DLCV hw3 電機碩一 R11921100 溫威領

Problem 1: Zero-shot image Classification with CLIP

1. Methods analysis

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.

利用 VGG、ResNet 等模型架構做影像分類的方法，是經由在特定資料集下訓練所得到的結果，因此只是在特定資料集下的 domain 去進行分類，當同樣的模型在分類其他資料集時，可能因 domain 有所差距而表現較差。

而使用 CLIP 進行 zero-shot 時，是用大量資料預訓練後的 text encoder 和 image encoder 將文字與圖片分別轉換到相同的 domain，再進行相似度比對，而非在特定資料集的 domain 下。我認為相較於分類，比較像是檢索問題，找 image 與哪個 text 最相近，所以在不同 domain 的資料集時，通常比 VGG、ResNet 等方法有較好的表現。

2. Prompt-text analysis

Please compare and discuss the performances of your model with the following three prompt templates:

i. “This is a photo of {object}” – acc: 60.86%

ii. “This is a {object} image.” – acc: 68.18%

iii. “No {object}, no score.” – acc: 56.30%

根據結果比較，正確率 ii > i > iii。

直覺上來看，iii 的 prompt templates，我認為會是最低的，因為有點雙重否定的感覺，可能較無法正確的 encoding，而且較不是一個常看到且正規的句子。

至於 i 和 ii 比較，我猜測是因為”a {object}”的關係，在”a”後的單字對於 CLIP model 可能有較高的重要性，所以 ii 的{object}緊接在”a”後，acc 會優於” a photo of {object}”。

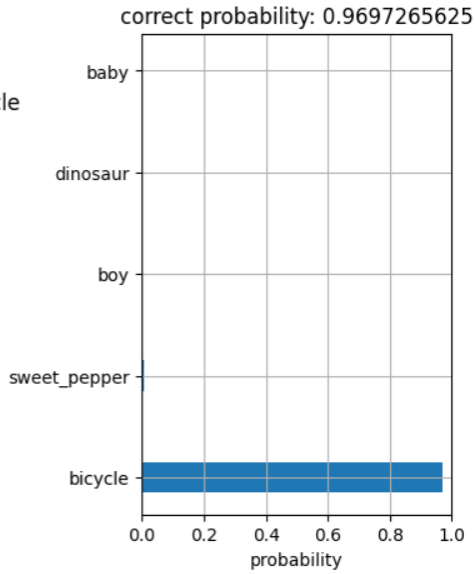
因此為了驗證這個猜測，稍微修改 i 的 prompt templates 為"a photo of a {object}." 可以達到 acc: 71.27%。

3. Quantitative analysis

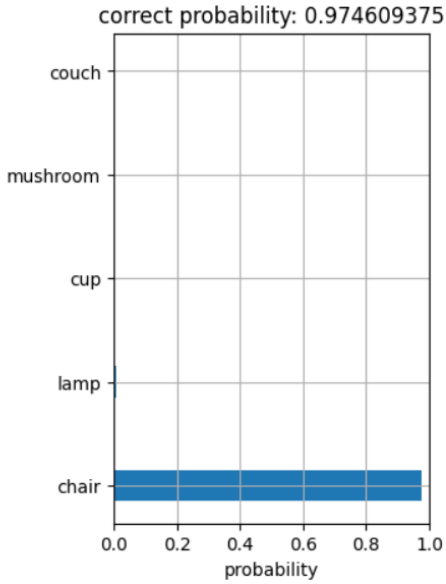
Please sample three images from the validation dataset and then visualize the probability of the top-5

*prompt templates: "a photo of a {object}."

