Deep Learning for Computer Vision

NTU, Fall 2024, Homework3

電機所碩一 李彥璋 **R13921093**

Problem 1: Zero-shot image captioning with LLaVA (9%)

1. **Please read the paper "Visual Instruction Tuning" and briefly describe the important components (modules or techniques) of LLaVA. (3%)**

   架構

   - 使用 CLIP ViT-L/14 作為視覺編碼器來處理圖像。
   - 使用 Vicuna (基於 LLaMA) 作為語言解碼器/模型。
   - 採用簡單的線性投影層來連接視覺特徵和語言嵌入。
   - 將視覺特徵轉換為與詞嵌入空間維度相同的語言嵌入標記。

   訓練過程

   - 第一階段:特徵對齊預訓練
     - 凍結視覺編碼器和語言模型權重。
     - 僅訓練投影矩陣以對齊視覺特徵和語言模型詞嵌入。
   - 第二階段:端到端微調
     - 保持視覺編碼器凍結但更新投影層和語言模型權重。
     - 針對多模態聊天機器人、科學問答進行微調。

2. **Please come up with two settings (different instructions or generation config). Compare and discuss their performances. (6%)**

Instruction : What are these?, num_beams : 3	
CIDEr	CLIPScore
0.002125662287249568	0.79108642578125
Instruction : What are these?, num_beams : 1	
CIDEr	CLIPScore
0.027123552847190226	0.752618408203125

比較相同instruction (What are these?) 在不同num_beams下的結果。

可以發現，CIDEr表現欠佳;而CLIPScore則還不錯。分析原因如下，

- CLIPScore高的原因:CLIP是通過圖文對比學習訓練的，專注於圖像和文字的語義匹配度，What are these? 這類問題通常會引導模型直接描述圖片中的主

要物體，這種直接的物體描述往往與CLIP的訓練目標很接近，因為CLIP就是學習圖像和文字描述的對應關係。

- CIDEr低的原因：CIDEr評分注重的是生成的描述與人工標註的參考描述的相似度，人工標註通常會包含更豐富的內容，如物體之間的關係、動作、場景描述等，What are these? 容易導致模型只列舉物體，缺乏對場景的完整描述，這種簡單的物體列舉與人類寫的豐富描述差異較大。

num_beams的部份，其效果為，

- num_beams = 1：等同於貪婪搜索，每步只選機率最高的詞。
- num_beams > 1：考慮更多可能的路徑，可能找到整體更好的描述。
- 較大的 num_beams 通常能產生更好的結果，但也會增加計算時間。

在實驗結果當中，CLIPScore在num_beams較高的情況下分數上升；而CIDEr在num_beams較高的情況下反而下降，顯示其並非絕對。

Problem 2: PEFT on Vision and Language Model for Image Captioning (10%)

1. **Report your best setting and its corresponding CIDEr & CLIPScore on the validation data. Briefly introduce your method. (5%)**

Best setting

- batch_size = 16
- num_epochs = 10
- optimizer = AdamW
- scheduler = OneCycleLR, max_lr = 1e-3
- lora_rank = 58

CIDEr	CLIPScore
1.149792132760579	0.7446038818359375

Method

- 視覺編碼器
  - 使用 CLIP ViT-Large (OpenAI 版本) 預訓練模型。
  - 從圖片中提取視覺特徵 (1024 維) 。
  - 訓練過程中完全凍結，不更新參數。
- 解碼器
  - 基於 GPT 架構，採用參數高效微調 (PEFT) 。
  - 主要組件
    - 投影層：將 1024 維視覺特徵轉成768 維，中間加入非線性層。
    - 使用 LoRA 的 Transformer 層：使用 rank-58 實現高效訓練。
  - 只訓練投影層和 LoRA 參數 (總參數量 < 10M) 。

- 輸入處理
    - 序列開頭加入 <|endoftext|> token。
    - caption 後加入 <|endoftext|> 作為結束符。
    - 將序列填充到固定長度 64，填充使用 <|endoftext|> token。
    - 標籤序列填充部分用 -100，使 loss 忽略這些位置。
- 生成機制
    - 輸入起始 token <|endoftext|>。
    - 每次取最後一個位置的預測結果。
    - 用 argmax 選擇最可能的下一個 token。
    - 將選擇的 token 加入序列，繼續預測。
    - 直到生成 <|endoftext|> 或達到長度限制 (30) 。

