DLCV Hw-3

Name: 李維釗

Student ID: R12942089

Problem 1: Zero-shot Image Captioning with LLaVA

1. Please read the paper “Visual Instruction Tuning” and briefly describe the important components (modules or techniques) of LLaVA.
   - Instruction Tuning通過更明確的多模態或多任務指令來微調模型，旨在提升模型在不同任務下的泛化能力，特別是能實現zero-shot或few-shot。與Prompt Tuning相比，Instruction Tuning 注重通用性和跨任務能力的提升。
   - 為了解決缺少詳細數據的問題，作者利用了chatGPT等已經訓練好的模型，透過給予更多圖像的資訊（從不同角度描述視覺場景、邊界框等）來拓展LLM給予的詳細資訊。
   - 而在模型的部分則是整合了Vision以及LLM兩個單模態的大模型來對齊兩個模態之間的embedding。
   - 透過這樣的思路，或許可以應用到各種跨模態的大模型上，以整合不同模型。
2. Please come up with two settings (different instructions or generation config). Compare and discuss their performances.

	CIDEr	CLIPScore
Setting 1	1.1529808500176097	0.77809814453125
Setting 2	0.7197498966877777	0.7927020263671875

   1. Setting 1

      `USER: <image>\nBriefly describe this image in a sentence. Thank you. ASSISTANT:`

   2. Setting 2

      `USER: <image>\nIn one sentence, describe the key objects in the image and their state. For example, 'A red apple lies on a wooden table' or 'A dog sleeps peacefully on a couch.' ASSISTANT:`

      雖然setting2中含有一些instruction來引導model回答的方式，會影響到生成結果的CIDEr；但可以看到的是CLIPScore則有不錯的表現。說明了有更多的Instruction可以引導生成caption的準確度；但相反的太少樣的instruction則會限制caption的格式，讓CIDEr的分數因此下降

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. Briefly introduce your method. (TA will reproduce this result)

   - Optimizer: Adam, Lr=5e-5, lora rank=32, training epoch=10
   - Lora：只有在transformer中的attention以及mlp內的linear layers使用。

```
self.c_attn = lora.Linear(cfg.n_embd, 3 * cfg.n_embd, cfg.rank)
self.c_proj = lora.Linear(cfg.n_embd, cfg.n_embd, cfg.rank)
```

```
self.mlp = nn.Sequential(collections.OrderedDict([
    ('c_fc', lora.Linear(cfg.n_embd, 4 * cfg.n_embd, cfg.rank)),
    ('act', nn.GELU(approximate='tanh')),
    ('c_proj', lora.Linear(4 * cfg.n_embd, cfg.n_embd, cfg.rank))
]))
```

   - Vision encoder使用的是timm的`vit_large_patch14_clip_224`並且使用同款的transform。
     而壓好的vision embedding則用一個trainable的linear layer來做to text embedding的映射。
   - 轉換過後的vision embedding會先與text embedding concatenate，然後再加上positional embedding。

	CIDEr	CLIPScore
Best	0.8844035428513431	0.6921499633789062

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

   a. Setting 1: LoRA rank = 32, training epochs = 10, beam search = 5
   model架構同best setting

   b. Setting 2: LoRA rank = 32, training epochs = 5, beam search = 5
   model架構同best setting

   c. Setting 3: LoRA rank = 64, training epochs = 3, beam search = 5
   model架構同best setting

d. Setting 4: LoRA rank = 32, training epochs = 7, beam search = 5
與setting 1, 2, 3相比，將attention linear改成MergedLinear，並多加了lm_head的lora。
vision embedding排除cls後才送入model中。

```
self.c_attn = lora.MergedLinear(cfg.n_embd, 3 * cfg.n_embd, r=cfg.rank,
                                enable_lora=[True, False, True],
                                # lora_alpha=cfg.lora_attn_alpha,
                                lora_dropout=cfg.lora_dropout,)
```

