HW3

Done	□
Due date	@November 21, 2022
Subject	DLCV

Part 1

1. Methods analysis

由於 CLIP 的訓練目標是**把 text 跟 image 映射到同一個空間**，而 CLIP 在訓練時用了很多不同領域的文字圖片做訓練，所以在影像辨識中可以透過列出 **“This is a photo of { class }”**，找出關聯性最大的 text 找出照片的分類。

2. Prompt-text analysis

- “This is a photo of { object }”

  Acc : 67.67

- “This is a { object } image.”

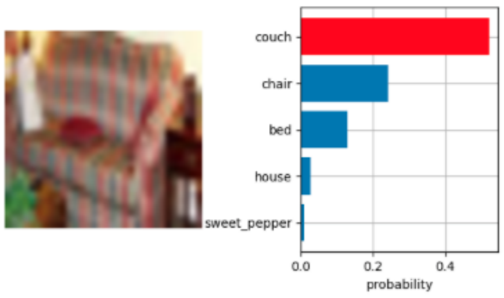
  Acc : 72.88

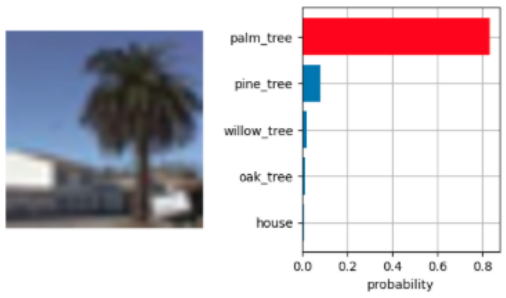
- “No { object }, no score.”

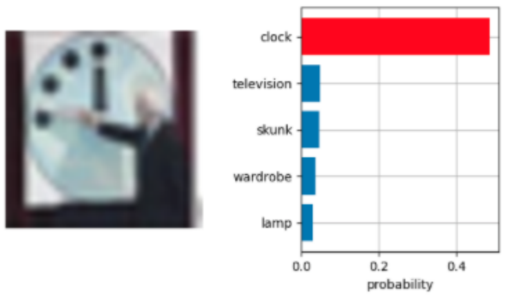
  Acc : 45.88

第三個 text 很明顯的在語意上跟第一二個 text 有極大的差距，所以導致 performance 有落差。

3. Quantitative analysis







Part 2

2. Report your best setting and its corresponding CIDEr & CLIPScore

使用 pretrained encoder with lr = 3e-5

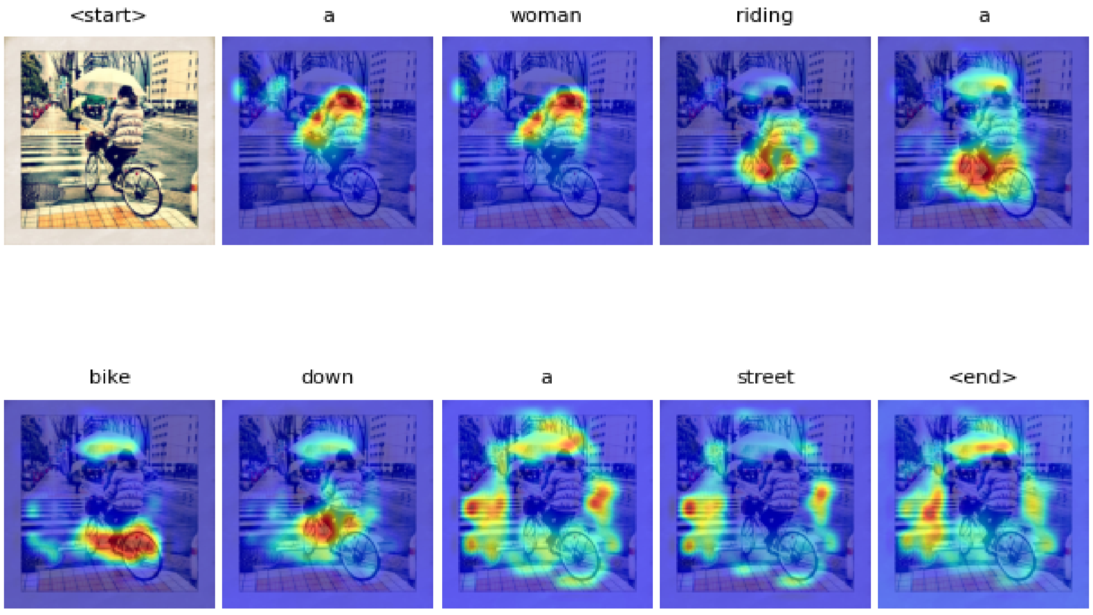
CIDEr: 0.7733517323228953 | CLIPScore: 0.6859251292075705

