DLCV HW3 Report

R11921078 李鎮宇

Problem 1: Zero-shot image classification with CLIP

1. Methods analysis (3%)

   CLIP 的做法是先將影像和文字投射到空間中，透過對比訓練的方式，將相似的影像和文字拉近，不相干的則推遠。CLIP 因為使用了網路上大量的影像和文字資料，增強了模型的理解能力。以上的特性讓 CLIP 能夠對於未見過的資料進行分類，只要找到影像跟大量文字中相似度最高的 pair 即可進行預測。

2. Prompt-text analysis (6%)

Prompt-text	Accuracy
“a photo of a {object}”	**71.24%**
“This is a photo of {object}”	60.80%
“This is not a photo of {object}”	65.44%
“No {object}, no score.”	56.36%

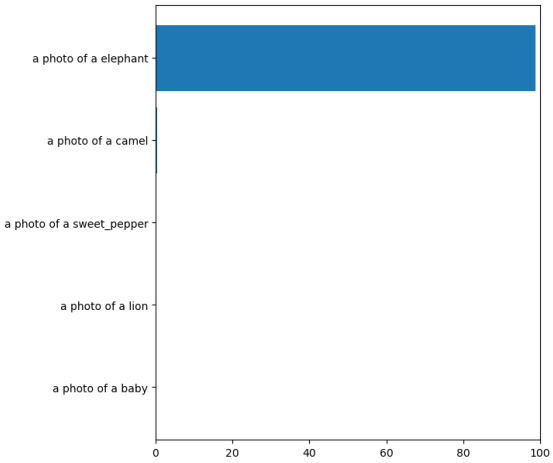
   "a photo of a {object}"的準確度最高，推測是因為 prompt 有明確的提示模型要從 photo 裡面找到與 object 有關的資訊，而且結構相似，都是 a … of a …，所以不會特別依賴前面的字詞，而這也說明為何分類表現會比"This is a photo of {object}"來的好，因為在這個 prompt 中，模型可能會比較傾向於"This is a photo"的特徵，而沒有關注到 object 的資訊。

   此外，“This is not a photo of {object}”準確度也蠻高的原因推測是因為 prompt 中引入了否定詞"not"，可以讓 CLIP 這種會將不相關資訊分離的模型能更好的區分"不是 object"的影像，進而提升分類準確度。

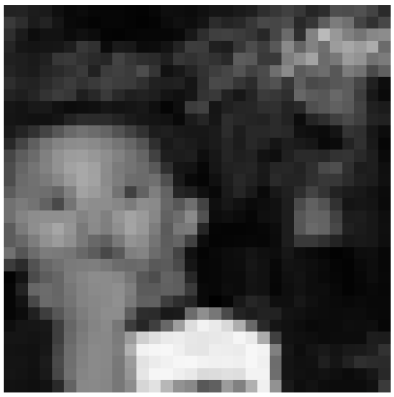
   "No {object}, no score."表現最差的原因應該是 prompt 缺乏了對 object 的描述(例如 a photo of…)，而且"no score"也沒有很明確的定義和影像之間的關係，所以導致這個 prompt 的效果最差。

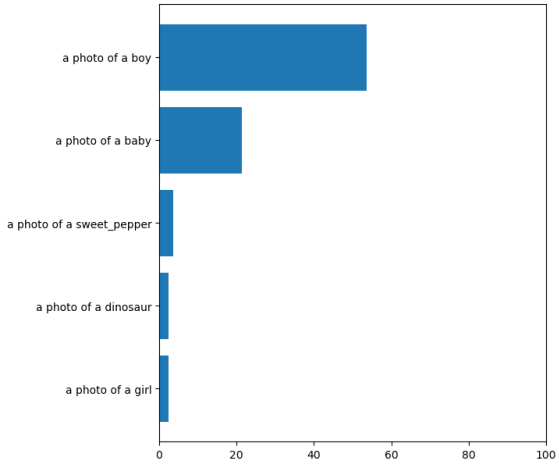
3. Quantitative analysis (6%)

- (44_473.png)

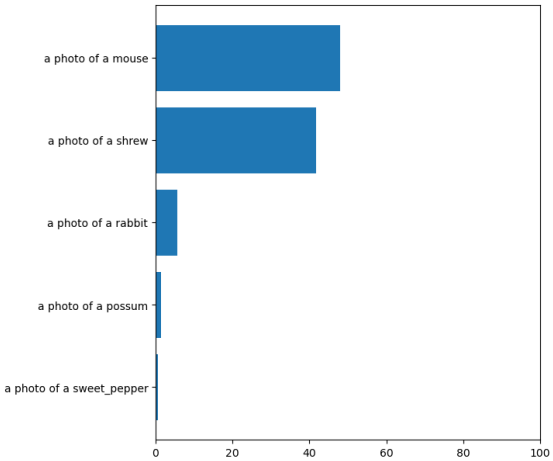


- (35_458.png)





