Hw3_report

陳敬和 R11922166

Problem 1 Zero-shot image Classification with CLIP

1. Methods analysis: Explain why CLIP could achieve competitive zero-shot performance on a great variety of image classification datasets.

CLIP透過Natural Language Supervision，用來學習image的feature。有別於之前的dataset，CLIP使用「圖片+文字描述」的資料集，這種文字描述的模糊性可以讓 AI 模型訓練的時候有更好的描述性。獲取的方法是利用大量的query，拿去網路上搜尋取得相對應的圖片，有了大量且廣泛的訓練資料，在做zero-shot classification時也就能有不錯的表現。

2. Prompt-text analysis: Please compare and discuss the performances of your model with three prompts

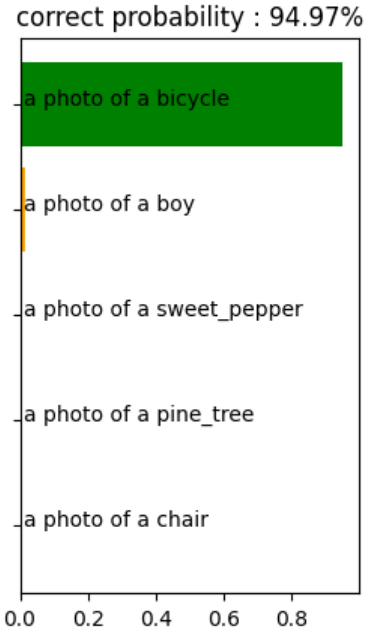
	Accuracy
“A image of a {object}”	71.64%
“A photo of {object}”	71.08%
i. *“This is a photo of {object}”*	60.92%
ii. *“This is a {object} image.”*	68.28%
iii. *“No {object}, no score.”*	56.32%
“This is a image of {object}”	64.76%

除了report要求的i, ii, iii三個prompt-text，另外實驗了”photo”和”image”對於accuracy的變化。從上面實驗結果可以知道，以”A”為開頭的prompt-text的準確率是比其他高的；而相對於”photo”，prompt-text中含有”image”的accuracy會比較高。至於iii，推論是因為不是一句直述句，而是有兩句話，且是語意沒什麼幫助的文本，因此準確率並不是很高。