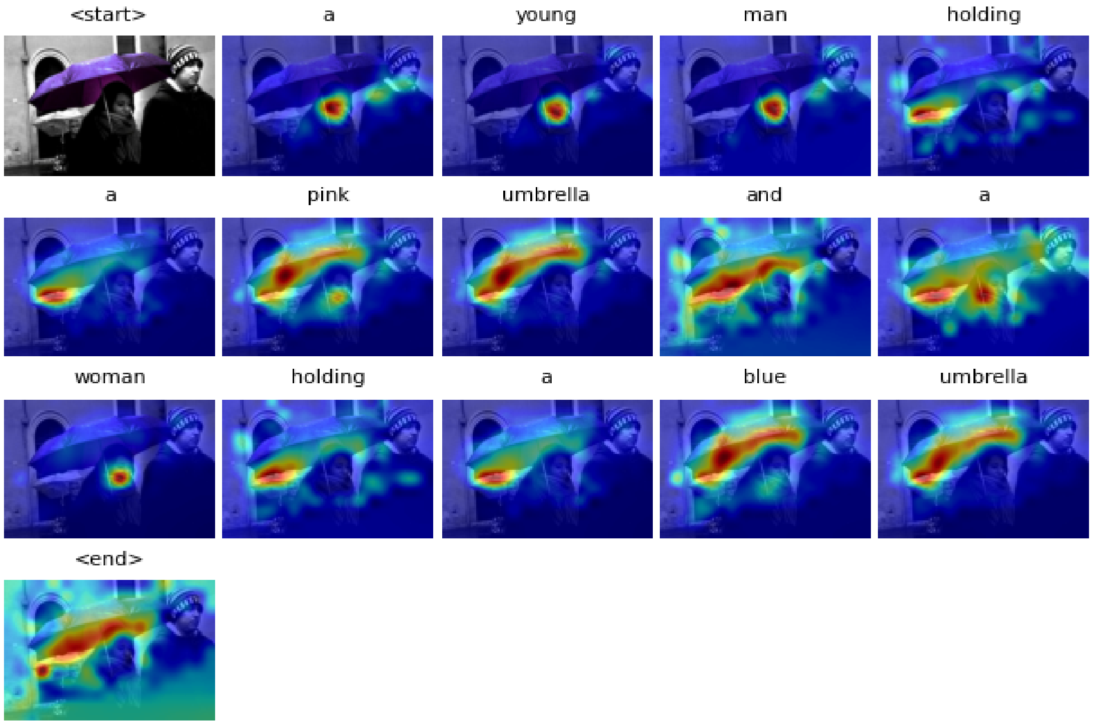
3. Report other 3 different attempts

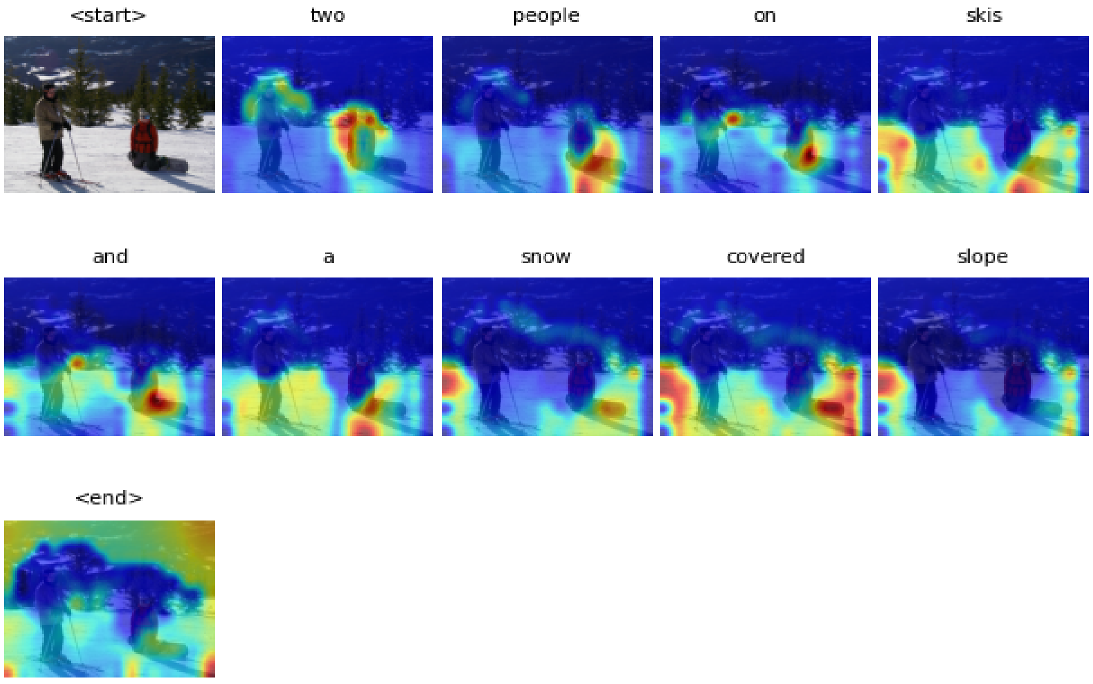
- 嘗試不使用 pretrained encoder ：**CIDEr: 0.1504148431717719 | CLIPScore: 0.4706820407481484**
- 調整 lr 為 1e-4 : **CIDEr: 0.22528638752785649 | CLIPScore: 0.5120282415706662**
- 更換 encoder 為 pretrained maxxvit : **CIDEr: 0.6667963790890161 | CLIPScore: 0.6843195873109105**

