r10921a36 石子仙

Problem 1:

1. Methods analysis (3%)

   Previous methods (e.g. VGG and ResNet) are good at one task and one task only, and requires significant efforts to adapt to a new task. Please explain why CLIP could achieve competitive zero-shot performance on a great variety of image classification datasets.

   I survey this question with some CLIP related paper, the source is in the reference part. CLIP is pre-trained by large-scale image-text pairs. It extracts both features of input images and texts by independent encoders, and aligns the paired ones within the same embedding space. When doing classification, given a new dataset with images of "unseen" classes, CLIP constructs the textual inputs by the category names. Then, CLIP converts the original classification task into an image-text matching problem. That's why CLIP could achieve competitive zero-shot performance on a great variety of image classification datasets.

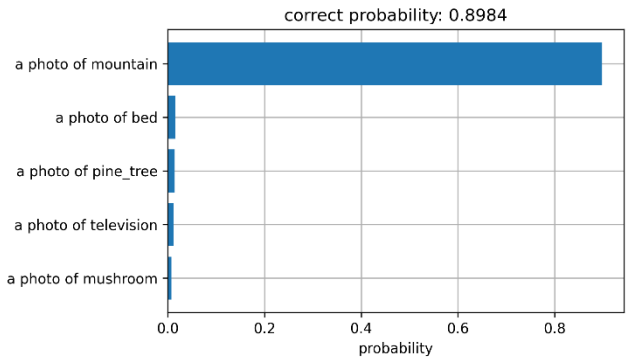
2. Prompt-text analysis (6%)
   - Please compare and discuss the performances of your model with the following **three** prompt templates:
   - *"This is a photo of {object}"*

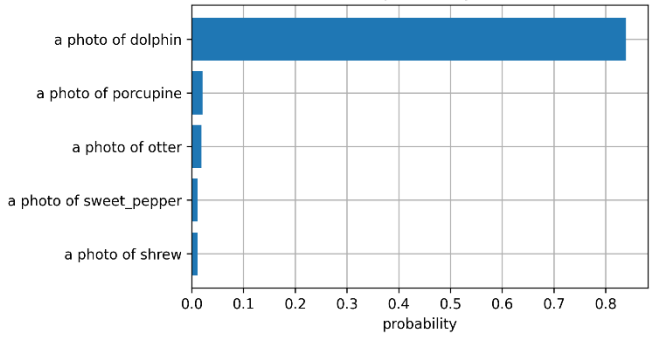
     acc: 0.8136
   - *"This is a {object} image."*

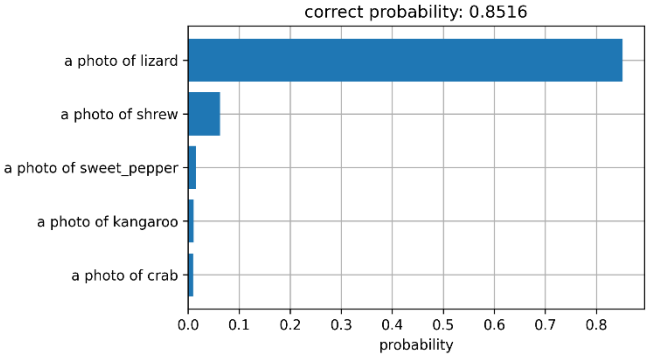
     acc: 0.7076
   - *"No {object}, no score."*

     acc: 0.4536
   - By previous problem, I think the performance is related to the pre-trained model, and the text training set in pre-trained model is related to our usual usage of speaking. Maybe most of texts in text-image pairs is similar with *"This is a photo of {object}."* For *"No {object}, no score.,"* which is not a common usage when saying to an image, that's why it has worst performance.

