DLCV

Homework 2

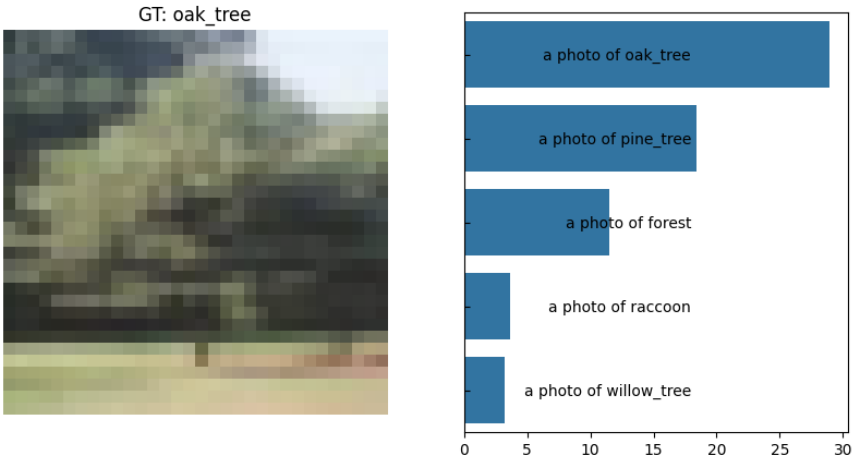
R12922147 魏子翔

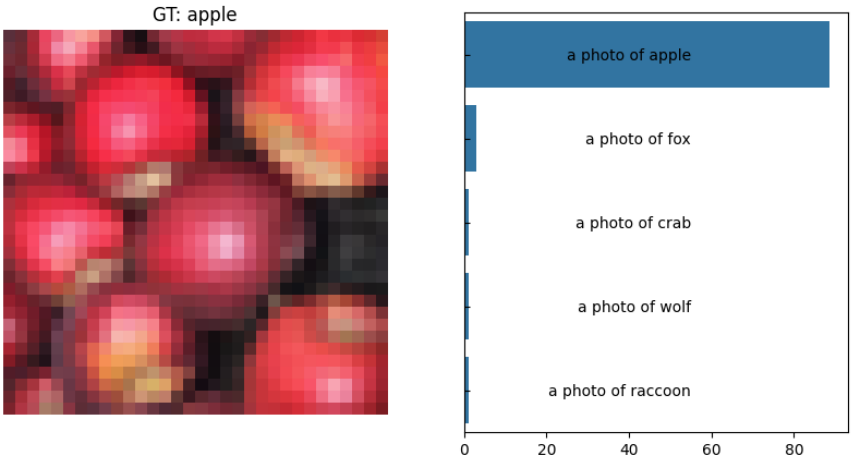
Problem 1.

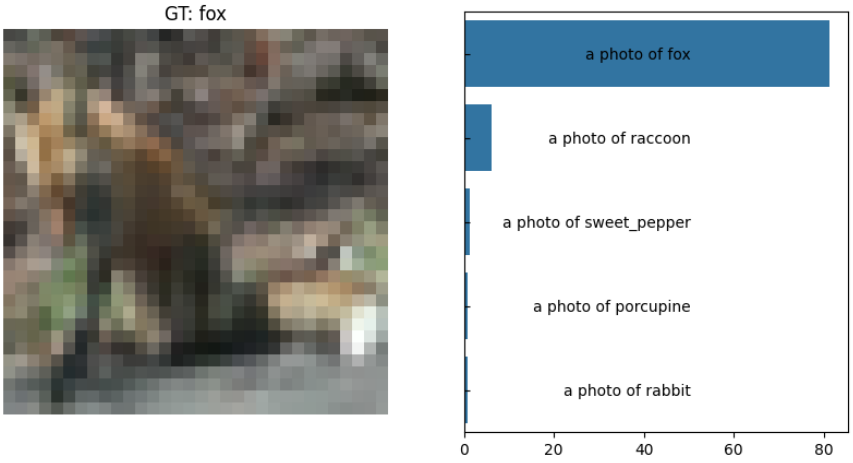
1. Methods analysis
   - Model Architecture: Clip use transformer architecture, compare to VGG and ResNet, it is more robust.
   - Alignment of Text and Images: Clip learns to understand the relationship between images and diverse textual descriptions from the internet. This approach helps CLIP develop a more generalized understanding of visual concepts across various domains.
   - Scale and Data Diversity: CLIP was trained on a vast and diverse dataset containing billions of image-text pairs from the internet. This scale and diversity contribute to its ability to capture a broad range of visual concepts and linguistic nuances.
2. Prompt-text analysis

prompts	accuracy
This is a photo of {object}	0.6784
This is not a photo of {object}	0.6952
No {object}, no score.	0.4524

   Surprisingly, the prompt "This is not a photo of {object}" gets the highest score among the three prompts. I think it is mainly due to the dataset bias. Since the dataset used to train CLIP is collected from web, without detailed data cleaning, there might be a lot of text is giving the negative statement of the corresponding image, which result in the model giving high similarity scores to those pairs.
3. Quantitative analysis







Problem 2.