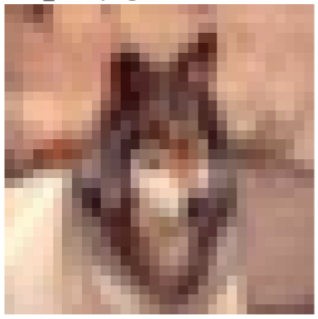
val/0_497.png, correct label: bicycle

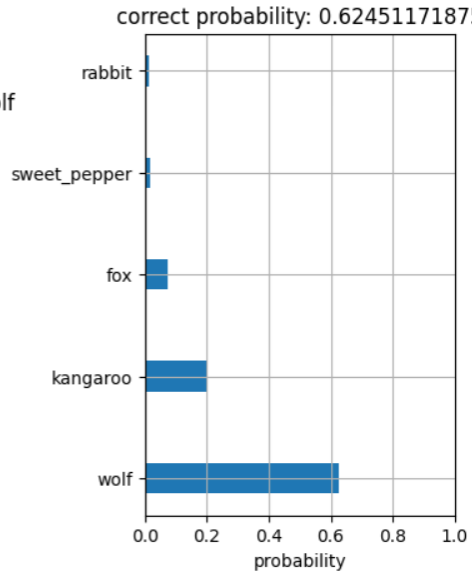


val/1_497.png, correct label: chair



val/25_497.png, correct label: wolf





Problem 2: Image Captioning with VL-model

Model Reference: https://github.com/saahiluppal/catr

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

Catr – backbone(ResNet101, pretrained)、transformer(non-pretrained)、memory hidden 256

CIDEr: 0.5119327468212173 | CLIPScore: 0.6185759559831785

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

i. Pretrained or not: all models weights are non-pretrained

Catr – backbone(ResNet101, non-pretrained)、transformer(non-pretrained)

CIDEr: 0.3521865168480471 | CLIPScore: 0.5320885465804873

ii. freezing layers: freezing backbone

Catr – backbone(ResNet101, freezing-pretrained)、transformer(non-pretrained)

CIDEr: 0.3405997494609855 | CLIPScore: 0.5261007774016123

iii. Different model architecture:

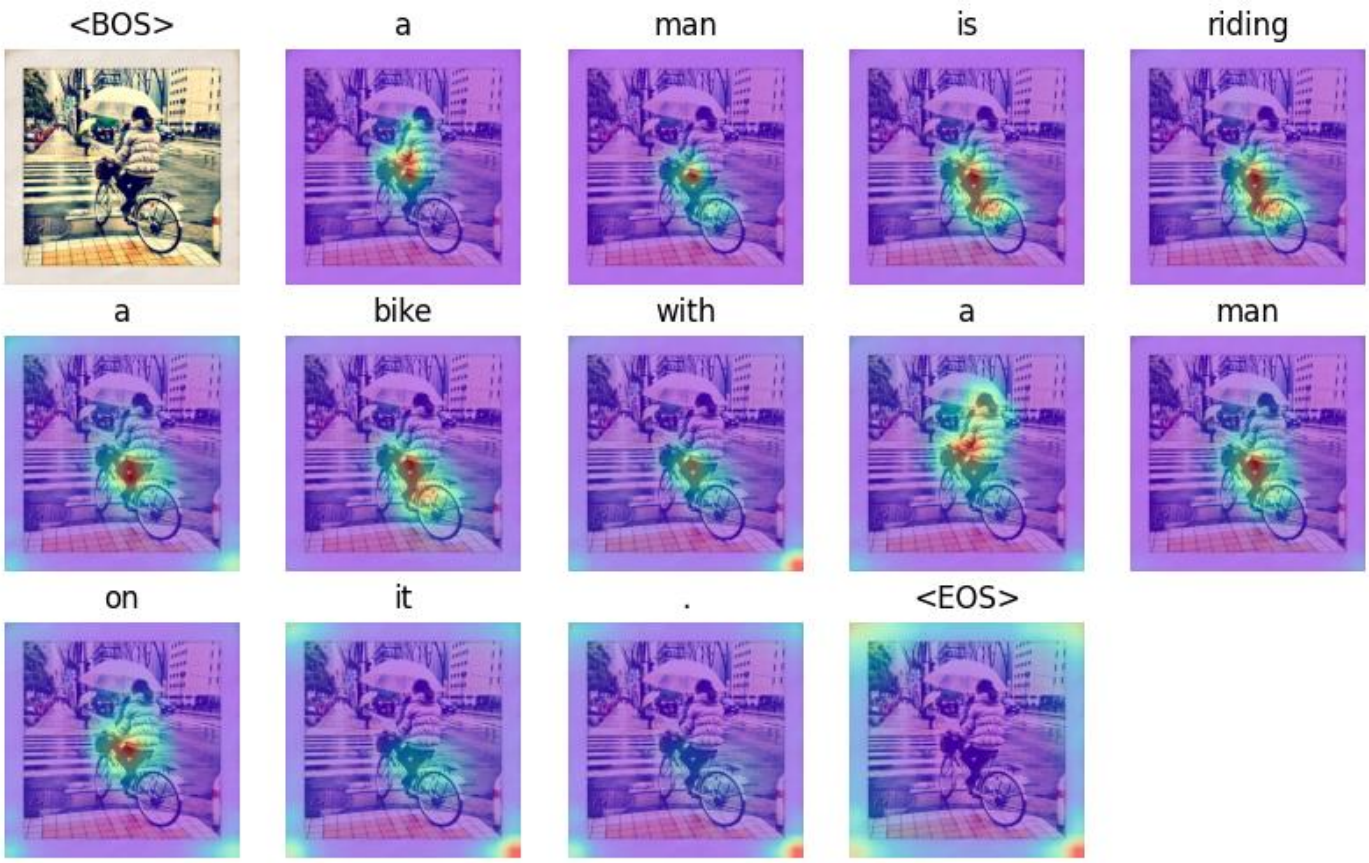
Catr – backbone(ResNet101, pretrained)、transformer(non-pretrained)、memory hidden 512

