DLCV HW1 Report

B10901184 錡亭勳

Problem 1: Zero-shot Image Captioning with LLaVA

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

Ans:

The main components of LLaVA is a ViT and a LLM. It proposed a new method to gather image-text pair instruction-tuning data from GPT-4, which is a text only model. The model design is simple, but due to the large training dataset gathered using its proposed method, LLaVA achieved SOTA.

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

Ans:

Setting 1:

max_new_token: 60, min_length: 15, num_beams: 3

Result: CIDEr: 1.1441 | CLIPScore: 0.7884

Setting 2:

max_new_token: 80, min_length: 45, num_beams: 3

Result: CIDEr: 1.0936 | CLIPScore: 0.7946

Comparison:

Setting longer output length results in higher CLIPScore, since the model is able to talk more about the image details. However, it also results in lower CIDer since the frequency of important words

decreases as the model talks longer. Therefore, the choice of maximum output length depends on specific needs.

Discussion:
The difference in quality scores between two settings also show the problem of trying to use one metric to evaluate the performance of a model. If we only focus on CIDer, we will lose the opportunity to make the model dive into the details of the images. If we only focus on CLIPScore, the model might speak about some incorrect or irrelevant things.

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) (5%)**

Ans:
Setting:
ViT: vit_large_patch14_clip_224.laion2b,
Rank: 16,
Alpha: 8,
Epochs: 10,
Optimizer: Adam,
Scheduler: CosineAnnealing(T_0=EPOCHS, T_mult=2, eta_min=1e−6)

Result: CIDer: 0.9513 | CLIPScore: 0.7345

Method: The most important thing, after discussion with the TA, is the choice of pretrained ViT. The rest is just setting everything from “nn.” to “lora.” I trained the LoRA with teacher–forcing method since it lets the model learn faster, but it is also prone to overfitting.

---

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

Ans:

Rank = 16, Alpha = 8	Rank = 32, Alpha = 1
CIDer: 0.9513 | CLIPScore: 0.7345	CIDer: 0.9279 | CLIPScore: 0.7273

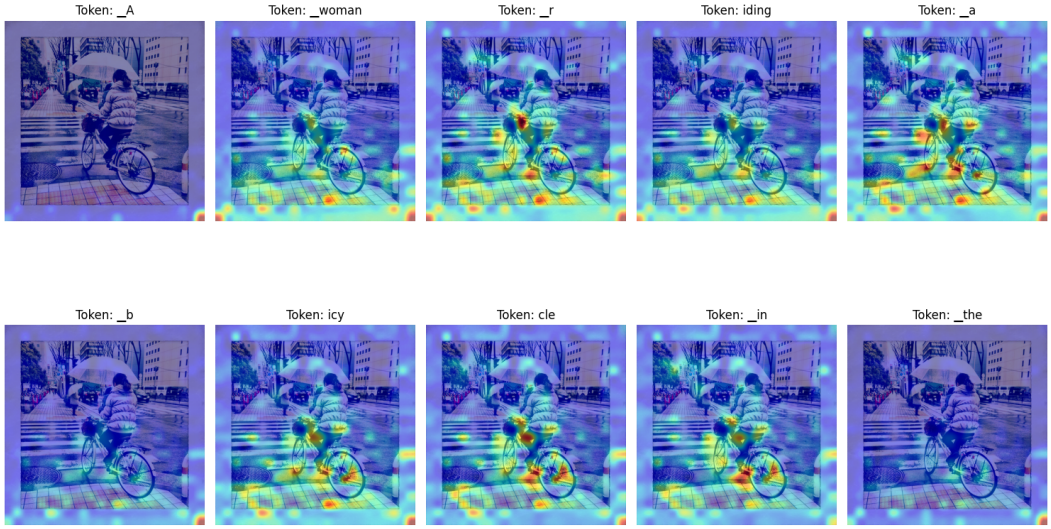
Problem 3: Personalization

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)