3. Quantitative analysis (6%)
   - Please sample **three** images from the validation dataset and then visualize the probability of the top-5 similarity scores as following example:







Problem 2:

1. Report your best setting and its corresponding CIDEr & CLIPScore on the validation data. (TA will reproduce this result) (2.5%)
   - Score:
     - CIDEr: 0.9037382654475961
     - CLIPScore: 0.7324988086374197
   - Setting:
     - Encoder: pretrained timm vit_large_patch14_224_clip_laion2b
     - Optimizer: Adam
     - Learning rate: 1e-4
     - Num of Decoding layer: 6 layers
     - Loss: CrossEntropy

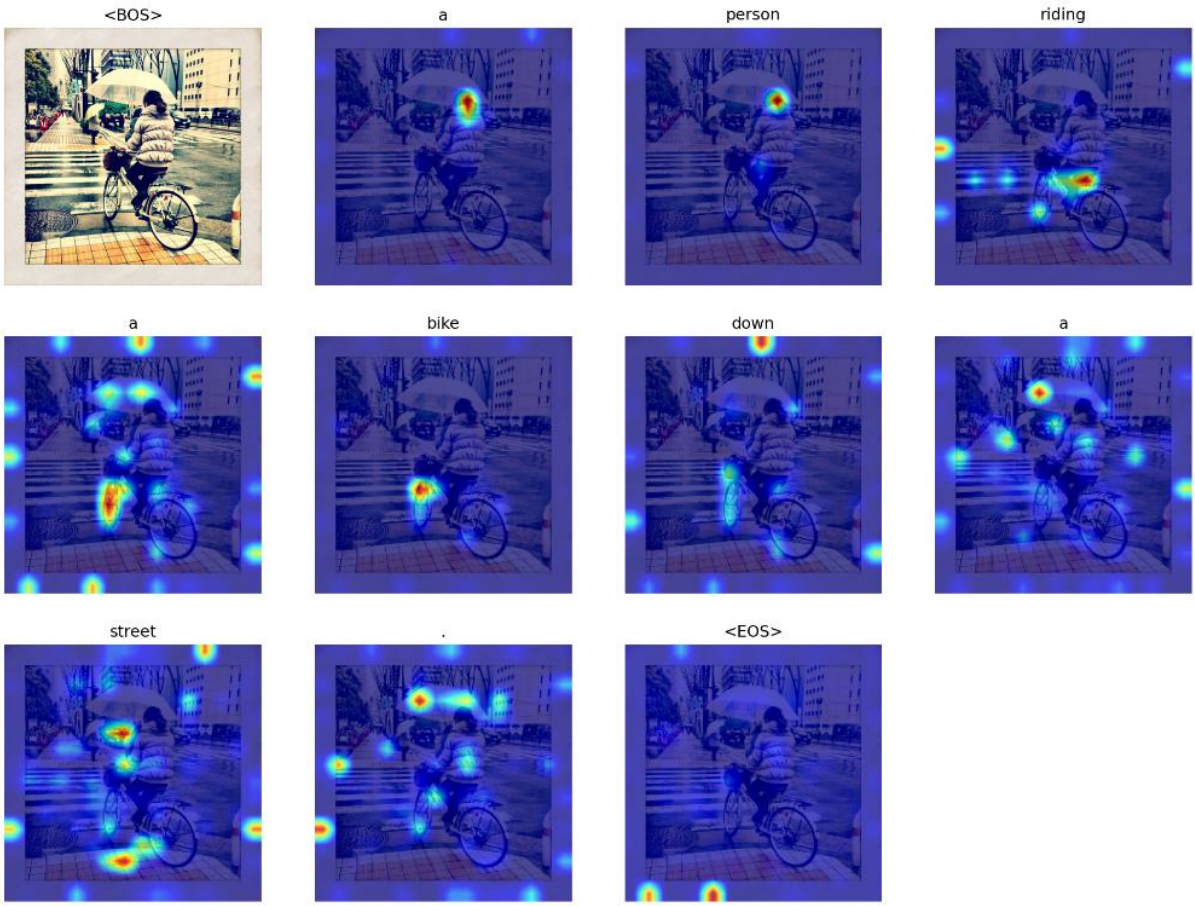
- Decoding strategy: My best setting is using beam search, but it will take very long time to decoding. Hence, I choose using greedy search for decoding.