CIDEr: 0.4270766084542133 | CLIPScore: 0.5564406146011651

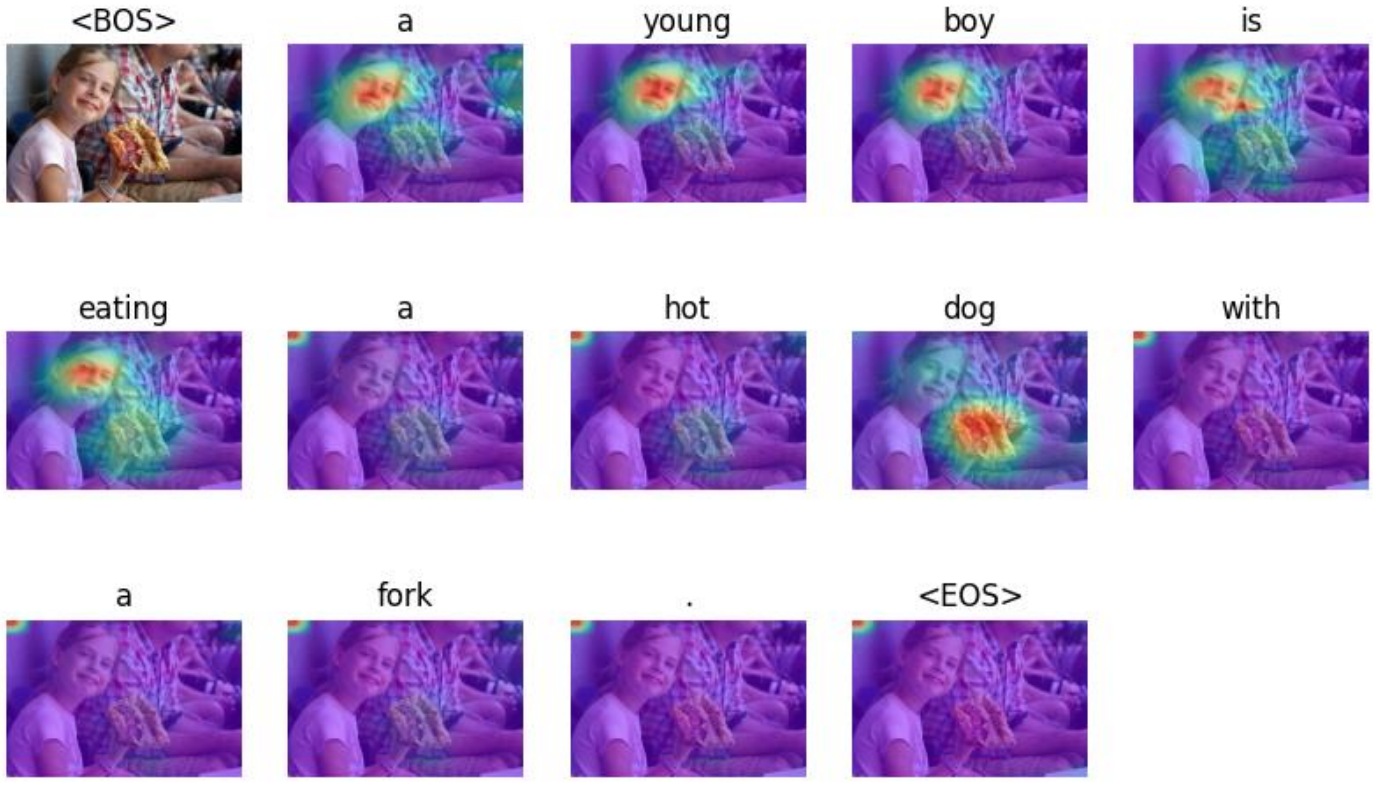
Problem 3: Visualization of Attention in Image Captioning