2. **Report 2 different attempts of LoRA setting (e.g. initialization, alpha, rank…) and their corresponding CIDEr & CLIPScore. (5%, each setting for 2.5%)**

lora_rank = 29	
CIDEr	CLIPScore
1.0288681047730925	0.7278727722167969
lora_rank = 1	
CIDEr	CLIPScore
1.1095884202943644	0.7464169311523438

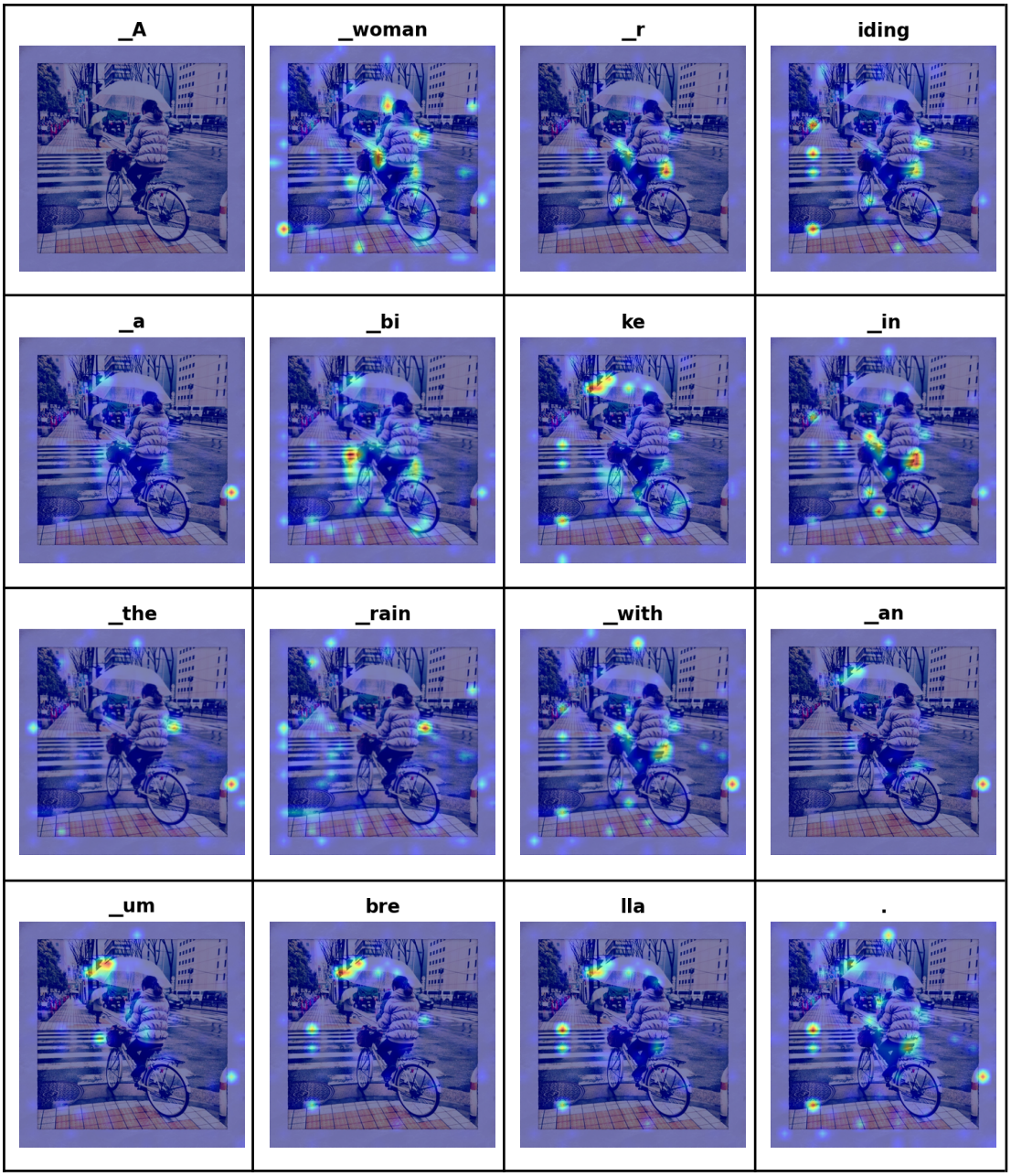
Problem 3: Visualization of Attention in Image Captioning (26%)