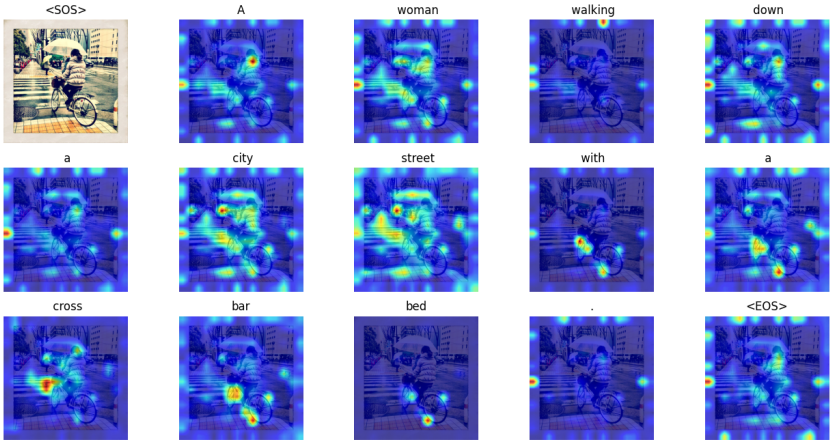
Ans: (next page)

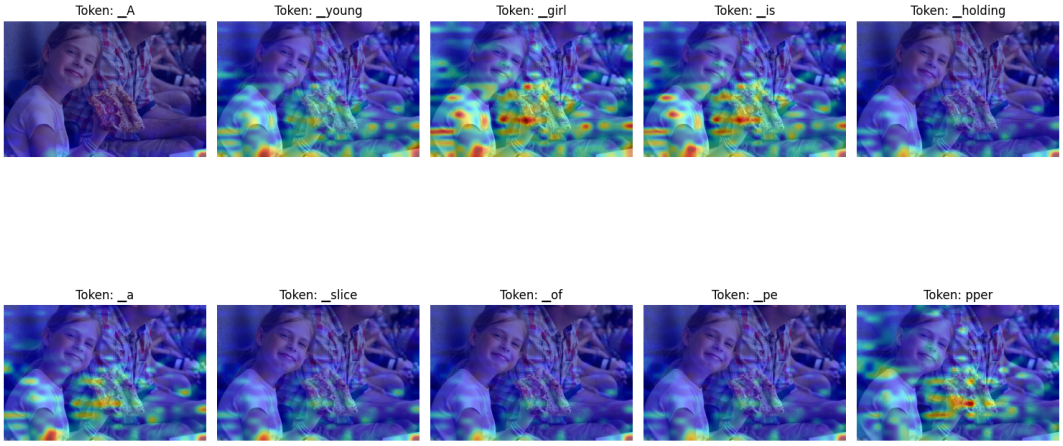
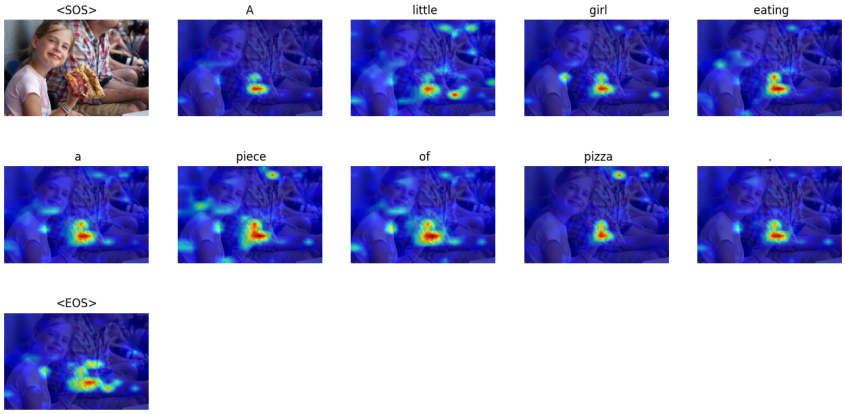
Bike.png