1. please visualize the predicted caption and the corresponding series of attention maps in your report

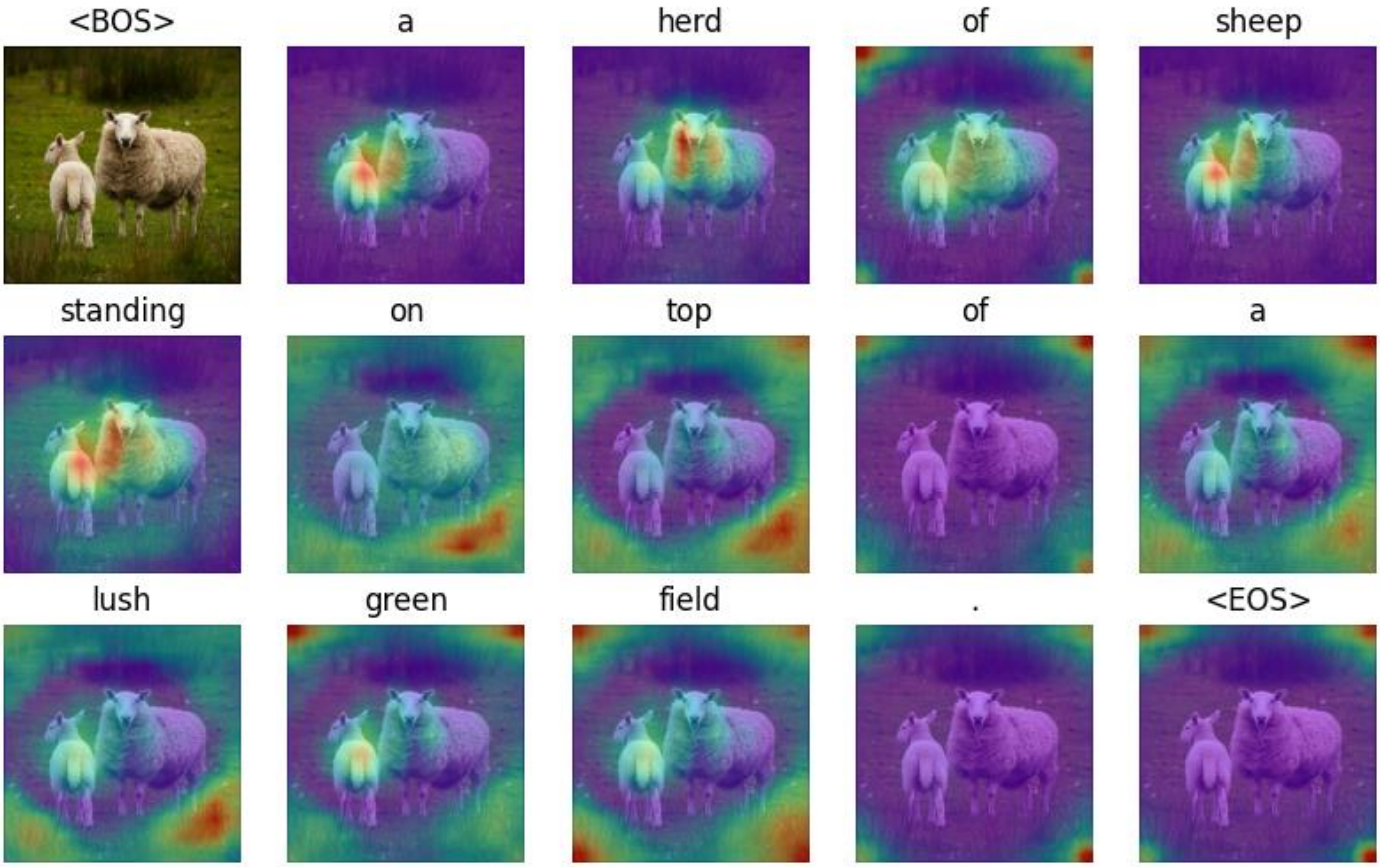
