead = 8，num_layers = 4，並使用 data augmentation，decoding strtegy 為 greedy。CIDEr 和 CLIPScore 分別為 0.955 和 0.739。

2. **Report other 3 different attempts (e.g. pretrain or not, model architecture, freezing layers, decoding strategy, etc.) and their corresponding CIDEr & CLIPScore.**

   A. freeze encoder
   B. encoder 由 vit-base 改成 vit-clip-large
   C. data augmentation，如下圖

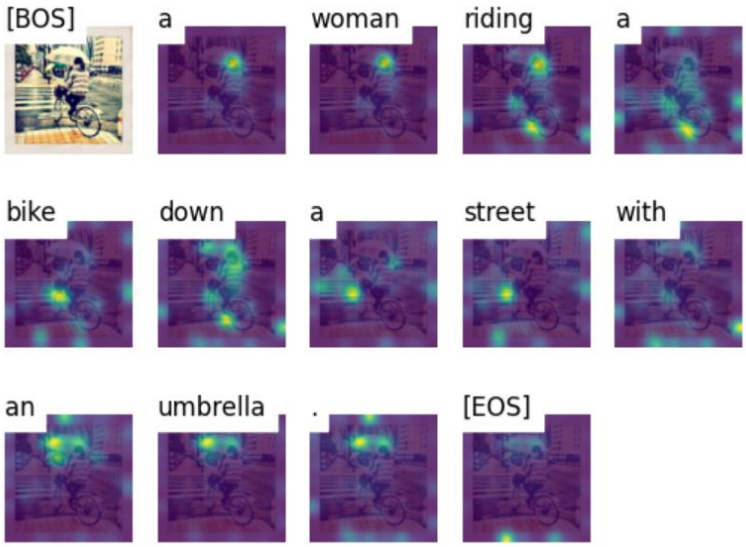
   baseline 為不使用 ABC 任何 attmpt，encoder 為 vit-base。

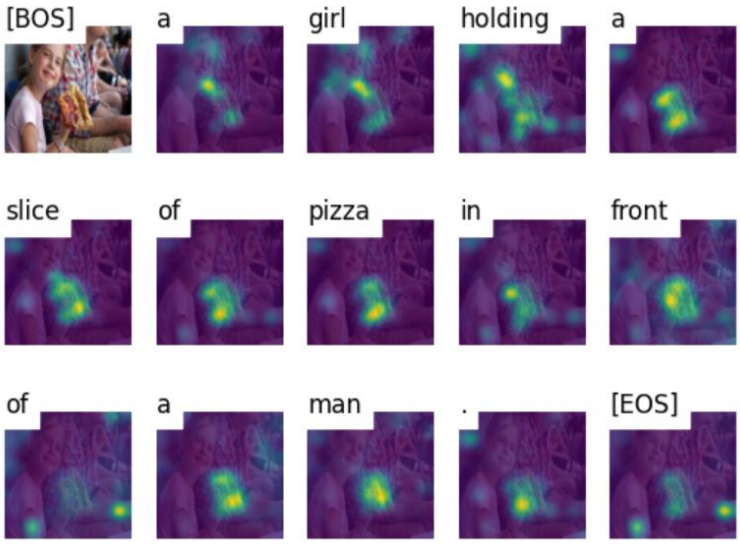
	CIDEr	CLIPScore
baseline	0.696	0.706
baseline+A	0.755	0.716
baseline+B	0.825	0.724
baseline+C	0.722	0.709
baseline+ABC	0.955	0.739

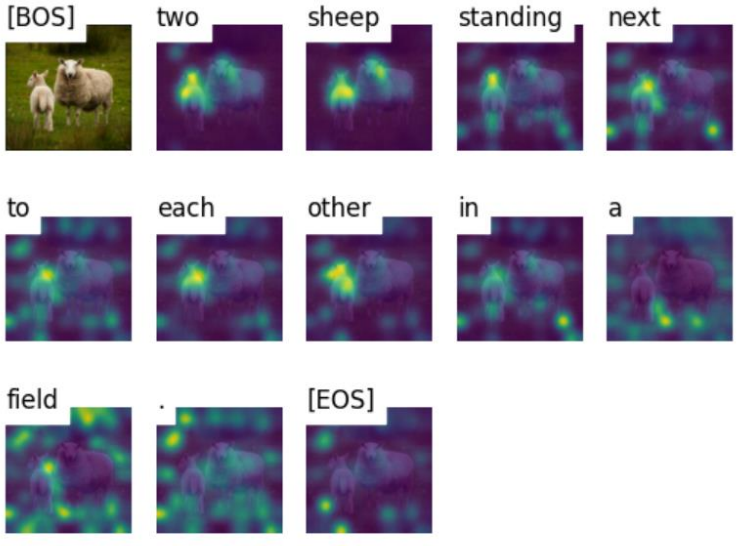
```
transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ColorJitter(brightness =[0.5, 1.3],
                           contrast =  [0.8, 1.5],
                           saturation =[0.2, 1.5]),
    transforms.RandomHorizontalFlip(),

    transforms.ToTensor(),
    transforms.Normalize(0.5, 0.5),
])
```