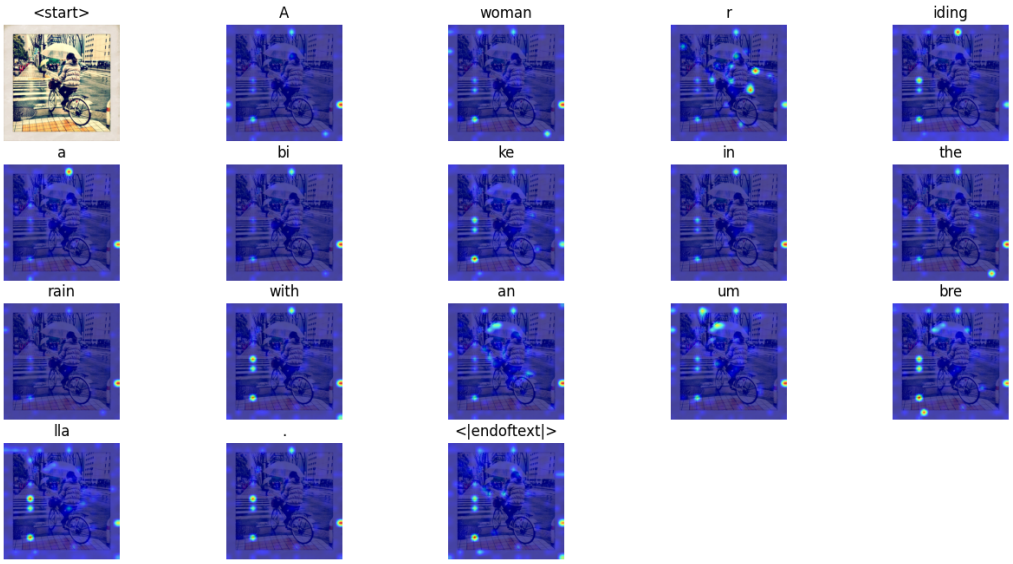
e. Setting 5: LoRA rank = 32, training epochs = 13, beam search = 5
model架構與setting 3相同

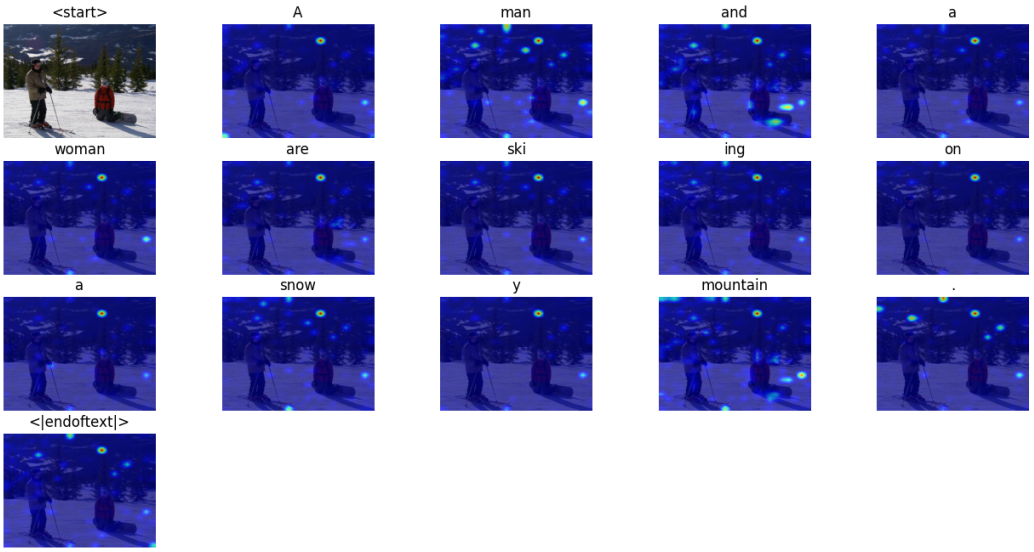
	CIDEr	CLIPScore
Setting 1	0.8844035428513431	0.6921499633789062
Setting 2	0.8513301122849635	0.6806594848632812
Setting 3	0.8654818424667257	0.6828769683837891
Setting 4	0.7765376105405314	0.678280029296875
Setting 5	0.792413225587337	0.6929396057128906