bike.jpg



girl.jpg



sheep.jpg



ski.jpg

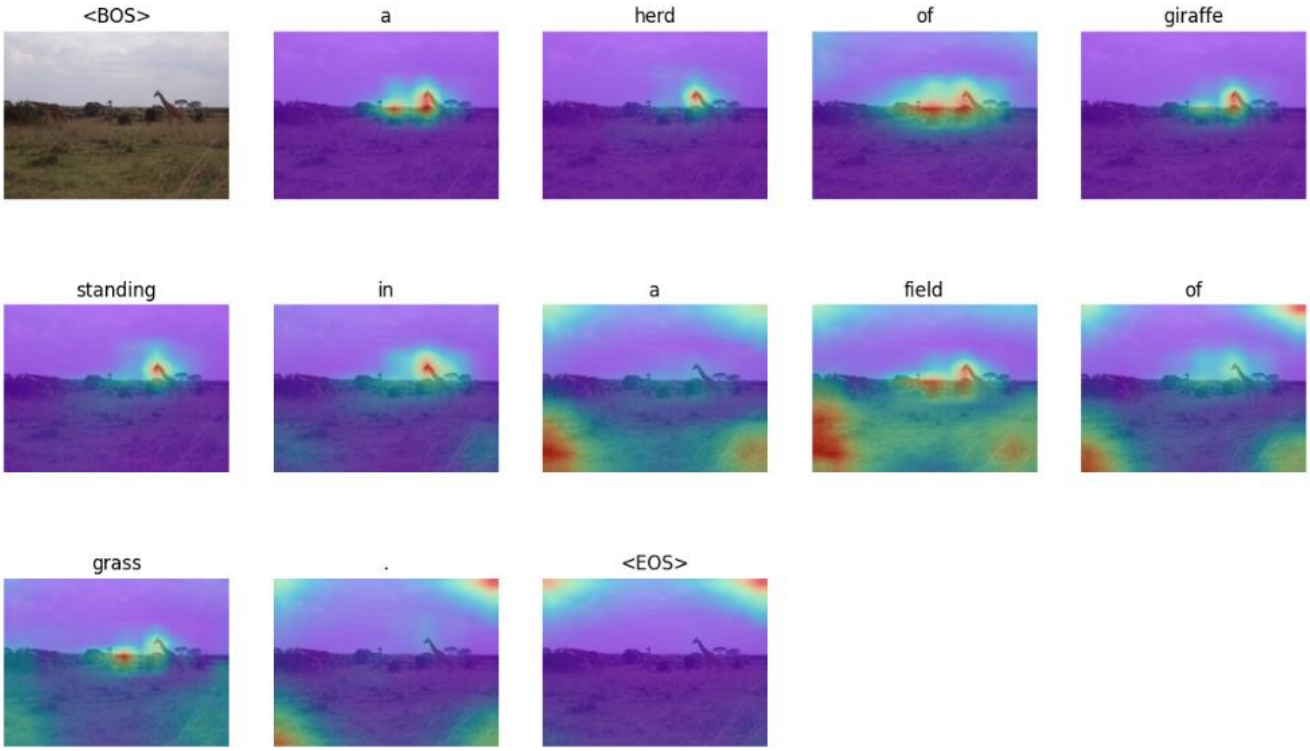
umbrella.jpg

2. According to CLIPScore, you