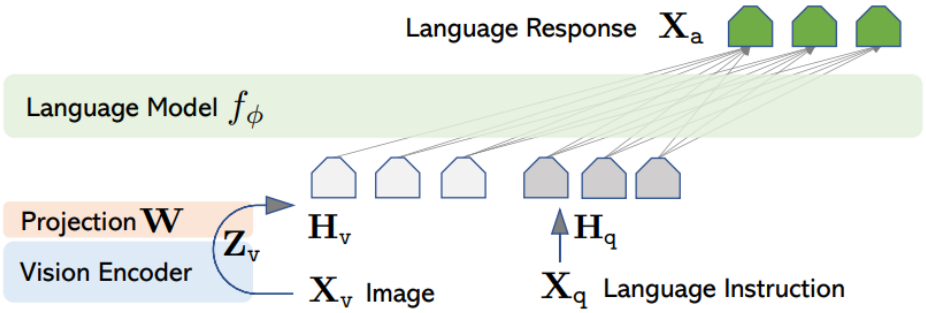
R12521504 吳玉文

Problem 1

1. *Please read the paper "Visual Instruction Tuning" and briefly describe the important components (modules or techniques) of LLaVA.*



使用預訓練的 CLIP 視覺編碼器 ViT-L/14 處理輸入圖像 X_v，生成視覺特徵 $Z_v = g(X_v)$，然後透過線性層和可訓練的投影矩陣 W 將這些特徵映射到詞嵌入空間，生成語言嵌入 token H_v，再與作為指引的 X_q 一同丟入到語言模型（這篇使用的是 Vicuna），最後生成回應 X_a。

2. *Please come up with two settings (different instructions or generation config). Compare and discuss their performances.*

	第一種	第二種
instruction	prompt = "<image>\n Generate a concise caption that describes the main subject, action, or context in the image. Focus on essential elements, such as people, objects, and activities, while keeping the description straightforward and informative.\nCaption:"	
generation config	"max_new_tokens": 20, "do_sample": True, "num_beams": 8, "top_k": 25	"max_new_tokens": 15, "do_sample": True, "num_beams": 8, "top_k": 25
results	CIDEr:1.163812992898 CLIPScore:0.78136657	CIDEr:1.180051765790 CLIPScore:0.77137695

從上面的結果發現，max_new_tokens 減少，會使得 CIDEr 分數增加，而 CLIPScore 分數減少（用其他 config 嘗試也得出類似結果）。這是因為 CIDEr 偏好較精簡的描述，而越長的句子通常越能符合 CLIPScore 所需要的圖像與文本的對應性。

Problem 2

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

CIDEr	*CLIPScore*
1.020476103911773	0.734683837890625

VIT 使用的是 vit_gigantic_patch14_clip_224。把圖像特徵經過可訓練 Projector 後所得到的 image token 跟以及 text embedding 的 token 進行 concat 後再餵給 Language Model。使用 LoRA 微調的部分是 Attention 中的 c_attn 和 c_proj。

2. *Report 2 different attempts of LoRA setting (e.g. initialization, alpha, rank…) and their corresponding CIDEr & CLIPScore.*