- (33_498.png)



Problem 2: PEFT on Vision and Language Model for Image Captioning

1. Report your best setting and its corresponding CIDEr & CLIPScore on the validation data. (TA will reproduce this result) (2.5%)

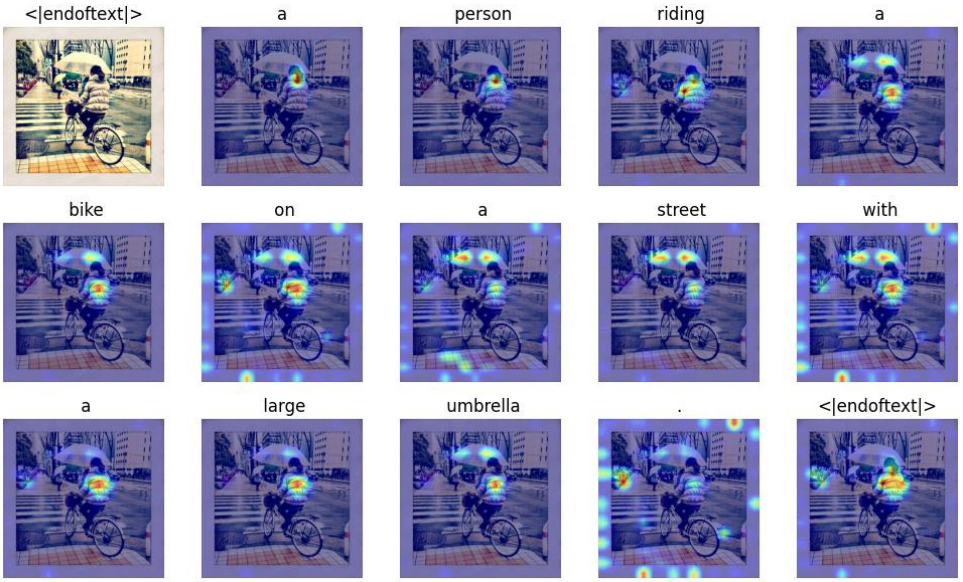
   ViT 使用 timm 中的"vit_huge_patch14_clip_224.laion2b_ft_in12k_in1k"這個 pre-trained model，將其設為 freeze，接一層 linear(1280, 768)轉換後，連接到 decoder。將 decoder 層數設定為 12，並對 linear、Attention 中的 c_attn 和 c_proj，以及 mlp 層皆使用 Lora 減少參數運算。訓練的參數設定：lr=0.00008, batch size=4, epochs=5。使用 cross entropy 作為 loss function、Adam 作為 optimizer。validation loss 最低的模型選為 best model，對應的分數為：**CIDEr=0.8762, CLIPScore=0.7207**。

2. Report 3 different attempts of PEFT and their corresponding CIDEr & CLIPScore. (7.5%, each setting for 2.5%)

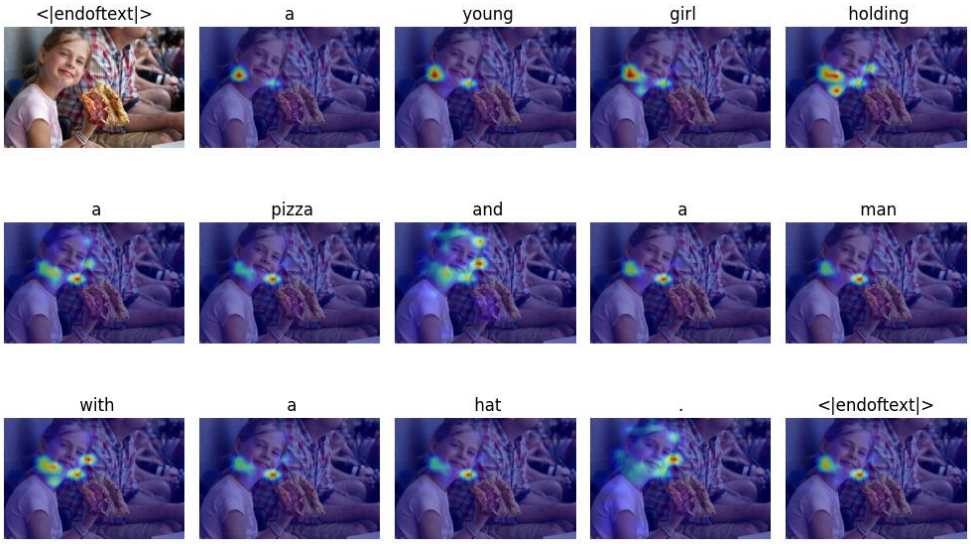
   (ps.以下 Table 中的 decoder 皆使用 6 層，並且都使用 best model 產生 caption)