need to visualize:

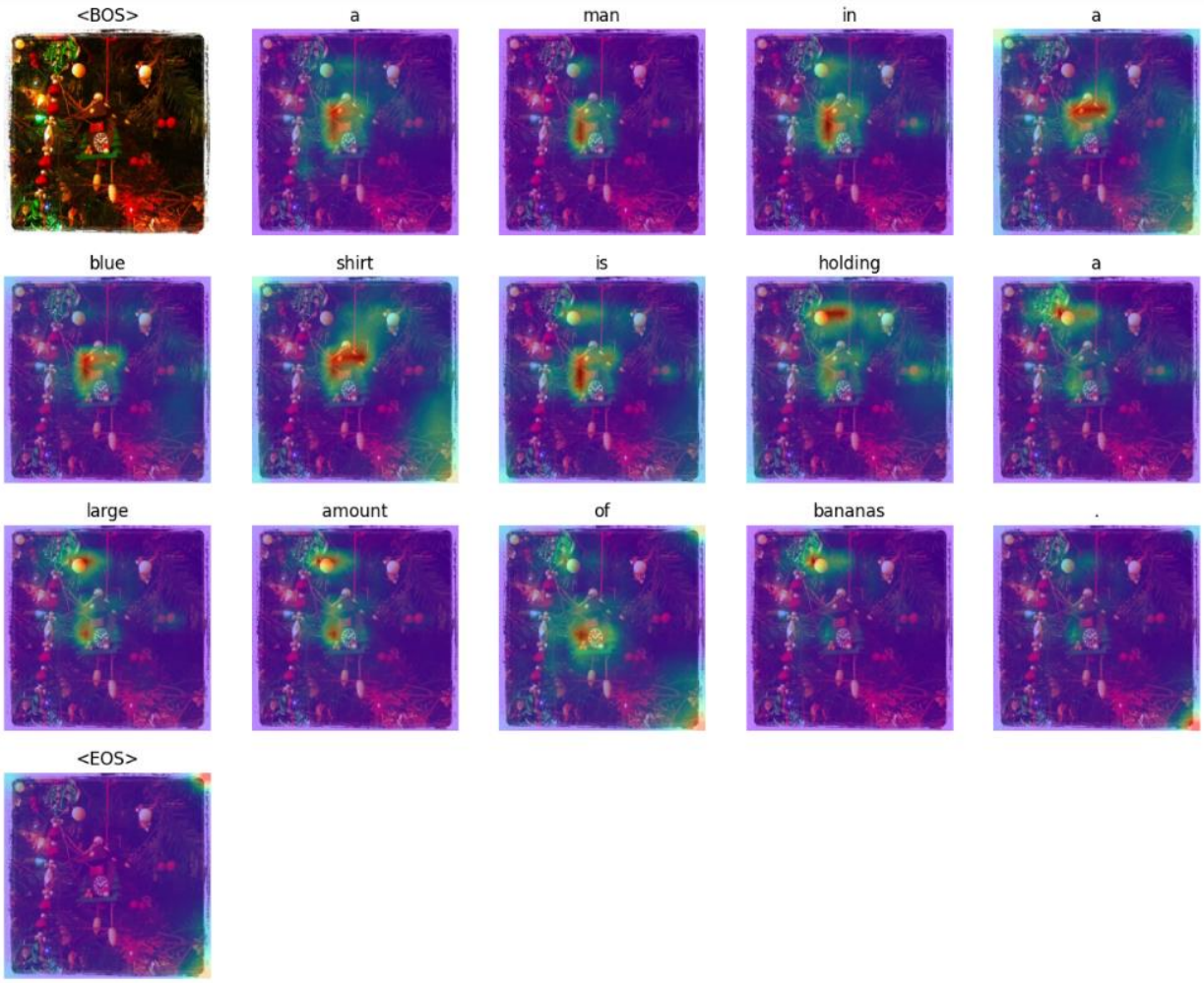
i. top-1 and last-1 image-caption pairs / ii. its corresponding CLIPScore