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

   I choose LoRA with $r = 4$ as my best setting, and below is my training details:

   image encoder: vit_large_patch14_clip_224.openai

   image embedding dimension: 1024

   transforms: follow encoder transforms

   cross attention: follow self attention but match the projection for k and v to image embedding and without mask

   numbers of decoder layers: 12

   numbers of attention heads: 12

   lr = 3e-5

   weight_decay = 1e-5

   batch_size = 64

   epochs = 10

   optimizer: AdamW

lr scheduler: CosineAnnealingLR

decode strategy: beam search with beam=3

peft type: LoRA

rank of Lora: 4

	CIDEr	CLIPScore
score	0.9328	0.7208

2. Report 3 different attempts of PEFT and their corresponding CIDEr & CLIPScore.

   attempt 1: follow the above setting with peft type = LoRA and rank = 8.

   attempt 2: follow the above setting with peft type = Adapter and adapter rank = 64, add adapter blocks after self attention and cross attention for every transformer layers.

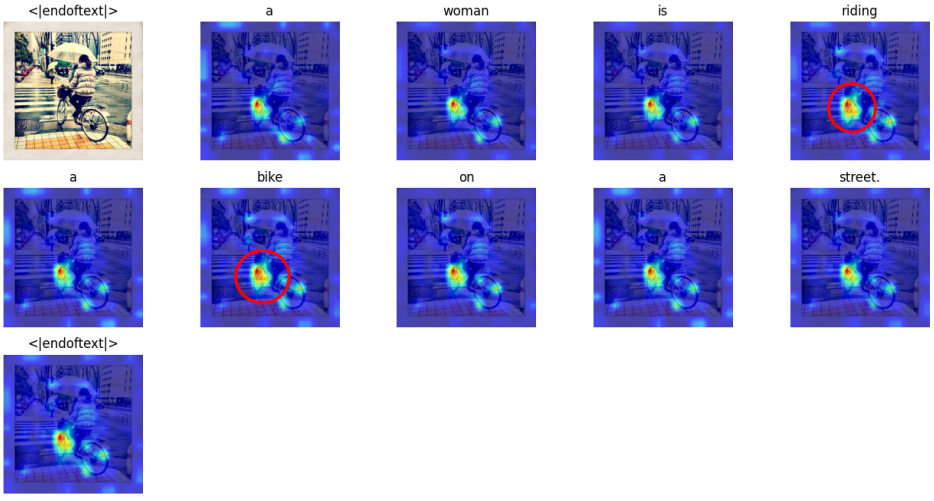
   attempt 3: follow the above setting with peft type = prefix tuning, add prefix length of 100.

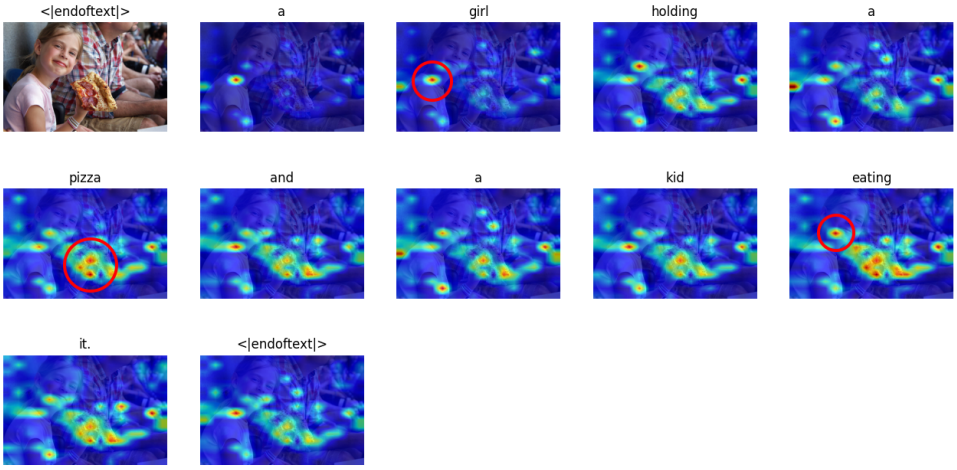
	CIDEr	CLIPScore
attempt 1 (lora, r=8)	0.9311	0.7181
attempt 2 (adapter)	0.9232	0.7203
attempt 3 (prefix tuning)	0.9303	0.7182