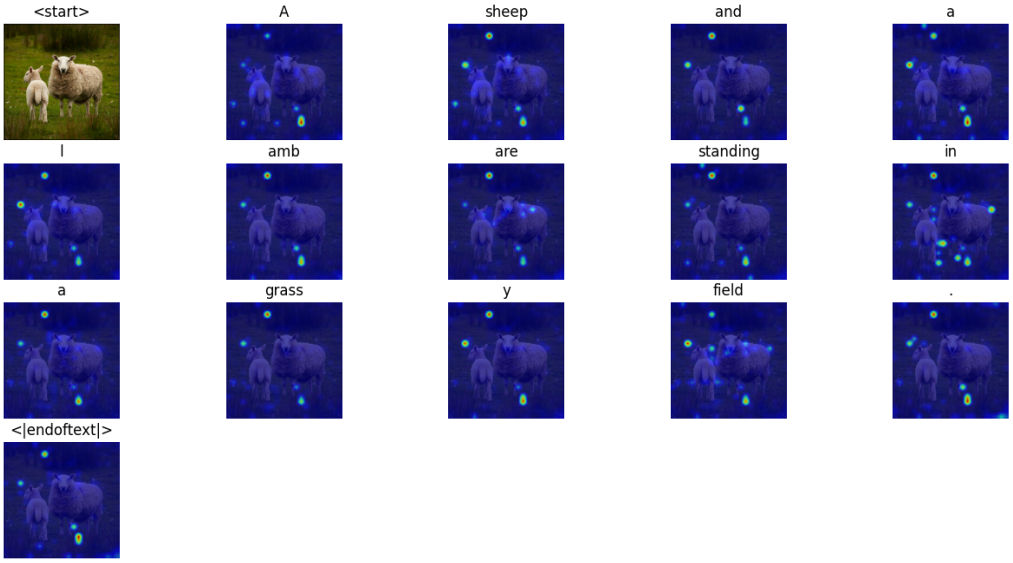
Problem 3: Visualization of Attention in Image Captioning

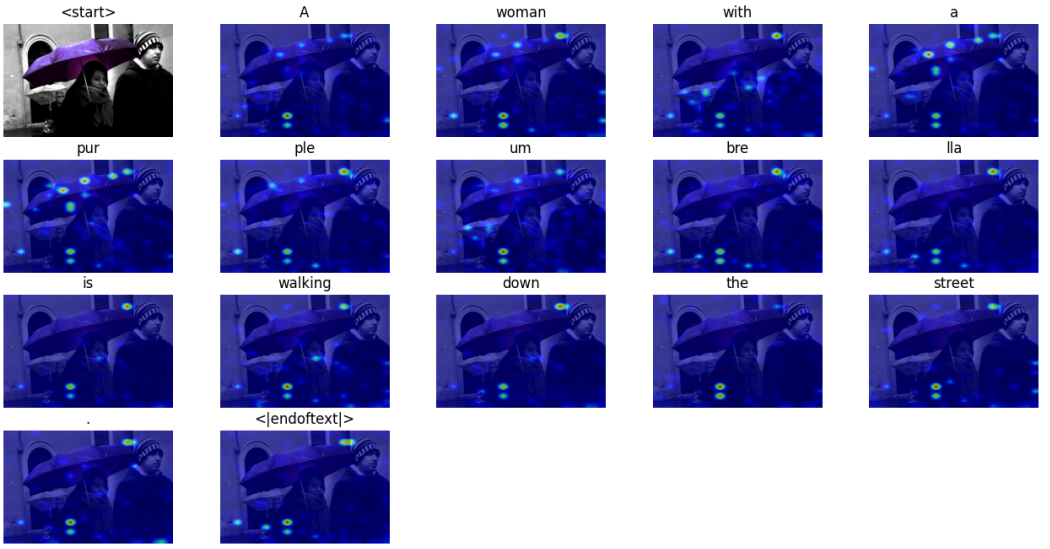
1. Given five test images `([p3_data/images/])`, and please visualize the predicted caption and the corresponding series of attention maps in your report with the following template:
(20%, each image for 2%, you need to visualize 5 images for both problem 1 & 2)