LLaVA



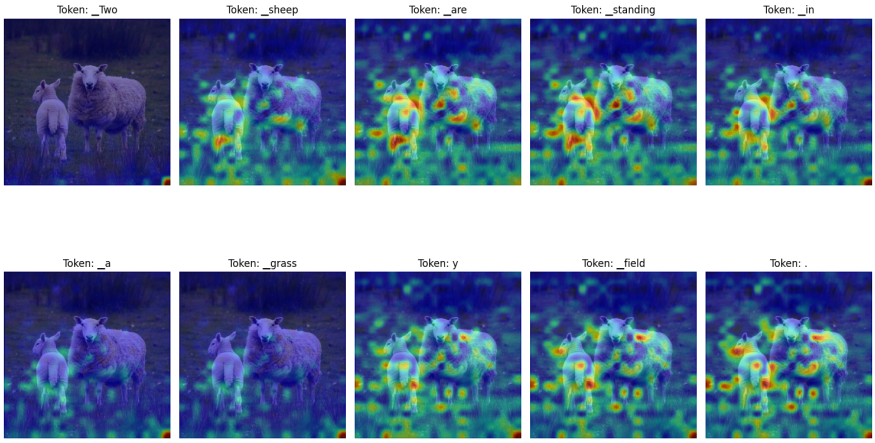
Mine



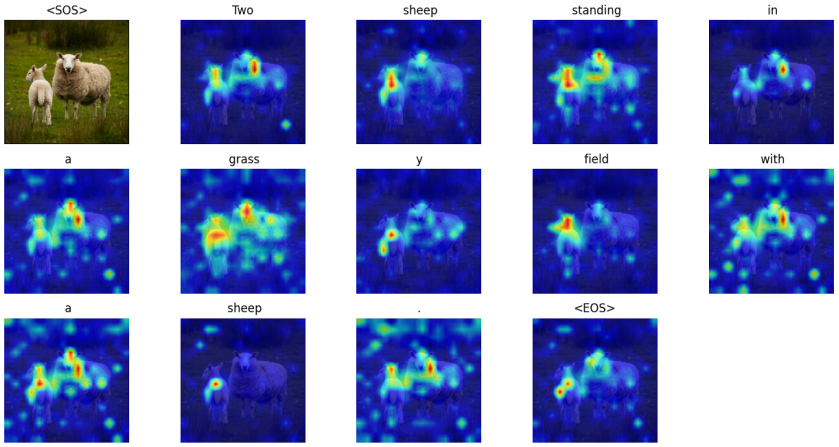
Girl.png	
LLaVA	
Mine	

Sheep.png

LLaVA

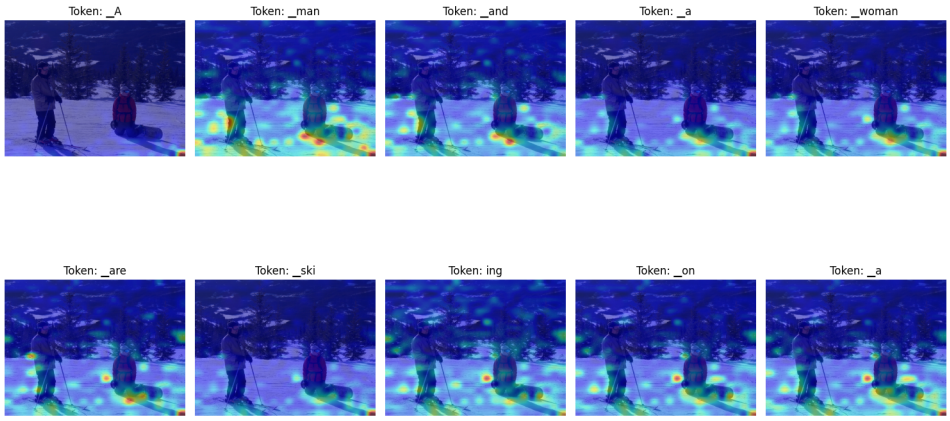


Mine

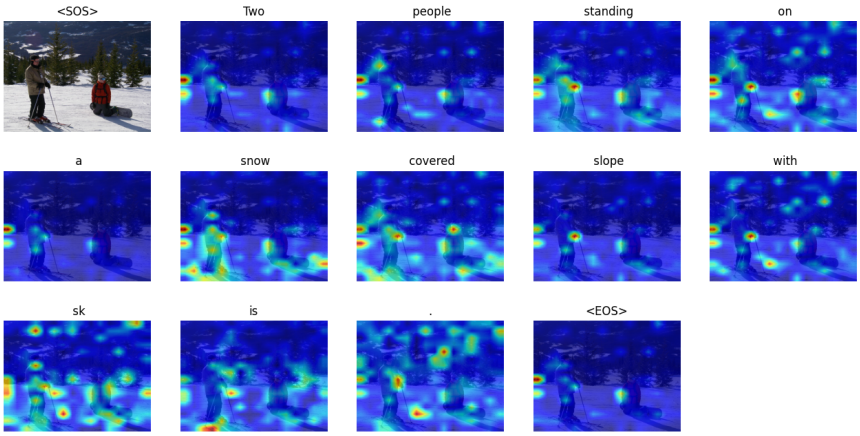


Ski.png

LLaVA

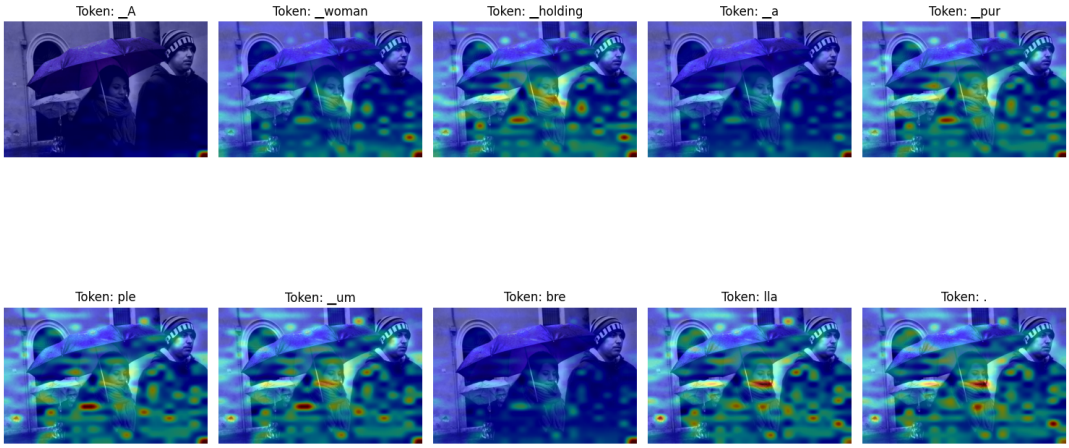


Mine

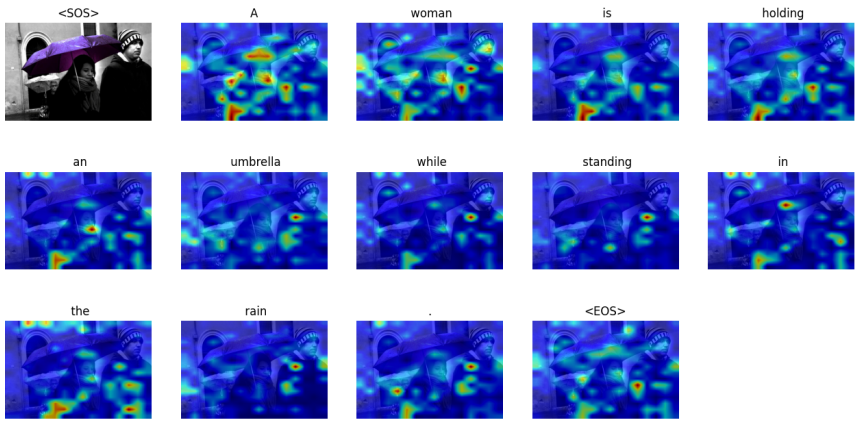


Umbrella.png

LLaVA



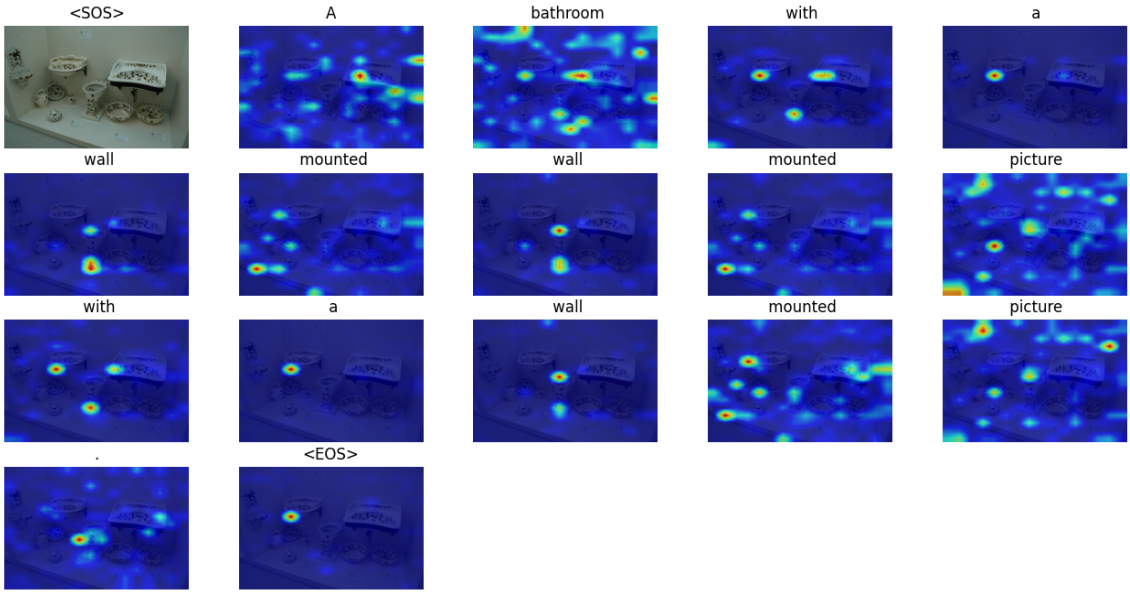
Mine



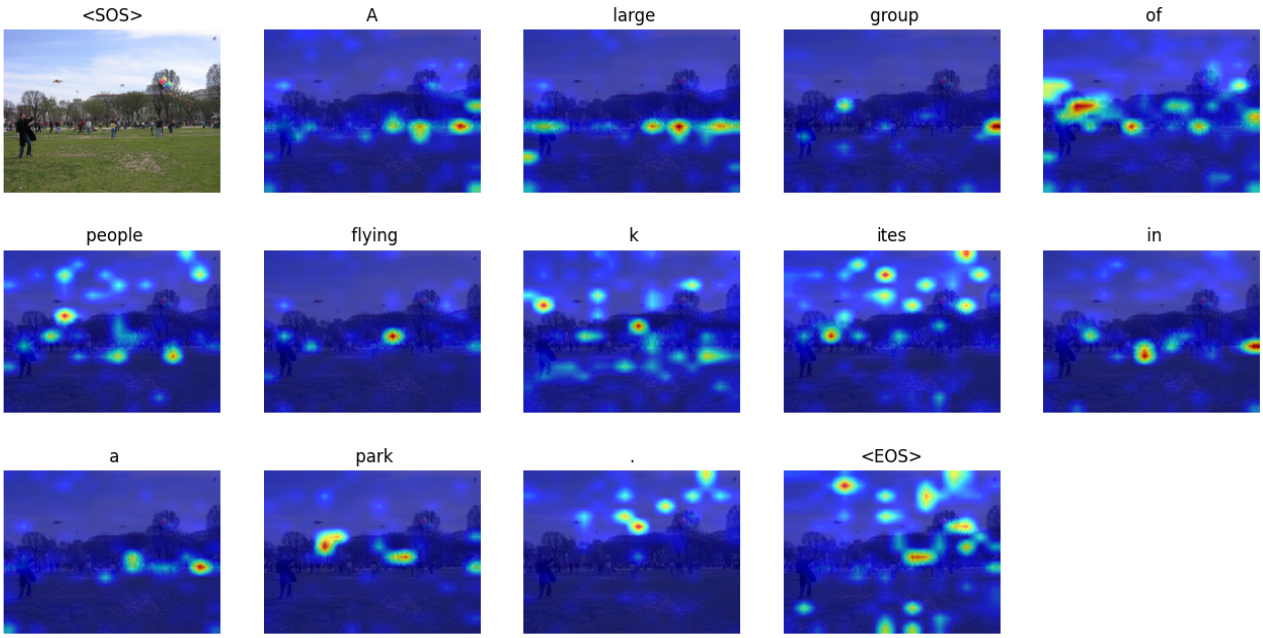
2. According to CLIPScore, you need to: 1. visualize top-1 and last-1 image-caption pairs. 2. report its corresponding CLIPScore in the validation dataset of problem 2. (3%)

Ans:

Worst Image: 000000000980, Worst Caption: A bathroom with a wall mounted wall mounted picture with a wall mounted picture., Worst Score: 0.40008544921875



Best Image: 000000001086, Best Caption: A large group of people flying kites in a park., Best Score: 1.06201171875



2. 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%)

Ans:

LLaVA: The attention maps do not show significant relation between the caption tokens and the images. I think this is due to the the multihead attention mechanism, the job of each head is not fully separated, so even if we average each head, or choose specific head, the overall attention map still does not show much information. However, we can still notice that the attention is mainly on the desired subjects. For example, the Girl.png shows that LLaVA focuses on the pizza the most.

Mine: The attention maps, which are the average of the second attention layer, show higher correlation with the “human interpretable” attention maps. For example, the “Best Image” in the previous question, the attention maps show strongest attention on the kites when the token is “k ite”, and it focuses on people when it says “a large group of people”, which is the main subject of the image.