1. **Given five test images, and visualize the predicted caption and the corresponding series of attention maps in your report. (20%, each image for 2%, you need to visualize 5 images for both problem 1 & 2)**

   Problem 1

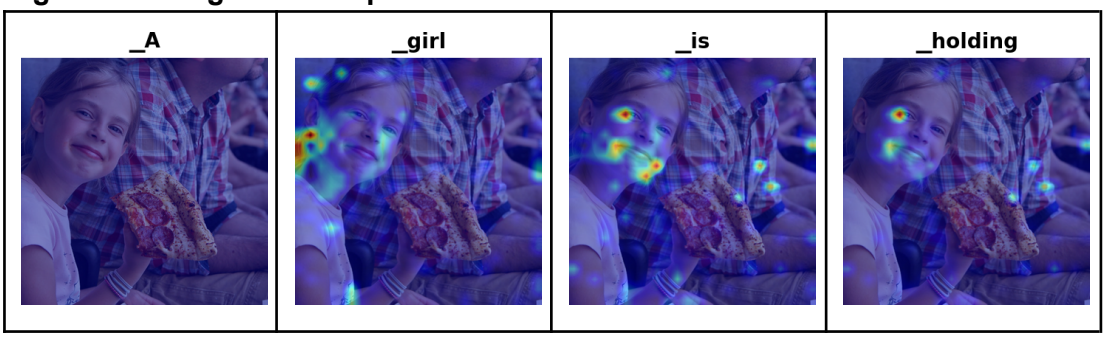
   bike.jpg

   A woman riding a bike in the rain with an umbrella.



girl.jpg

A girl is holding a slice of pizza in front of a man.

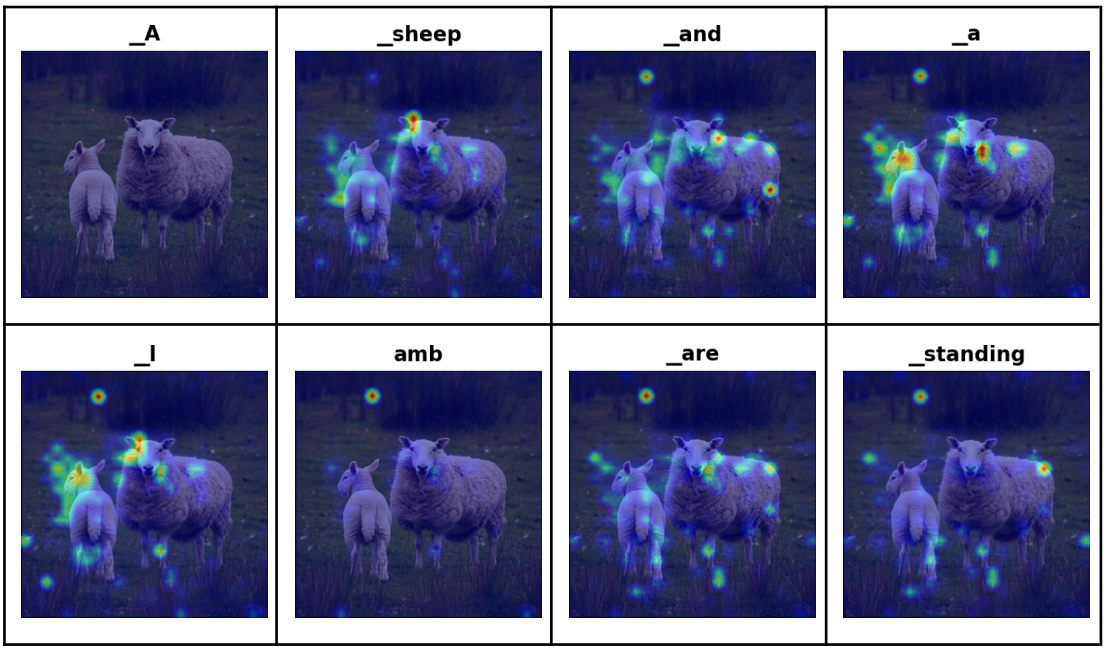


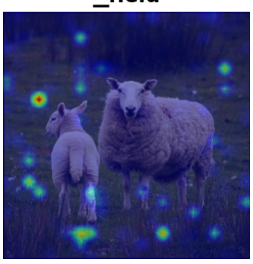
_a	_slice	_of	_p
izza	_in	_front	_of
_a	_man	.	

sheep.jpg

A sheep and a lamb are standing in a grassy field.