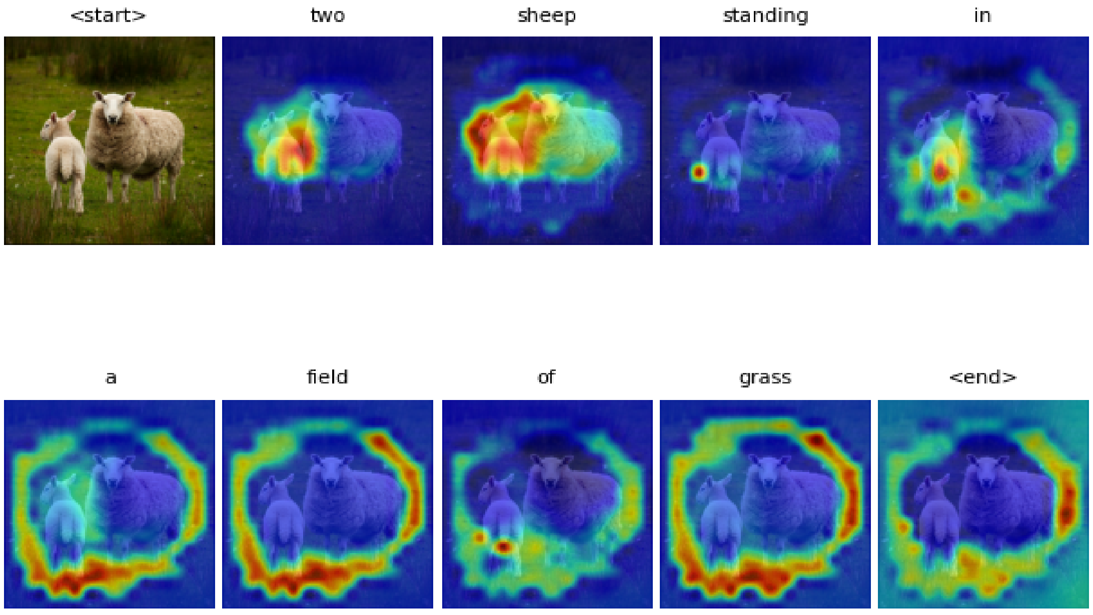
Part 3

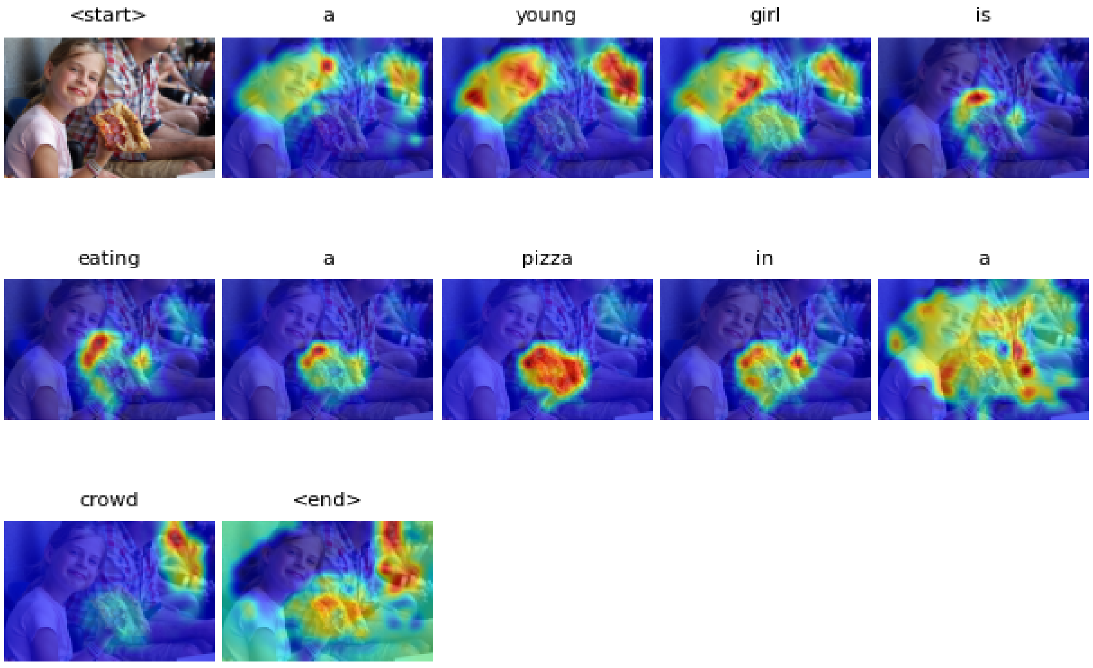
1. visualize the predicted caption and the corresponding series of attention maps