PEFT	CIDEr	CLIPScore
Adapter	0.7703	0.7124
Lora	0.7774	0.7107
Prefix Tuning	0.8079	0.7155

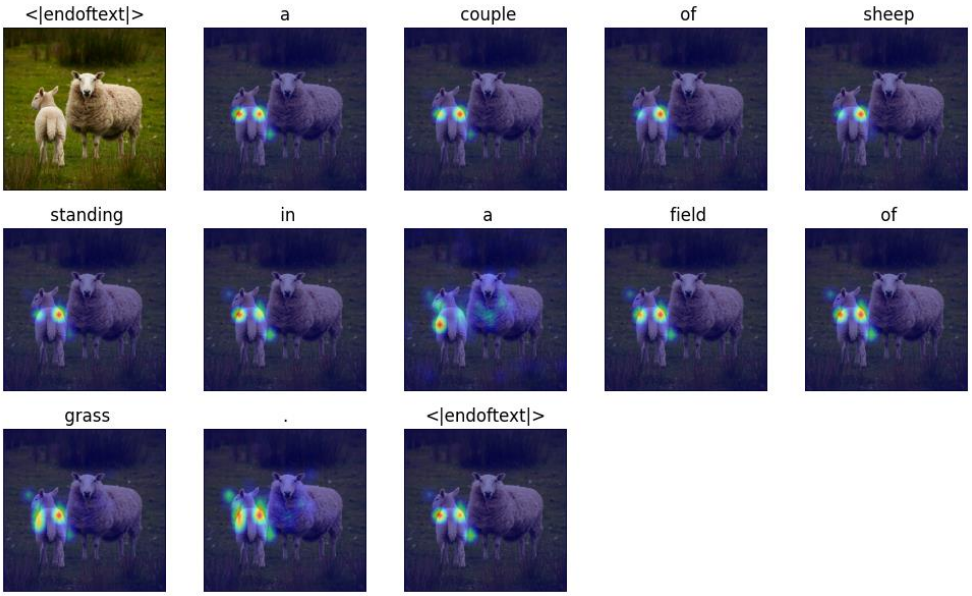
3. Visualize the **predicted caption** and the corresponding series of **attention maps.** (皆使用# 6 head 進行視覺化)