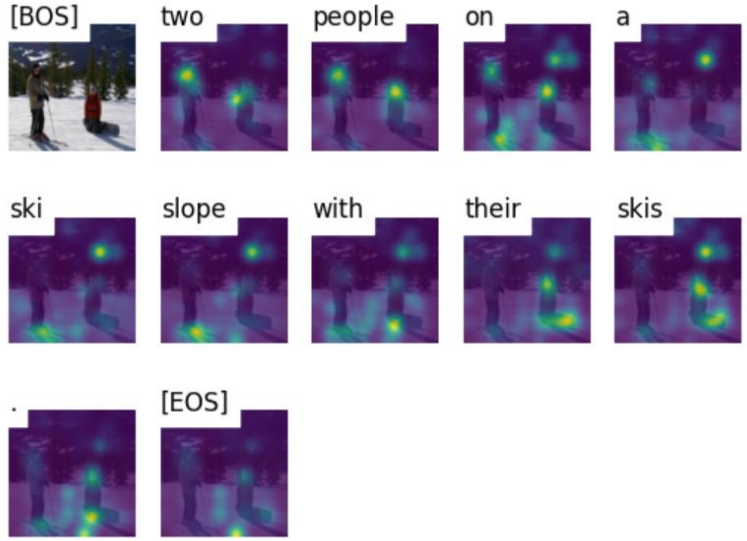
Problem 3.

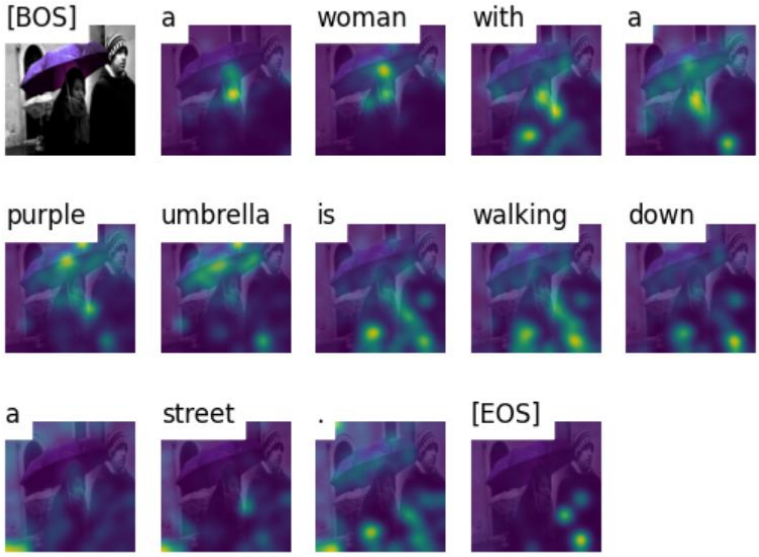
1. Please visualize the predicted caption and the corresponding series of attention maps.








[BOS]
two
people
on
a
ski
slope
with
their
skis
.
[EOS]