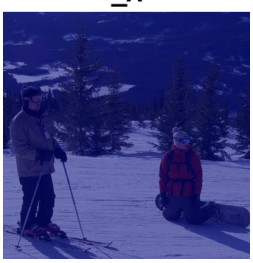
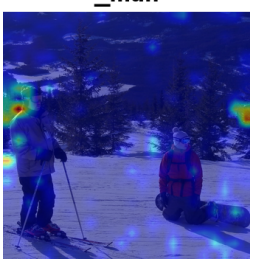
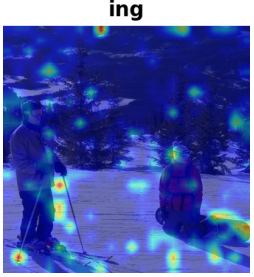
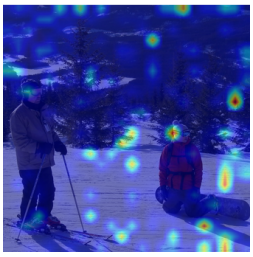
_A	_sheep	_and	_a
_l	amb	_are	_standing



_in	_a	_grass	y
			
_field	.		
			

ski.jpg

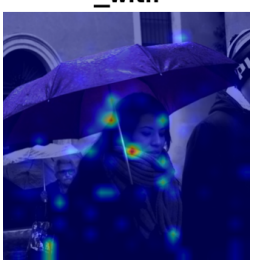
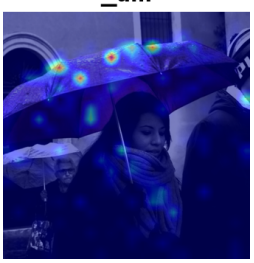
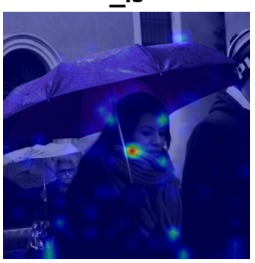
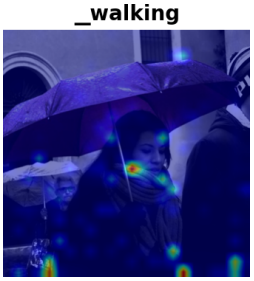
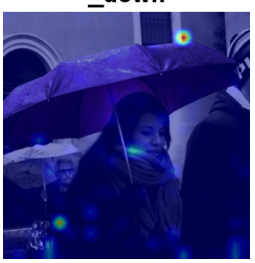
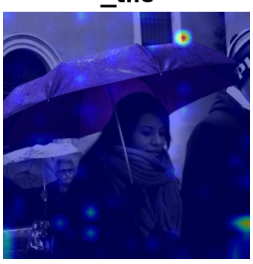
A man and woman are skiing on a snowy mountain.

_A	_man	_and	_woman
			
_are	_ski	ing	_on
			
_a	_snow	y	_mountain
			

. 			

umbrella.jpg

A woman with a purple umbrella is walking down the street.

_A 	_woman 	_with 	_a 
_pur 	ple 	_um 	bre 
lla 	_is 	_walking 	_down 
_the 	_street 	. 	