in the validation dataset of problem 2.

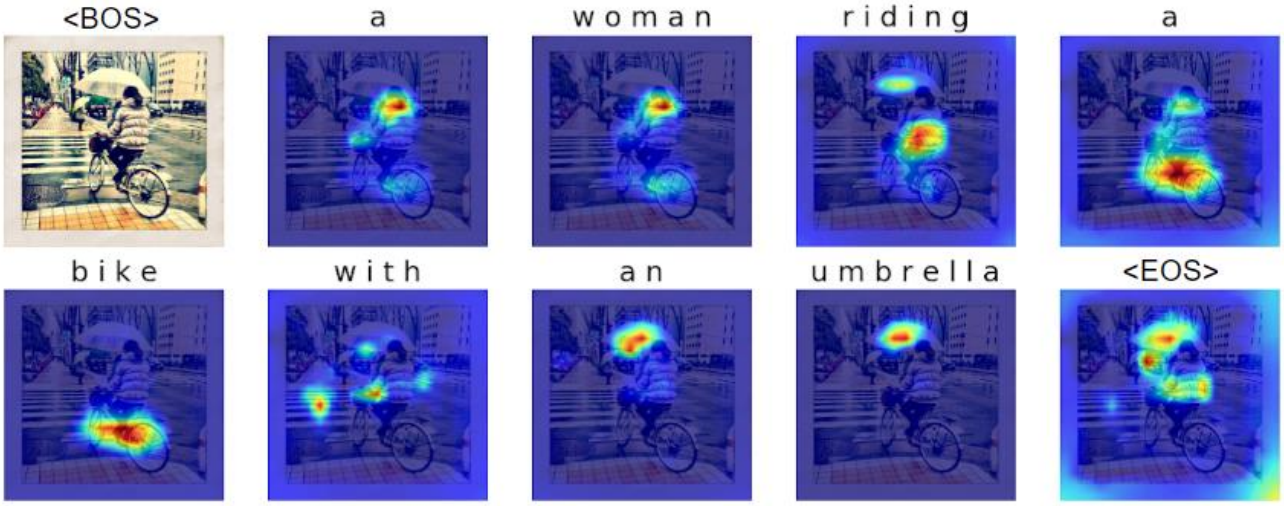
top-1: 000000330356.jpg, CLIPScore: 0.8770751953125