<start>
a
woman
riding
a
bike
down
a
street
<end>


<start>
a
young
man
holding
a
pink
umbrella
and
a
woman
holding
a
blue
umbrella
<end>


<start>
two
people
on
skis
and
a
snow
covered
slope
<end>



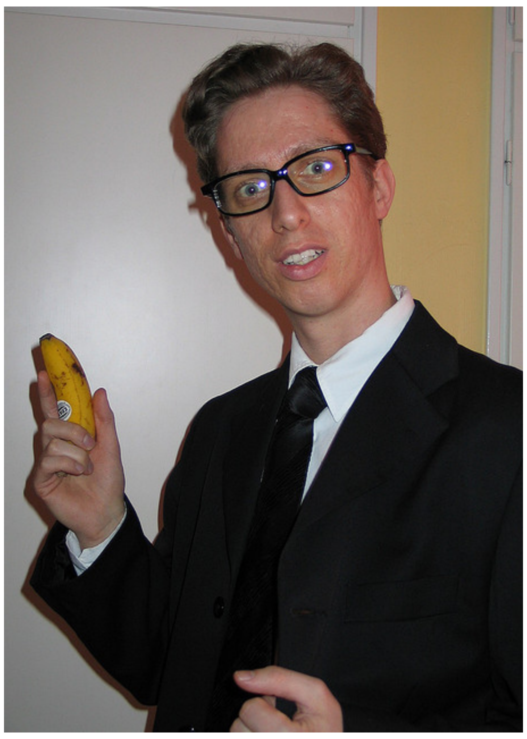


2. Visualize Top-1 and Last-1 image-caption pairs

Top 1

Caption: **A man wearing a suit and tie holding a banana .**

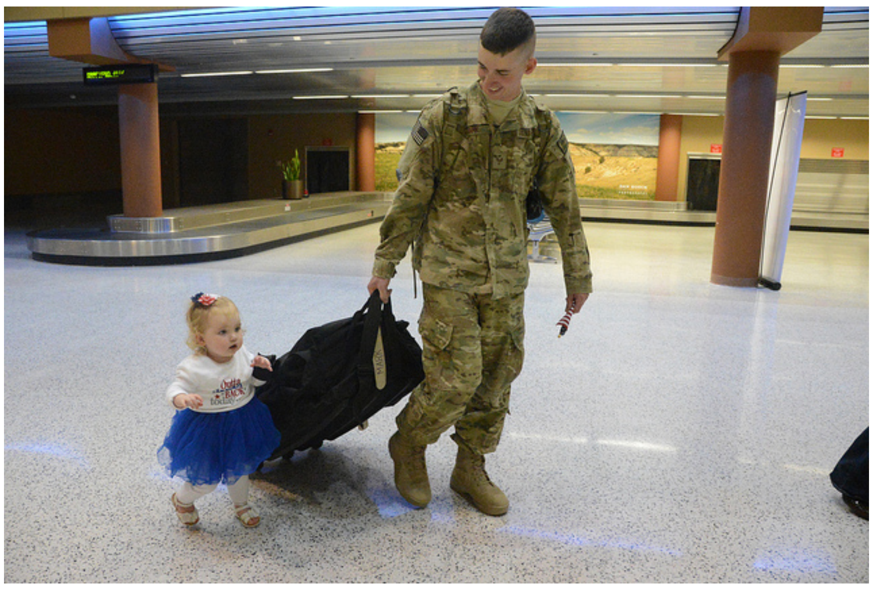
Score: **0.97900390625**



Last 1

Caption: **A man in a plaid shirt is walking through a workshop .**

Score: **0.34088134765625**



3. Analyze the predicted captions and the attention maps for each word according to the previous question.

Top 1 的圖可以看到模型成功的識別穿西裝的人跟手上握著的香蕉，但 Last 1 嘗試描述一個人走進某地，但無法識別出這是機場，也忽略了旁邊同樣是主體的小女孩。