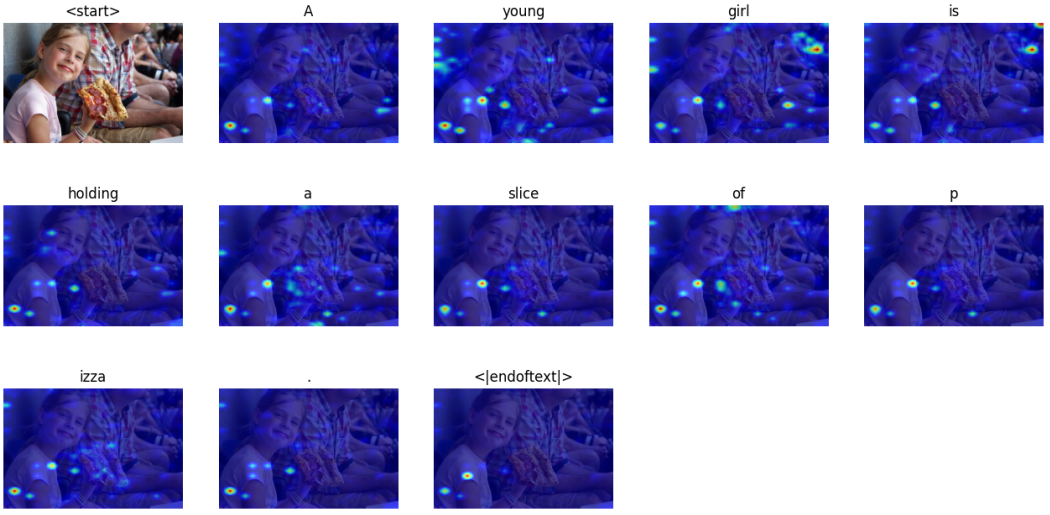
   a. P1 (對多頭注意力使用max pooling)