last-1: 000000244735.jpg, CLIPScore: 0.242919921875



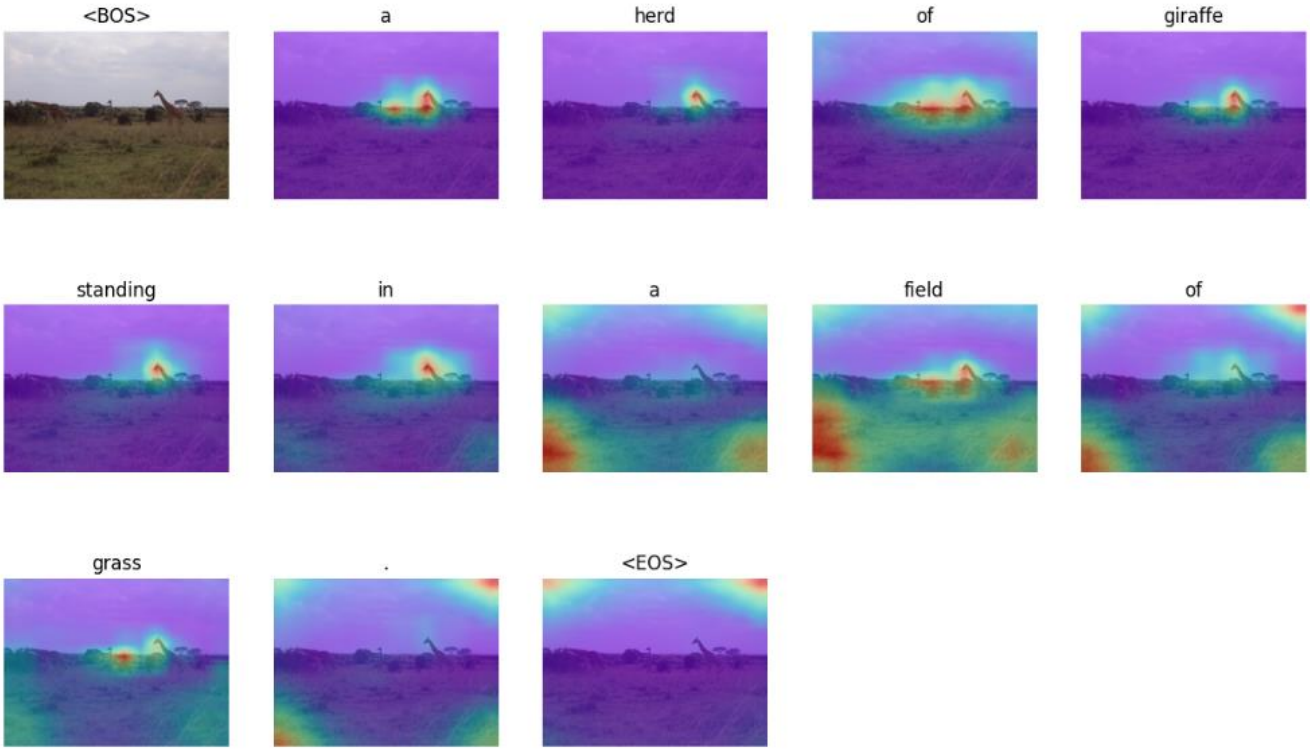
3.Analyze the predicted captions and the attention maps for each word according to the previous question. Is the caption reasonable? Does the attended region reflect the corresponding word in the caption?



範例

根據範例分析 predicted captions 和 the attention maps 之間的關係，名詞 text 都會對應到圖片上的 object，而”a”、”an”等冠詞 text 則會對應到下個名詞 text 所描述的 object，反之較抽象的動詞 text 或連接詞 text，可能會顯示出兩者物品對應的關係，或較無法看出有什麼關聯。

val/000000330356.jpg



從 CLIPScore 最高的 val/000000330356.jpg，也可以觀察 predicted captions 跟圖片中的 object 有很好的對應關係，可由 attention maps 解釋從圖片中的哪部分得到 text。