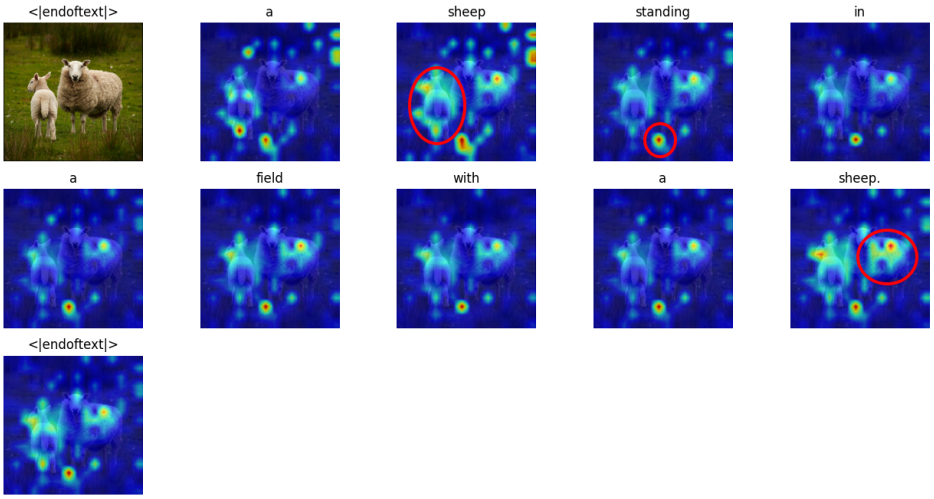
Problem 3.

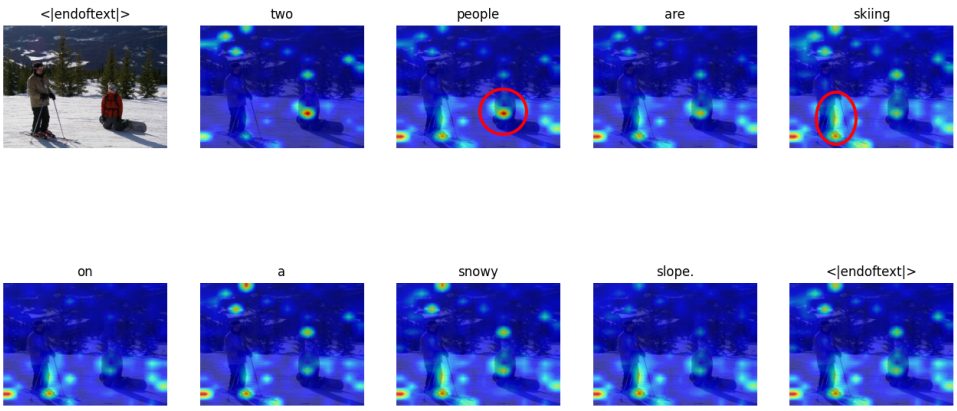
1. visualize the predicted caption and the corresponding series of attention maps.

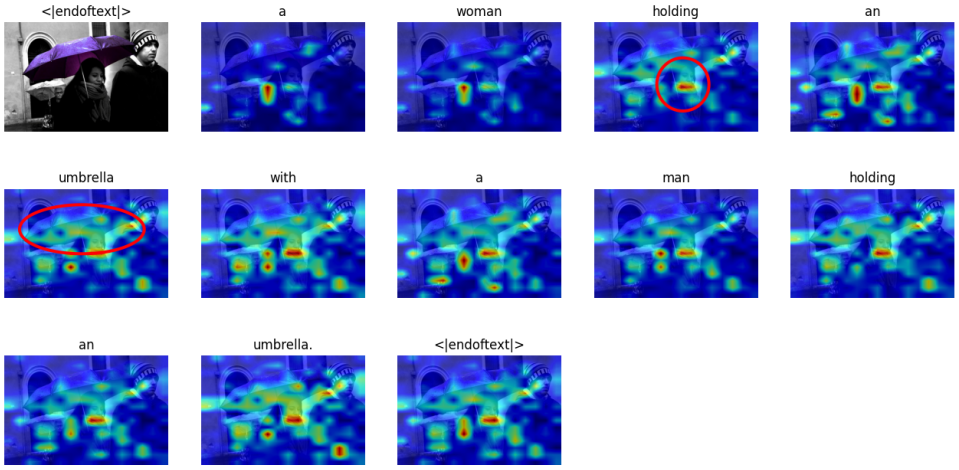
   The red circles are the most reasonable parts for each image and caption pair, which is for problem 3.