3. Quantitative analysis: Please sample three images and then visualize the probability of the top-5 similarity scores

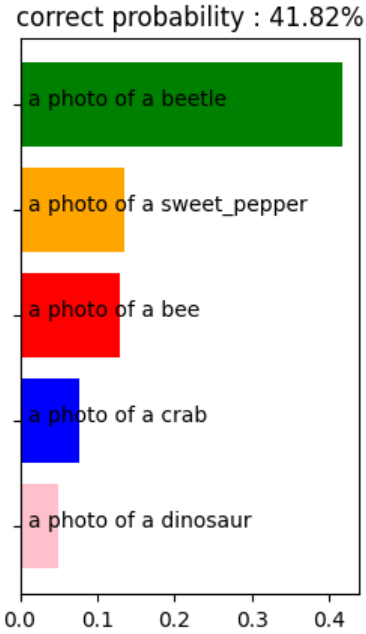
- Bicycle


correct label : bicycle

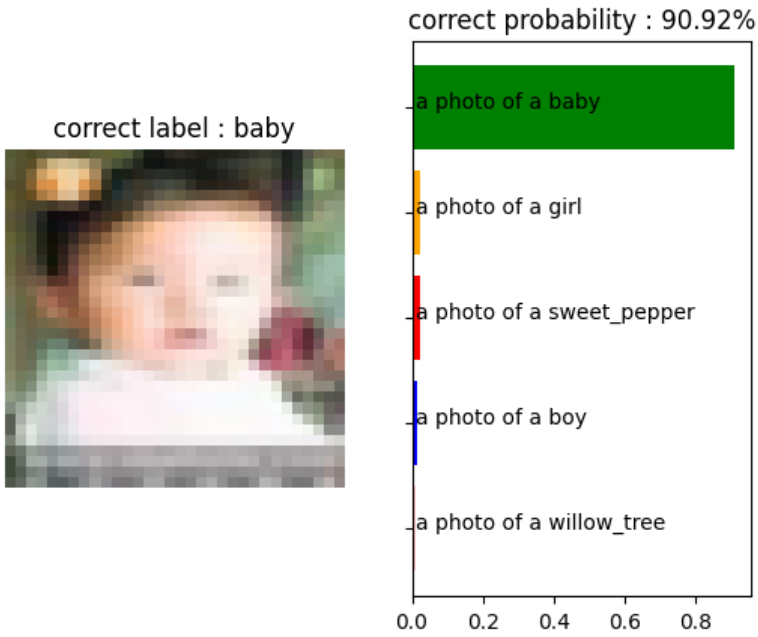

correct probability : 94.97%
a photo of a bicycle
a photo of a boy
a photo of a sweet_pepper
a photo of a pine_tree
a photo of a chair
0.0
0.2
0.4
0.6
0.8

- Bee


correct label : bee


correct probability : 41.82%
a photo of a beetle
a photo of a sweet_pepper
a photo of a bee
a photo of a crab
a photo of a dinosaur
0.0
0.1
0.2
0.3
0.4

- Baby



Problem 2 Image Captioning with VL-model

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

- Settings

Encoder	# Decoder layer	learning rate	optimizer	scheduler
CLIP-Vit-L/14@336px	5	1e-4	Adam	MultiStepLR

- Score

```
CIDEr: 0.9106683255232717 | CLIPScore: 0.7197610888153297
```

2. Report other 3 different attempts and their corresponding CIDEr & CLIPScore

1. Change learning rate

Encoder	# Decoder layer	learning rate	optimizer	scheduler
CLIP-Vit-L/14@336px	5	**3e-5**	Adam	MultiStepLR

```
CIDEr: 0.6663138535774273 | CLIPScore: 0.6692470765900119
```

可以發現在learning調高，其他配置不變時，在預測出來的json裡面可以看到一些caption會重複或是輸出沒有用的caption，如：

"a man in a black shirt and white horse and a man in a black shirt is riding a white photo ."

"a woman in a red dress and white dress is standing in front of a woman in a group of a group of a group of a group of a group of people ."

因此CIDEr與CLIPScore計算出來的都不高。

2. Change decoder layer數量

Encoder	# Decoder layer	learning rate	optimizer	scheduler
CLIP-Vit-L/14@336px	**3**	1e-4	Adam	MultiStepLR

```
CIDEr: 0.8681001566774383 | CLIPScore: 0.7104499010882476
```

Encoder	# Decoder layer	learning rate	optimizer	scheduler
CLIP-Vit-L/14@336px	4	1e-4	Adam	MultiStepLR

```
CIDEr: 0.8750356996336179 | CLIPScore: 0.7062281037656337
```