Problem 2

bike.jpg

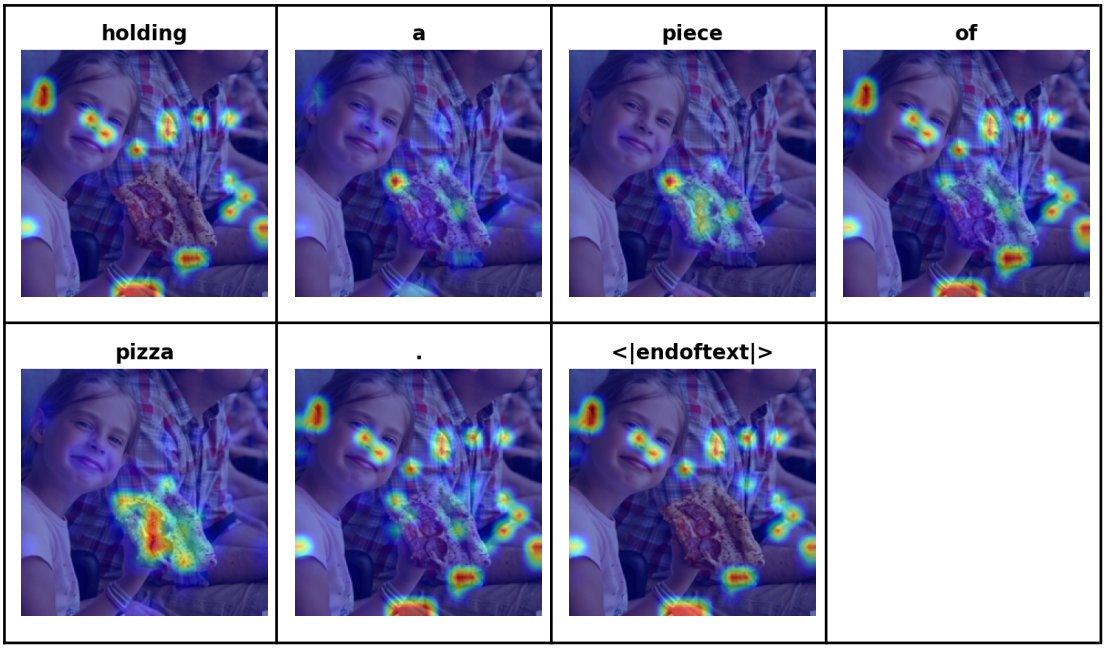
A woman riding a bike down a street with an umbrella.

<|endoftext|>	A	woman	riding
			
a	**bike**	**down**	**a**
			
street	**with**	**an**	**umbrella**
			
.	**<|endoftext|>**		
			

girl.jpg

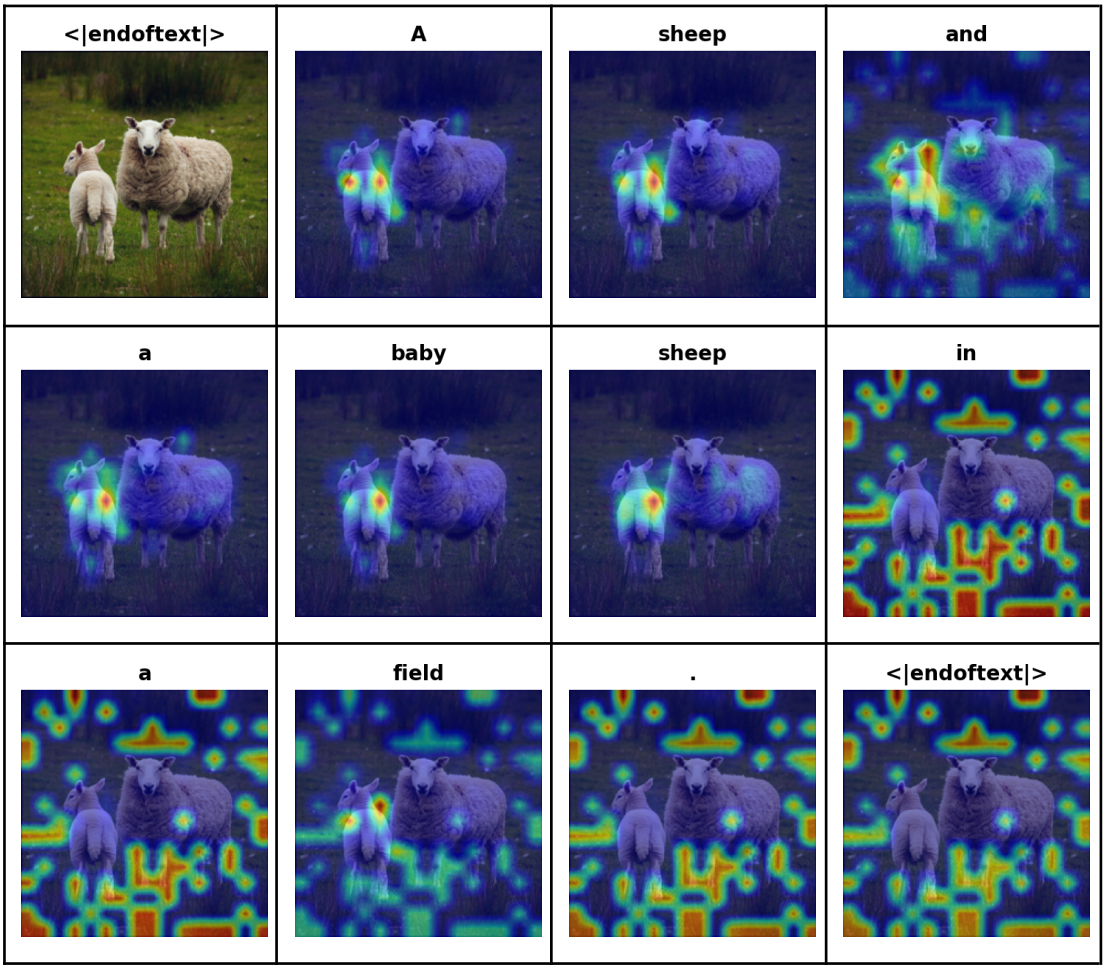
A little girl holding a piece of pizza.