<start>
A
woman
r
iding
a
bi
ke
in
the
rain
with
an
um
bre
lla
.
<|endoftext|>

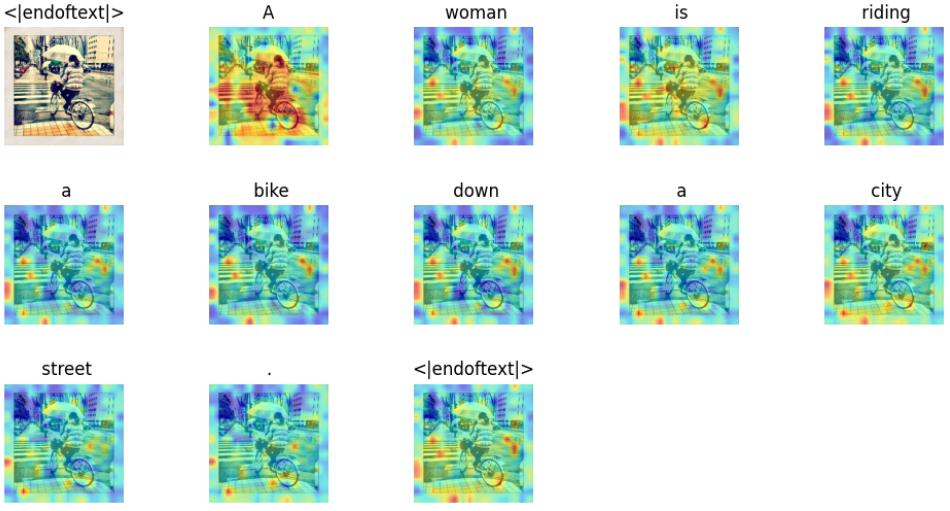

<start>
A
man
and
a
woman
are
ski
ing
on
a
snow
y
mountain
.
<|endoftext|>

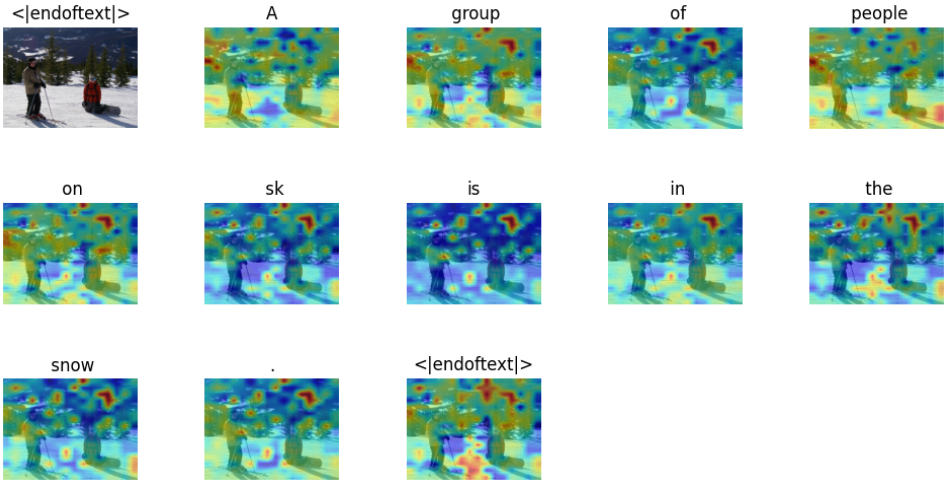

<start>
A
sheep
and
a
l
amb
are
standing
in
a
grass
y
field
.
<|endoftext|>

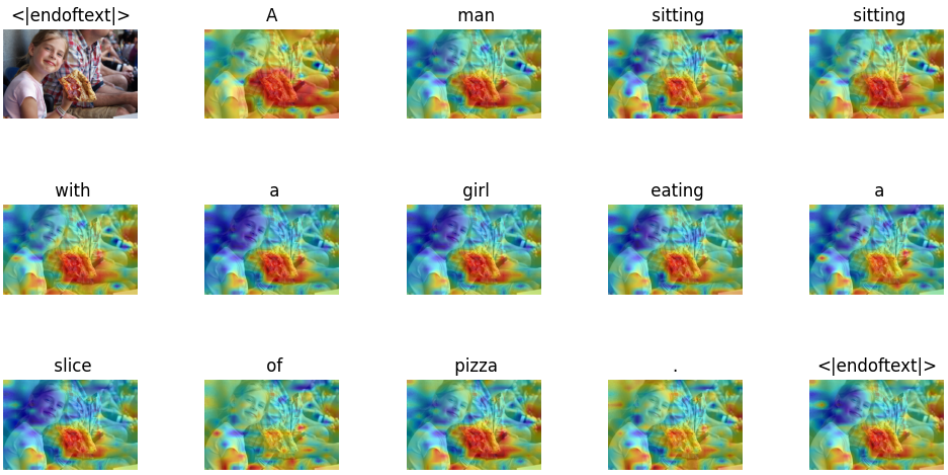

<start>
A
woman
with
a
pur
ple
um
bre
lla
is
walking
down
the
street
.
<|endoftext|>