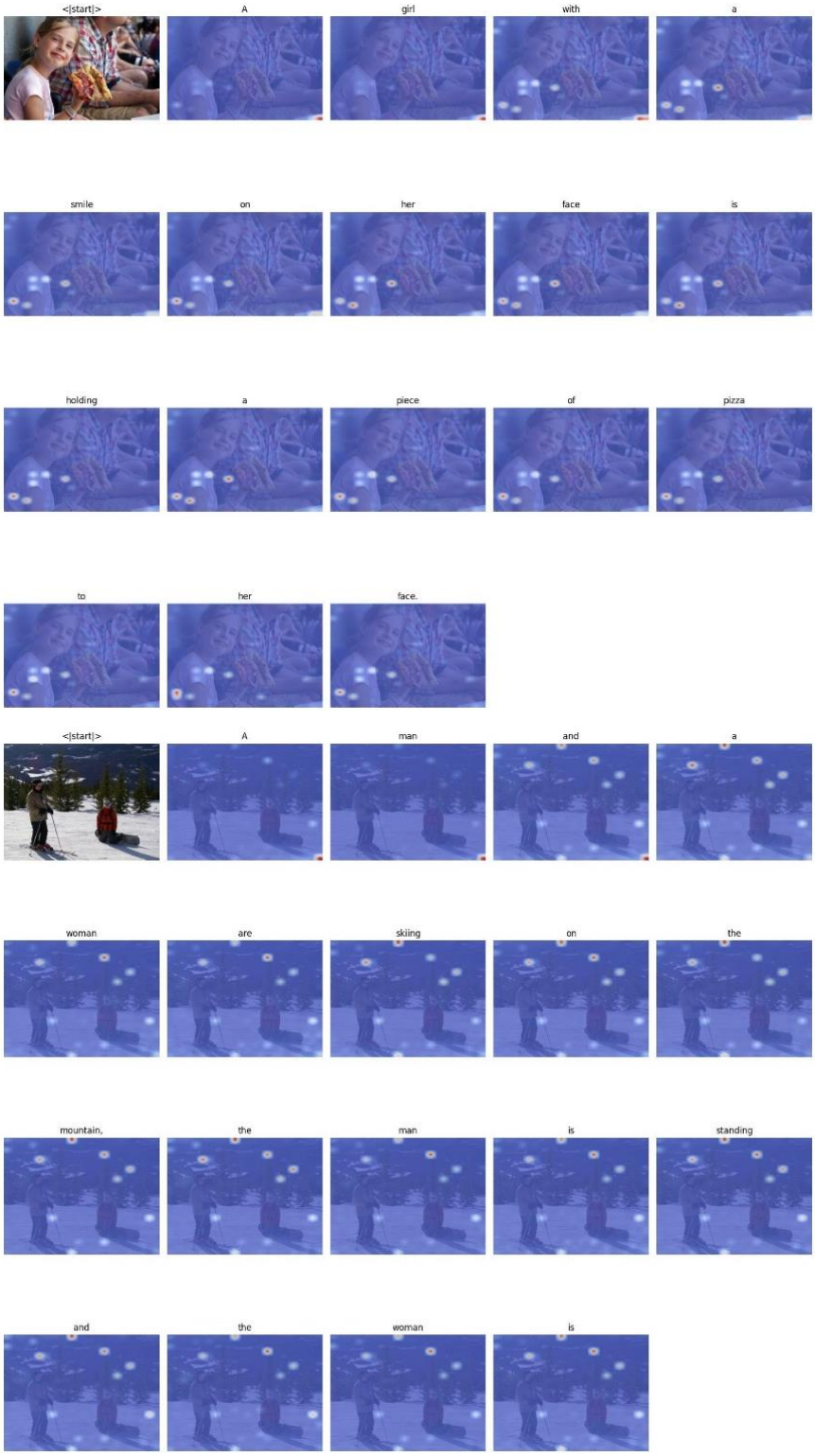
有使用 LoRA 的部分 r 均等於 40	
CIDEr	CLIPScore
1.020476103911773	0.734683837890625
有使用 LoRA 的部分 r 均等於 8	
CIDEr	CLIPScore
0.9156244241584307	0.7313275146484375