<|endoftext|>	A	little	girl
			



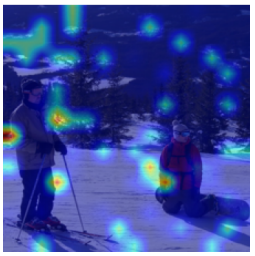
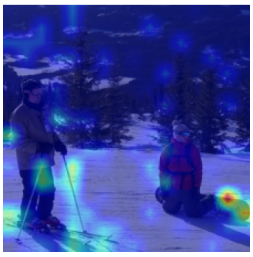
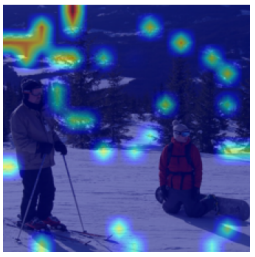
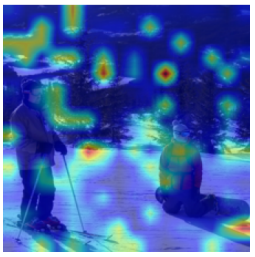
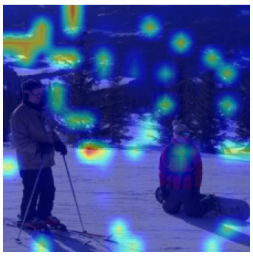
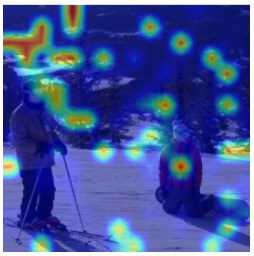
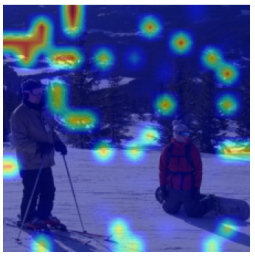
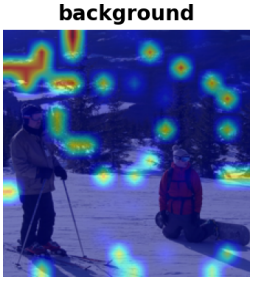
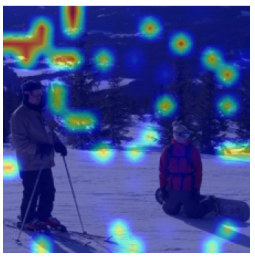
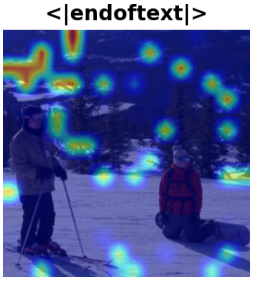
sheep.jpg

A sheep and a baby sheep in a field.