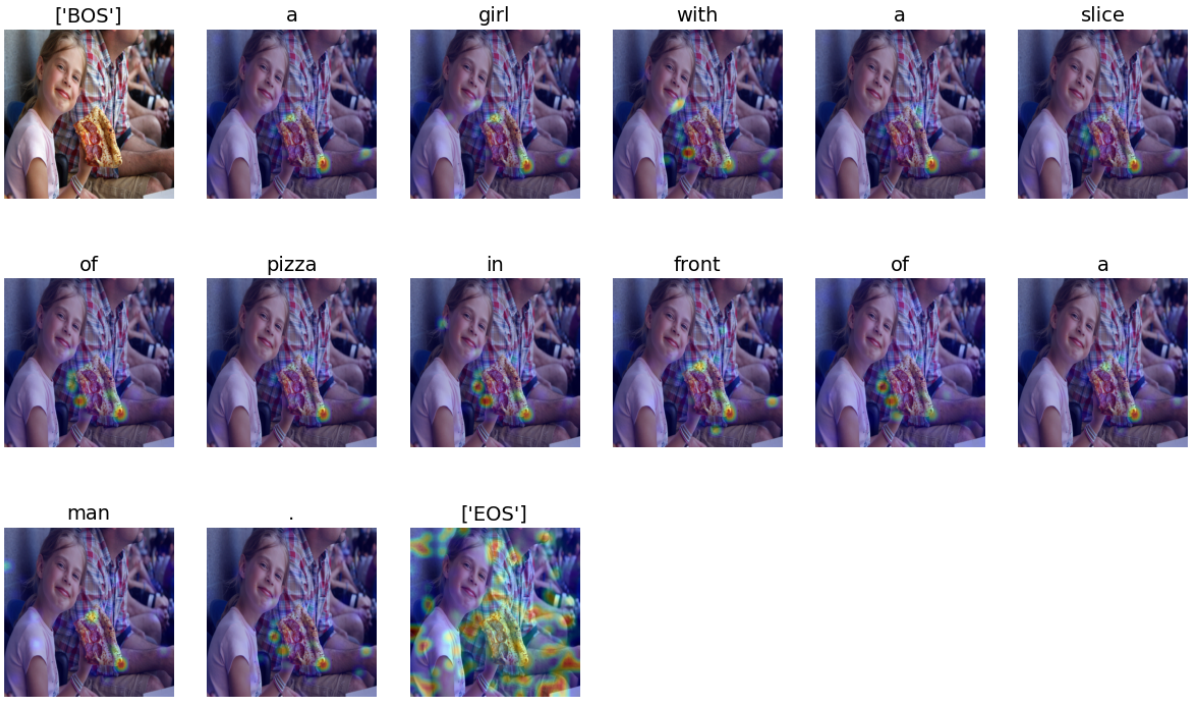
可以發現只改變decoder層數，CIDEr與CLIPScore的數值已經滿高，即使只有3層也達到接近strong baseline的水準。由上面兩點可以推論，learning rate對model表現的影響比較大。由於training時GPU在decoder layer為6時已達memory上限，若有能力增加更多數量的decoder layer就能更進一步測試層數與model表現之間的關係。

Problem 3 Visualization of Attention in Image Captioning

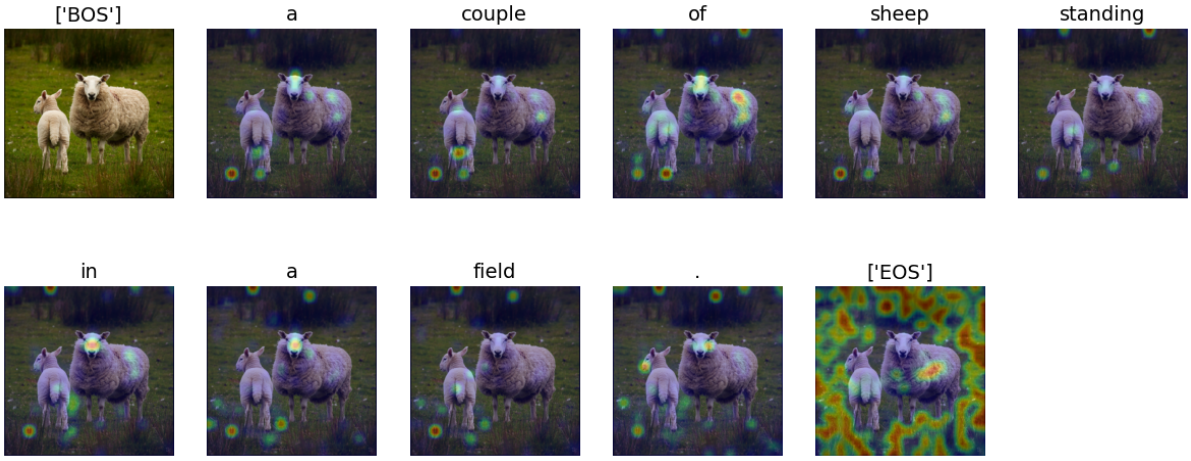
1. Please visualize the predicted caption and the corresponding series of attention maps