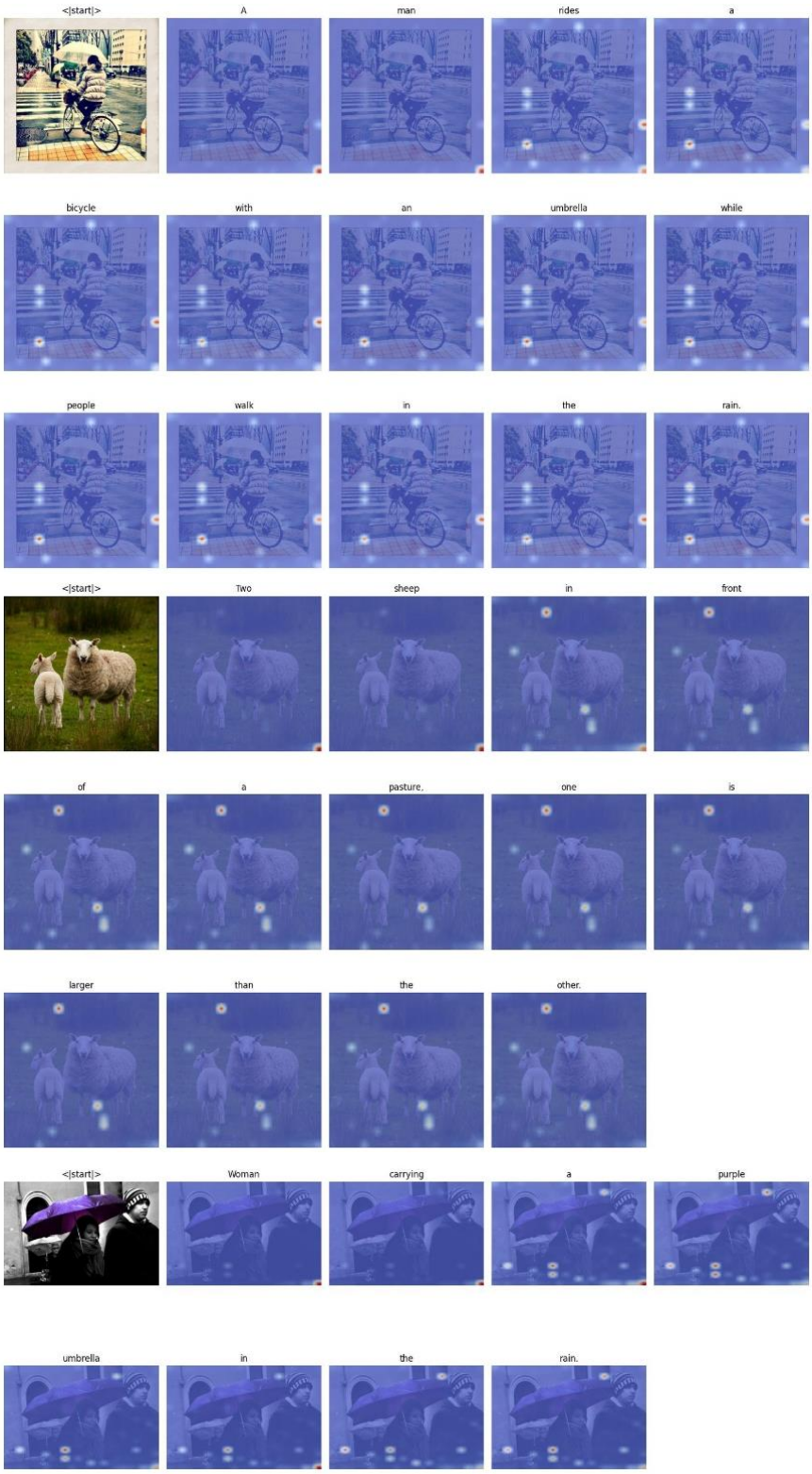
縮小 r 讓分數變的比較差，可能是因為可訓練的參數變少了，所以效果沒有那麼好。

Problem 3

1. *Given five test images ([p3_data/images/]), and please visualize the predicted caption and the corresponding series of attention maps.*