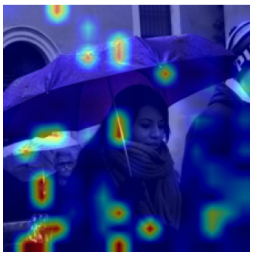
ski.jpg

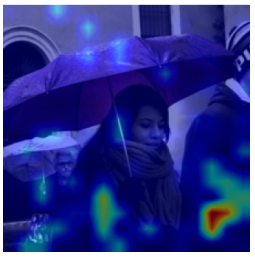
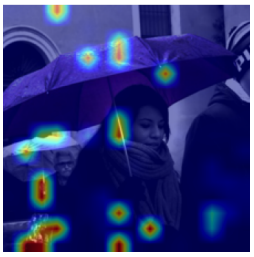
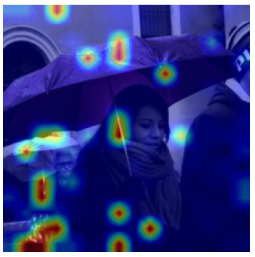
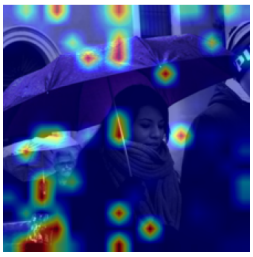
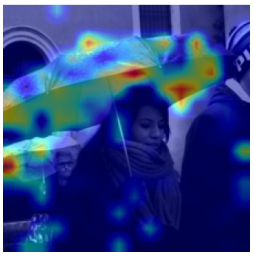
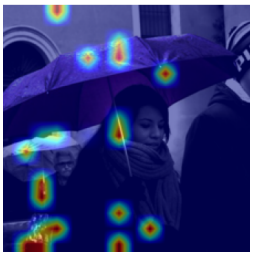
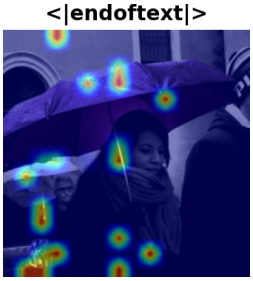
Two people on skis in the snow with trees in the background.

<|endoftext|>	Two	people	on
			
skis	in	the	snow
			
with	trees	in	the
			
background	.	<|endoftext|>	
			

umbrella.jpg

A couple of people standing in the rain with umbrellas.

<|endoftext|>	A	couple	of
			

people 	standing 	in 	the 
rain 	with 	umbrellas 	. 
<|endoftext|> 			

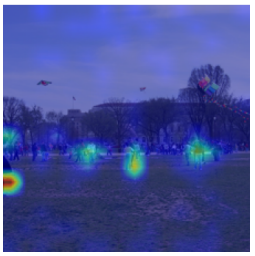
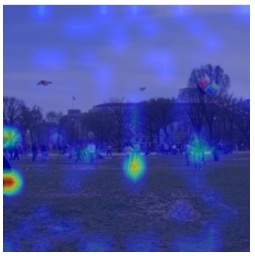
2. **According to CLIPScore, you need to: visualize top-1 and last-1 image-caption pairs and report its corresponding CLIPScore in the validation dataset of problem 2. (3%)**

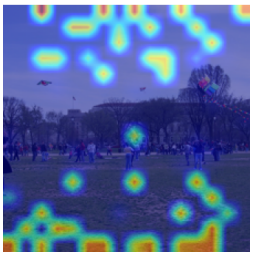
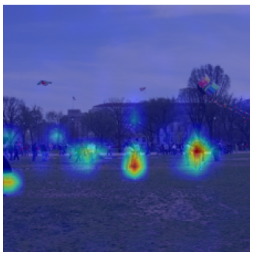
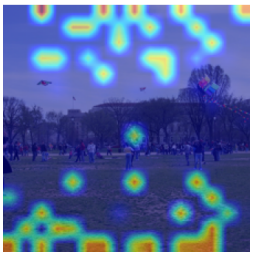
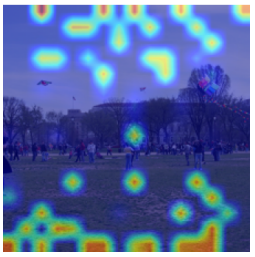
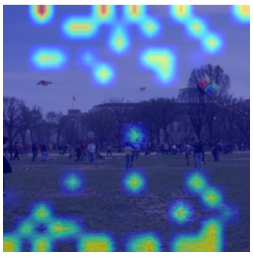
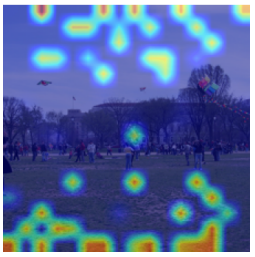
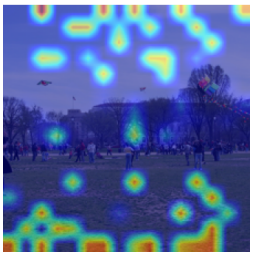
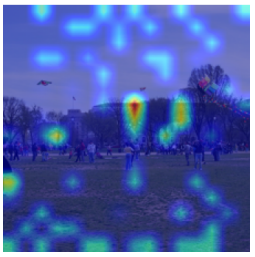
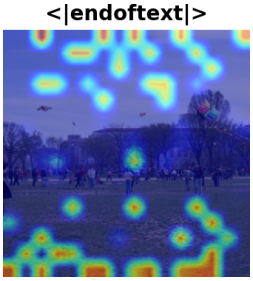
Top-1