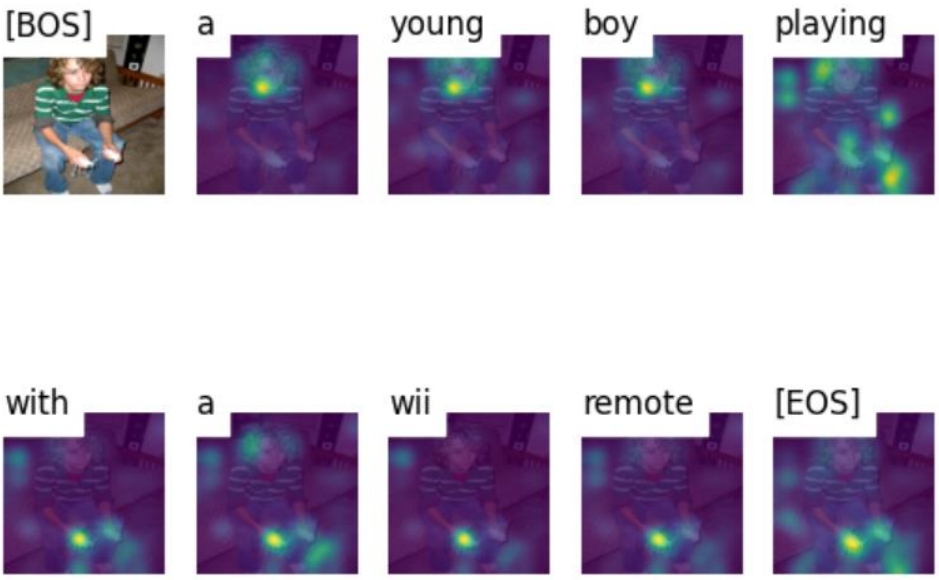

[BOS]
a
woman
with
a
purple
umbrella
is
walking
down
a
street
.
[EOS]

2. **According to CLIPScore, you need to visualize:**

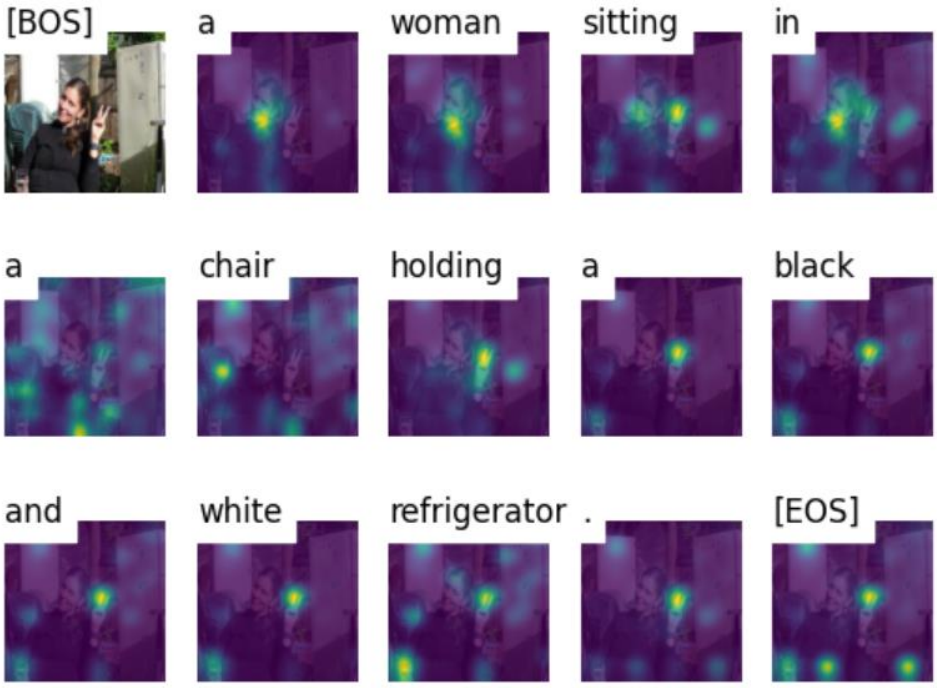
   i. **top-1 and last-1 image-caption pairs**

   ii. **its corresponding CLIPScore**

   in the validation dataset of problem 2.



top-1 pairs, CLIPScore = 1.00



last-1 pairs, CLIPScore = 0.394

3. **Analyze the predicted captions and the attention maps for each word according to the previous question. Is the caption reasonable? Does the attended region reflect the corresponding word in the caption?**

由上述視覺化的圖可看出，最好的結果的 caption 很符合圖片且合理，並且其 attention map 也相當準確，前段都落在小男孩頭上，後段落在小男孩手中的遙控手把上。最差的結果的 caption，並沒有很符合圖片，女人並沒有坐在椅子上，手上也沒有拿 registration，attention map 前段準確度還可以，但後段和文字較不符合。兩張圖冠詞和介係詞的位置，attention map 會和前後圖片差不多位置。這兩張圖片的 CLIPScore 和視覺化結果一致。