<|endoftext|>
a
girl
holding
a
pizza
and
a
kid
eating
it.
<|endoftext|>

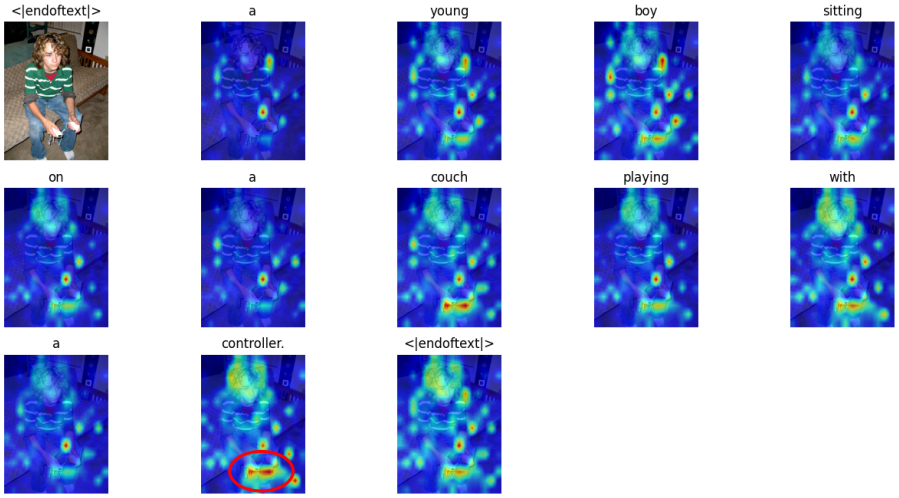

<|endoftext|>
a
sheep
standing
in
a
field
with
a
sheep.
<|endoftext|>


<|endoftext|>
two
people
are
skiing
on
a
snowy
slope.
<|endoftext|>

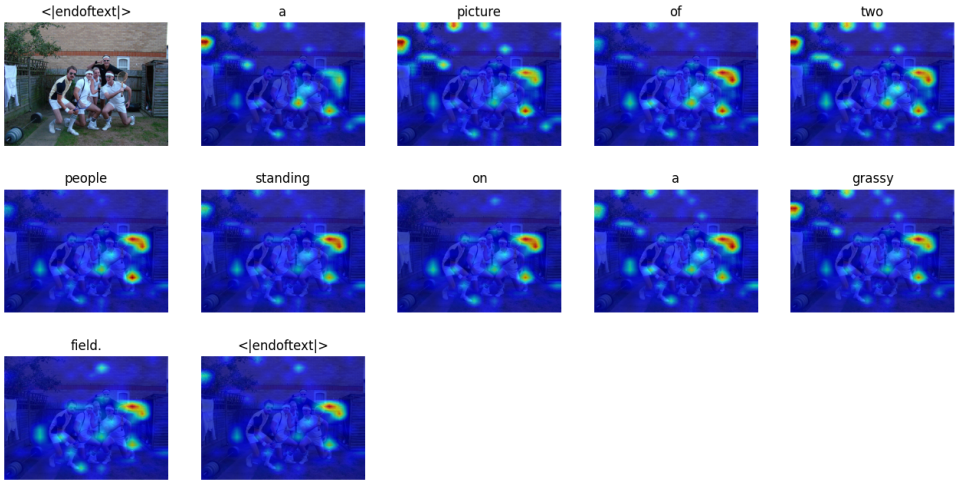


2. According to CLIPScore, you need to: visualize top-1 and last-1 image-caption pairs, report its corresponding CLIPScore in the validation dataset of problem 2.

   top 1, CLIP score: 0.9491:



last 1, CLIP score: 0.3812:



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?

   First of all, surprisingly, I found that if I get the attention scores from the last or the second last layer, the attention maps will be all similar, for example:

   Also, I found that get the attention scores from about the third or fourth transformer layers give the best result, so the visualizations for problem 1 is all from the fourth layer, which is decoder.transformer.h[3]. In my opinion, the model might finish the task since the fourth layer, so the rest layers did not do anything other than transfer the information, which do

not give the reasonable attention scores.

For the visualizations in problem 1, as mentioned above, the red circles represent the regions that reflect the corresponding word in the caption, I think it's quite specific. Also, the model always pay attention in the region of the main objects such as the pizza part for girl.jpg, and the two sheep in sheep.jpg.

For the visualizations in problem 2, follow my previous discussion, the controller part in the top 1 figure is the region that most corresponds to the caption, while the boy part is the main object in the picture so the attention score is always at a high level. However, for the last 1 figure, since the attention score is always high on some weird region such as the background, which may result in the low clip score.