- bicycle.jpg : ***“a woman riding a bike down a street .”***

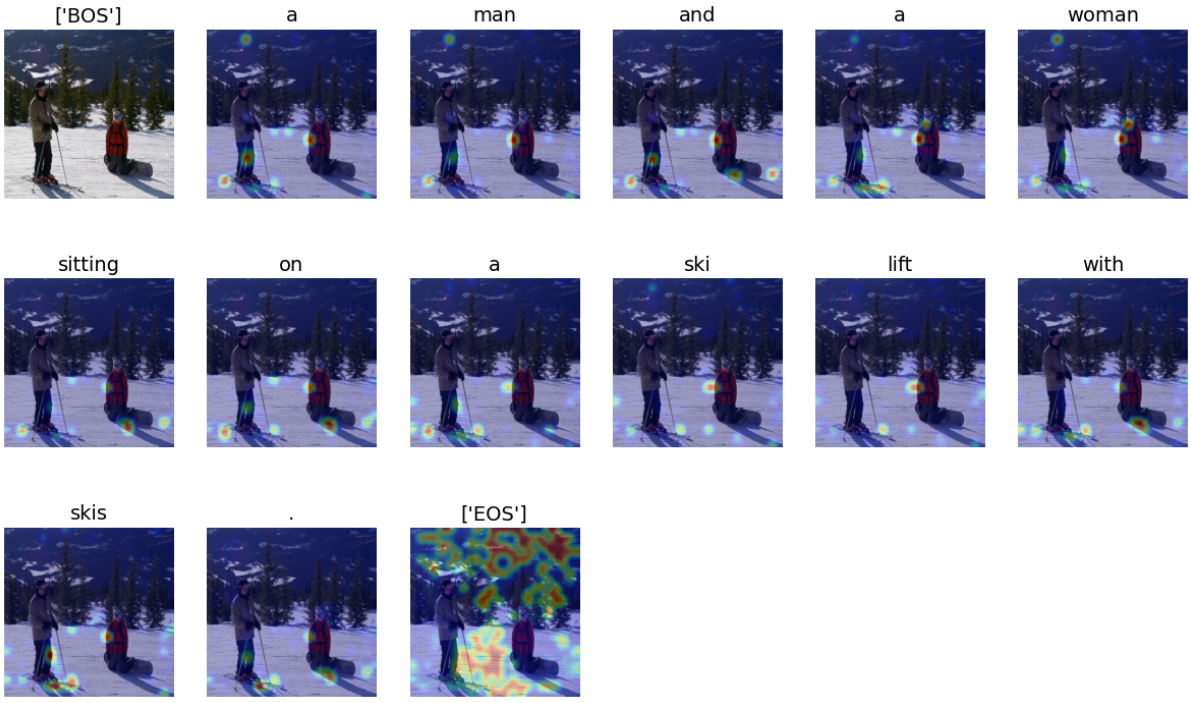
- girl.jpg : ***“a girl with a slice of pizza in front of a man .”***