000000001086.jpg

A large group of people flying kites in a park.

CLIPScore : 1.06201171875

<|endoftext|> 	A 	large 	group 

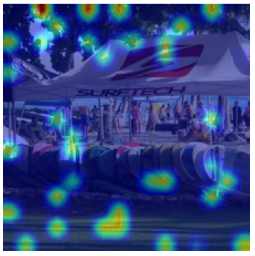
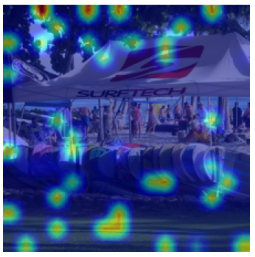
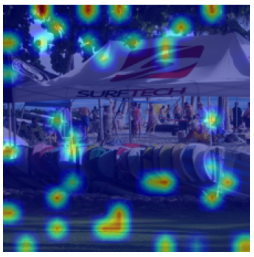
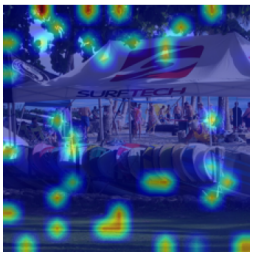
of 	people 	flying 	kites 
in 	a 	park 	. 
<|endoftext|> 			

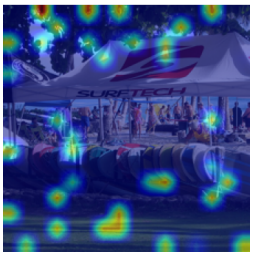
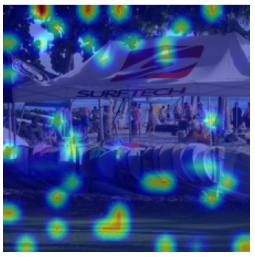
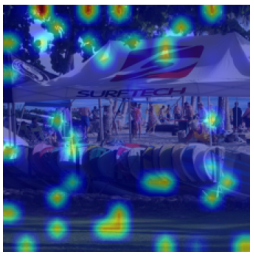
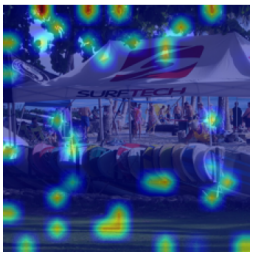
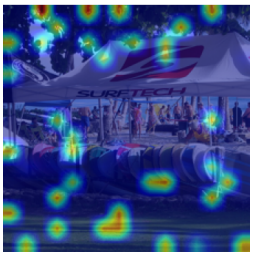
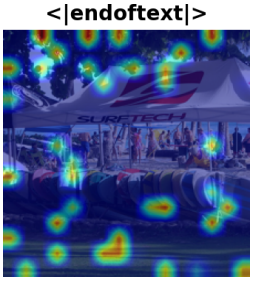
Last-1

000000001353.jpg

A group of people standing on top of a green field.

CLIPScore : 0.42144775390625

<|endoftext|> 	A 	group 	of 
people 	standing 	on 	top 

of	a	green	field
			
.	<|endoftext|>		
			

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? (3%)**
   Top-1的predicted caption合理；Last-1的則不合理。
   Top-1的attended region較符合corresponding word (people, kites等)，但我認為效果也沒有很好；Last-1沒有明顯符合的樣子。

Reference

1. claude
   https://claude.ai/new
2. chatgpt
   https://chatgpt.com/
3. https://huggingface.co/llava-hf/llava-1.5-7b-hf
4. https://huggingface.co/spaces/timm/leaderboard
5. https://huggingface.co/timm/vit_large_patch14_clip_224.openai_ft_in1k