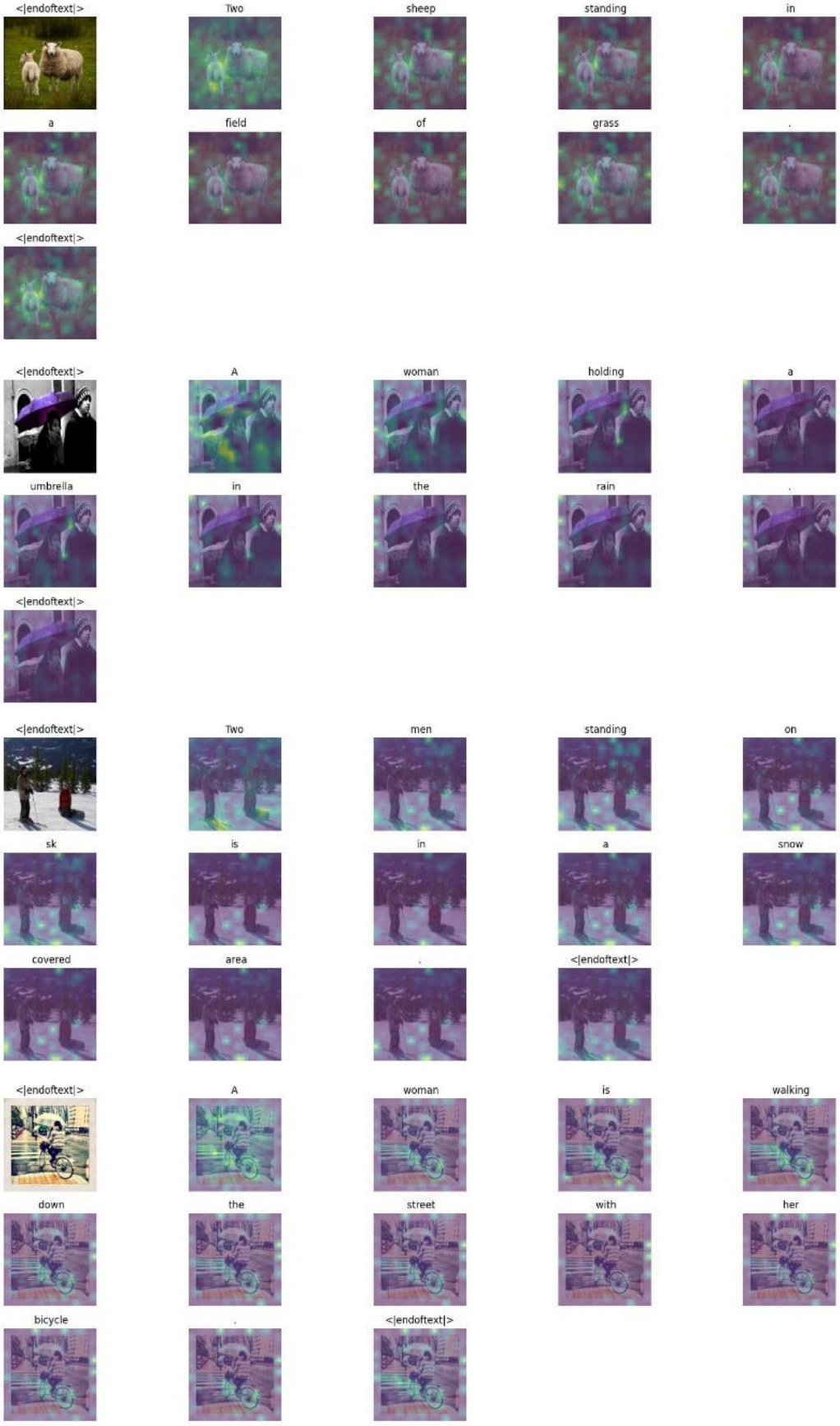
a. problem 1





LLaVA 模型可能在早期層就已經將圖像資訊充分融合到內部表徵中，而最後幾層更多是在於「語言內部的注意力」，或者只是進行微調的語義整合。換句話說，模型在最後一層的自注意力未必會明顯顯示「某個詞對應圖像中的某個區域」。這可能導致在最後一層進行可視化時，注意力圖「看起來沒有對上位置」。

b. problem 2


<|endoftext|>
Two
sheep
standing
in
a
field
of
grass
.
<|endoftext|>
<|endoftext|>
A
woman
holding
a
umbrella
in
the
rain
.
<|endoftext|>
<|endoftext|>
Two
men
standing
on
sk
is
in
a
snow
covered
area
.
<|endoftext|>
<|endoftext|>
A
woman
is
walking
down
the
street
with
her
bicycle
.
<|endoftext|>

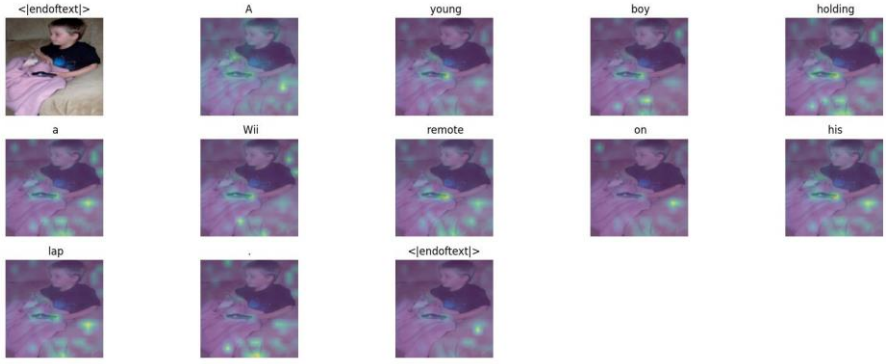
2. *According to CLIPScore, you need to:*

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

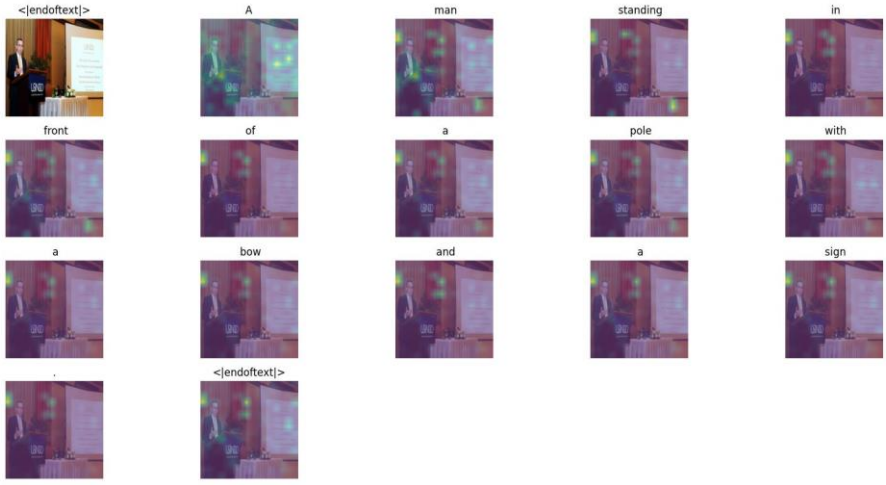
   ii. ii. report its corresponding CLIPScore

   in the validation dataset of problem 2

Top-1 (Image: 000000001288.jpg, CLIPScore: 0.996)



Last-1 (Image: 000000001523.jpg, CLIPScore: 0.316)



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?*

Top-1 的 caption 比起 Last-1 合理許多，attended region 也的確有對應到輸出的 token，反應出模型對於 Top-1 圖片的理解比較高，對於 Last-1 可能只是瞎猜。

感謝 ChatGPT 大神