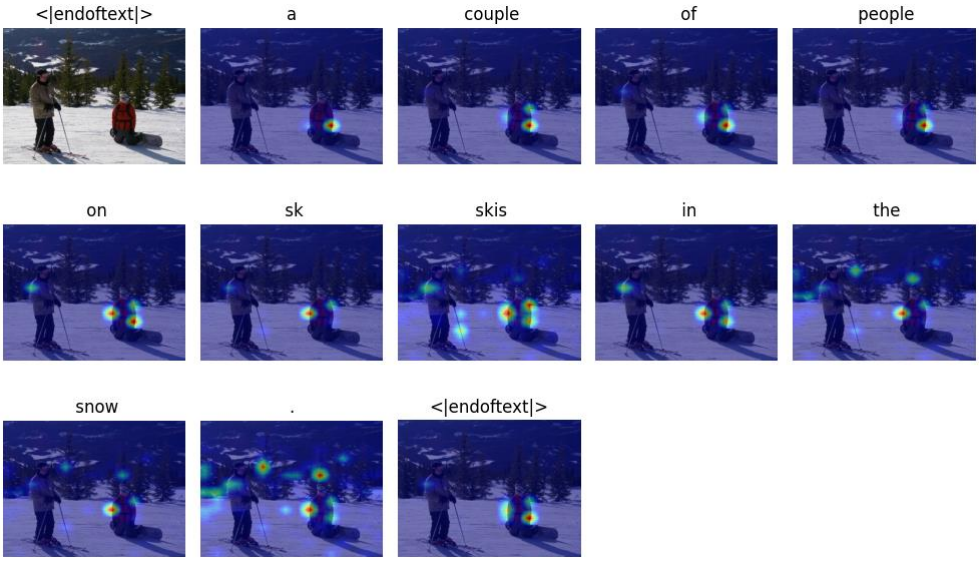
girl


<|endoftext|>
a
young
girl
holding
a
pizza
and
a
man
with
a
hat
.
<|endoftext|>

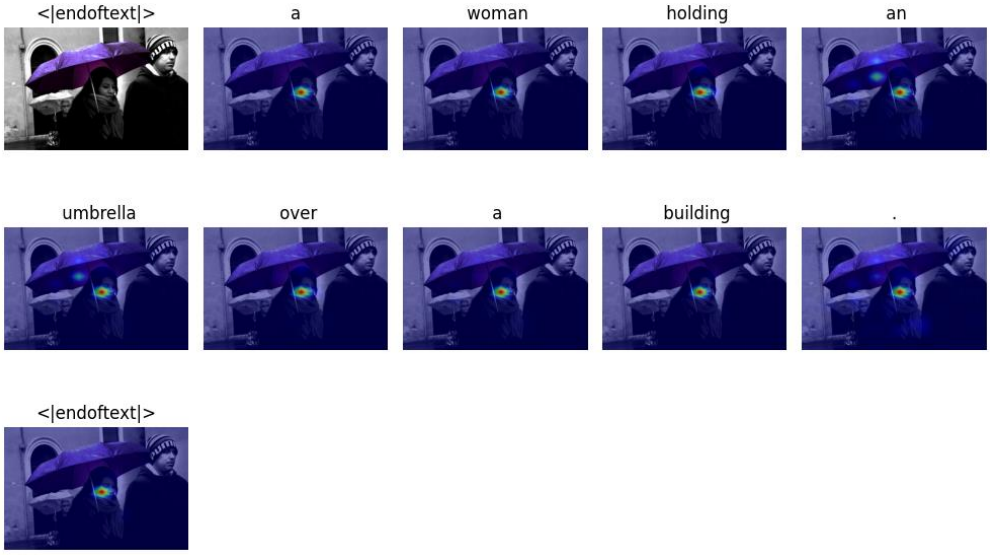
sheep


<|endoftext|>
a
couple
of
sheep
standing
in
a
field
of
grass
.
<|endoftext|>

ski



umbrella



4. According to **CLIPScore**, you need to:
   i. visualize top-1 and last-1 image-caption pairs
   ii. report its corresponding CLIPScore
   in the validation dataset of problem 2. (5%)

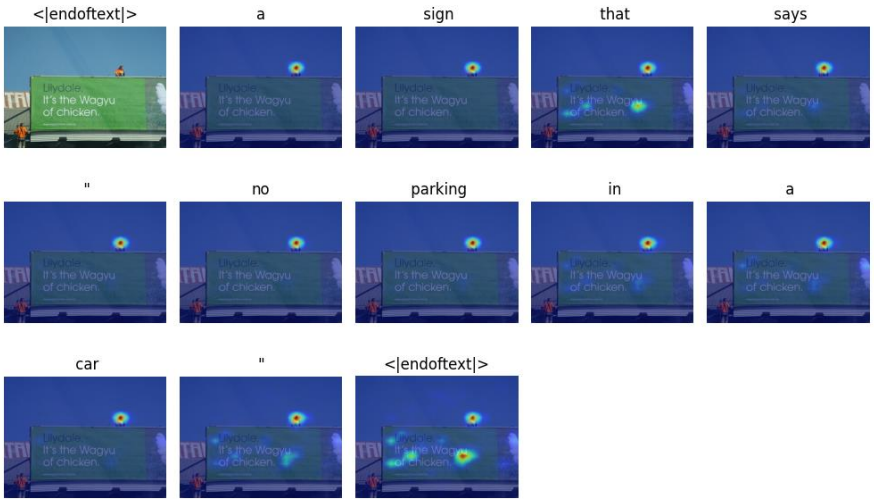
Top-1

Visualization: (# 11 head)

CLIPScore: 0.9789580851793289

Last-1

Visualization: (# 11 head)



CLIPScore: 0.4040498659014702

5. 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? (5%)

   從 top-1 可以看到在這個 attn map 中，模型比較著重在小區塊，其中香蕉的部分有被模型很好地看到，並且能反映出對應的 caption。在 last-1 則是一直將重點放在招牌上的人，而沒有去注意招牌內容，這可能也導致該影像預測的 caption 效果很差。不過從 top-1 或是 last-1 整體生成的 caption 和 heatmap 來看，模型其實沒有很好地反映 caption 所關注的位置，我想主要是因為 decoder 被 trained 的參數不多，模型不能完整地學習 image 和 caption 間的相似性，另一個部分則是因為只看 multi-head 中的其中一個 head，而每個 head 關注的都是影像中不同的特徵，因此一個 head 是無法反映出整體的狀況。

Reference:

1. prefix-tuning: https://github.com/kipgparker/soft-prompt-tuning
   decoder_ptuning.py 中的 SoftPrompt 參考該 github。