<start>
A
young
girl
is
holding
a
slice
of
p
izza
.
<|endoftext|>

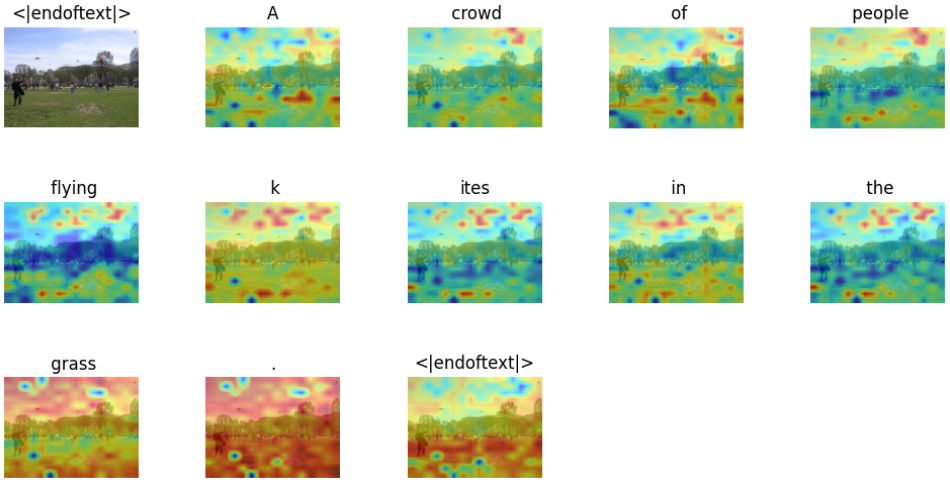
b. P2


<|endoftext|>
A
woman
is
riding
a
bike
down
a
city
street
.
<|endoftext|>

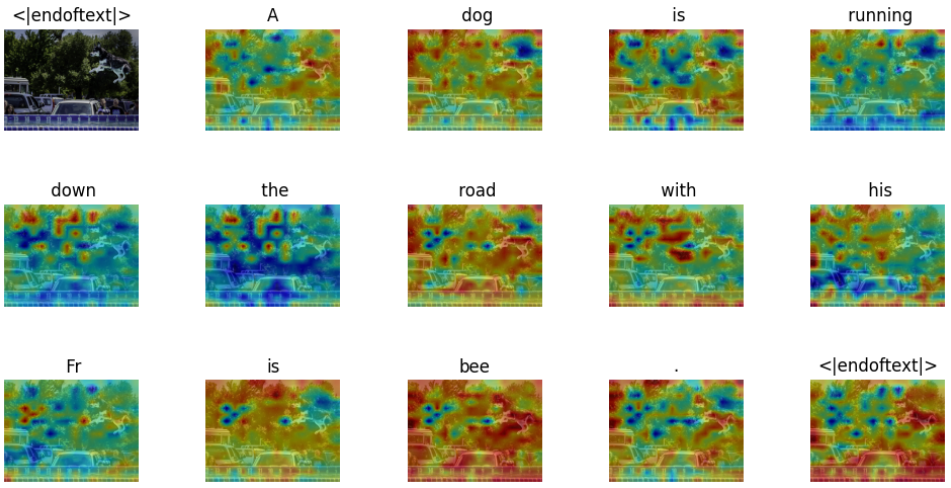

<|endoftext|>
A
group
of
people
on
sk
is
in
the
snow
.
<|endoftext|>
<|endoftext|>
A
couple
of
sheep
grazing
on
a
dry
grass
.
<|endoftext|>
<|endoftext|>
A
woman
is
holding
an
umbrella
in
the
rain
.
<|endoftext|>



2. According to **CLIPScore**, you need to: (in the validation dataset of problem 2.)
    a. visualize top-1 and last-1 image-caption pairs
        i. top-1



        ii. last-1



b. report its corresponding CLIPScore

	CLIPScore
top-1	1.002197265625
last-1	0.32196044921875

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%)

- 除了Girl的照片中model更傾向去認得後面的男人，並沒有敘述出正確的女孩描述，但吃披薩這部分倒是有正確；其餘caption效果都不錯。
- 例如在天空的風箏以及pizza對齊的不錯。
  但因為我並沒有針對多頭注意力來實現attention map因此會有些部分與實際看起來並不相符。