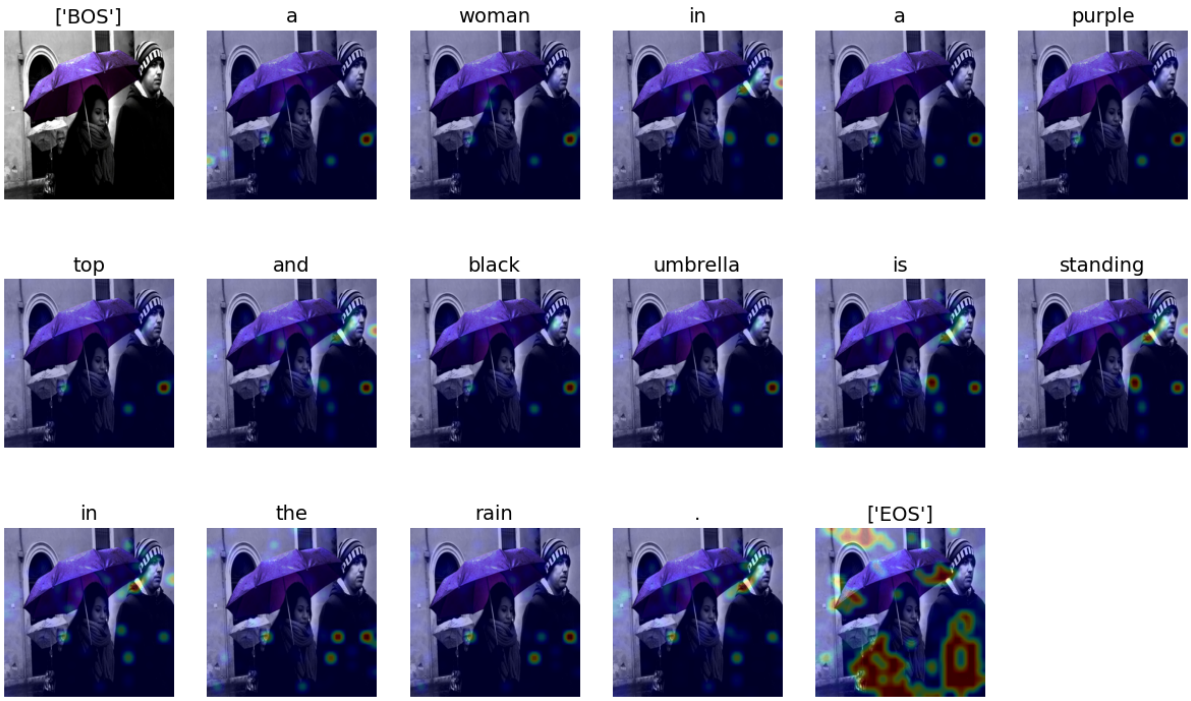
- sheep.jpg : ***“a couple of sheep standing in a field .”***



- ski.jpg : ***"a man and a woman sitting on a ski lift with skis ."***



- umbrella.jpg : ***"a woman in a purple top and black umbrella is standing in the rain ."***



2. Visualize top-1 and last-1 image-caption pairs and corresponding CLIPScore

- Top-1, CLIPScore : 0.9765625

- Last-1, CLIPcore : 0.37322998046875

a woman in a black dress and white photo

- Discussion

  在top1的圖片裡可以看到caption很符合圖片，CLIPScore高十分合理。在Last1的圖片本身就十分雜亂，因此CLIPScore不高也有其合理性。

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?

在上述幾個例子中可以看到每個caption都算合理，像是在bicycle.jpg和sheep.jpg的caption就能很清楚描述圖片；但是在其他例子中，雖然整體caption都有達到語意，有些細節就沒有描述好，像是girl.jpg裡應該是***"next to"***而非***"in front of"***，在umbrella.jpg裡雨傘的顏色也描述錯誤。

而在attention map中，圖片著重的地方大多都位於物體上，像腳踏車、pizza、羊、人、雪橇等等，而EOS的時候則主要著重在背景，作為caption的結束。但在每個字的attention map卻不如想像中的合理，或許是因為有些caption沒有描述好的緣故，attention的地方與每個字的關聯沒有那麼大。不過主要著重的object都有attention，若有機會改一下model讓CIDEr與CLIPScore分數更高可能會讓attention map畫得更好看。

Reference

1. https://github.com/openai/CLIP https://github.com/openai/CLIP
2. https://github.com/jadore801120/attention-is-all-you-need-pytorch/tree/master/transformer