2. Report other 3 different attempts (e.g. pretrain or not, model architecture, freezing layers, decoding strategy, etc.) and their corresponding CIDEr & CLIPScore. (7.5%, each setting for 2.5%)

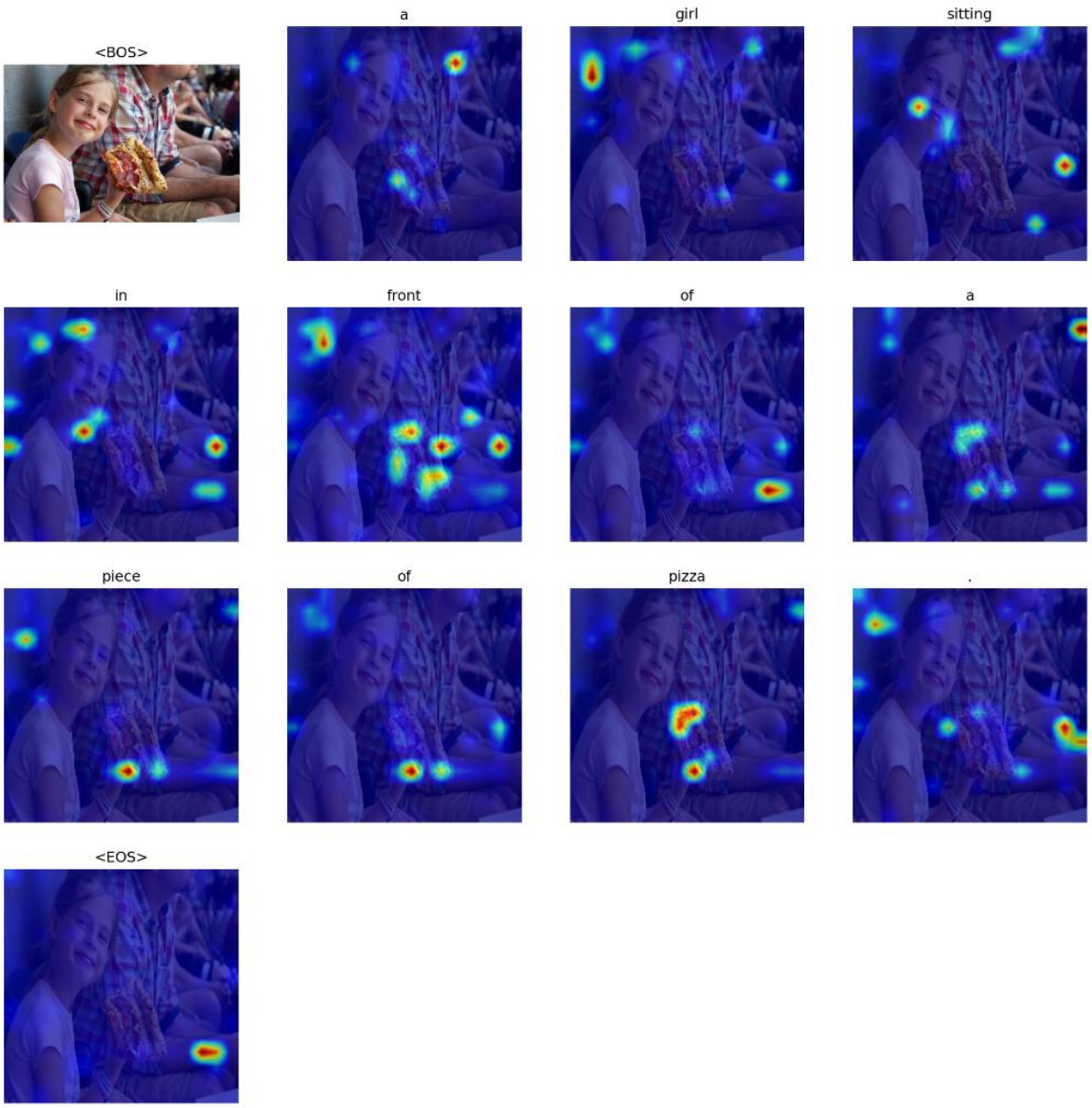
Setting	**CIDEr**	**CLIPScore**
Beam search with 3	0.9516	0.7280
All layer no pretrained	0.0860	0.4812
Less layers (3 layers)	0.7394	0.6962

Problem 3:

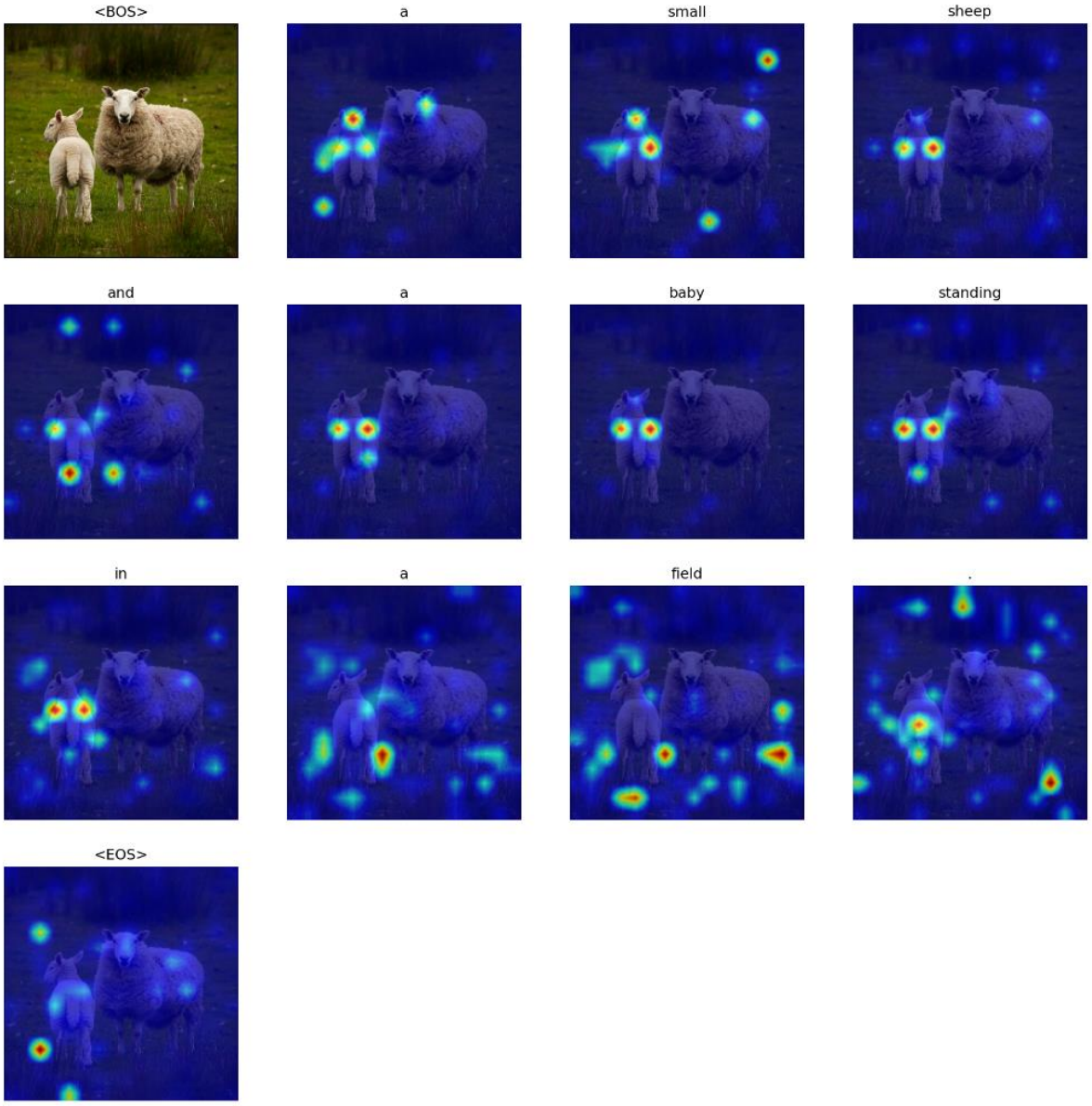
1. TA will give you five test images ([p3_data/images/]), and please visualize the **predicted caption** and the corresponding series of **attention maps** in your report with the following template: (10%, each image for 2%)
   - bike.jpg



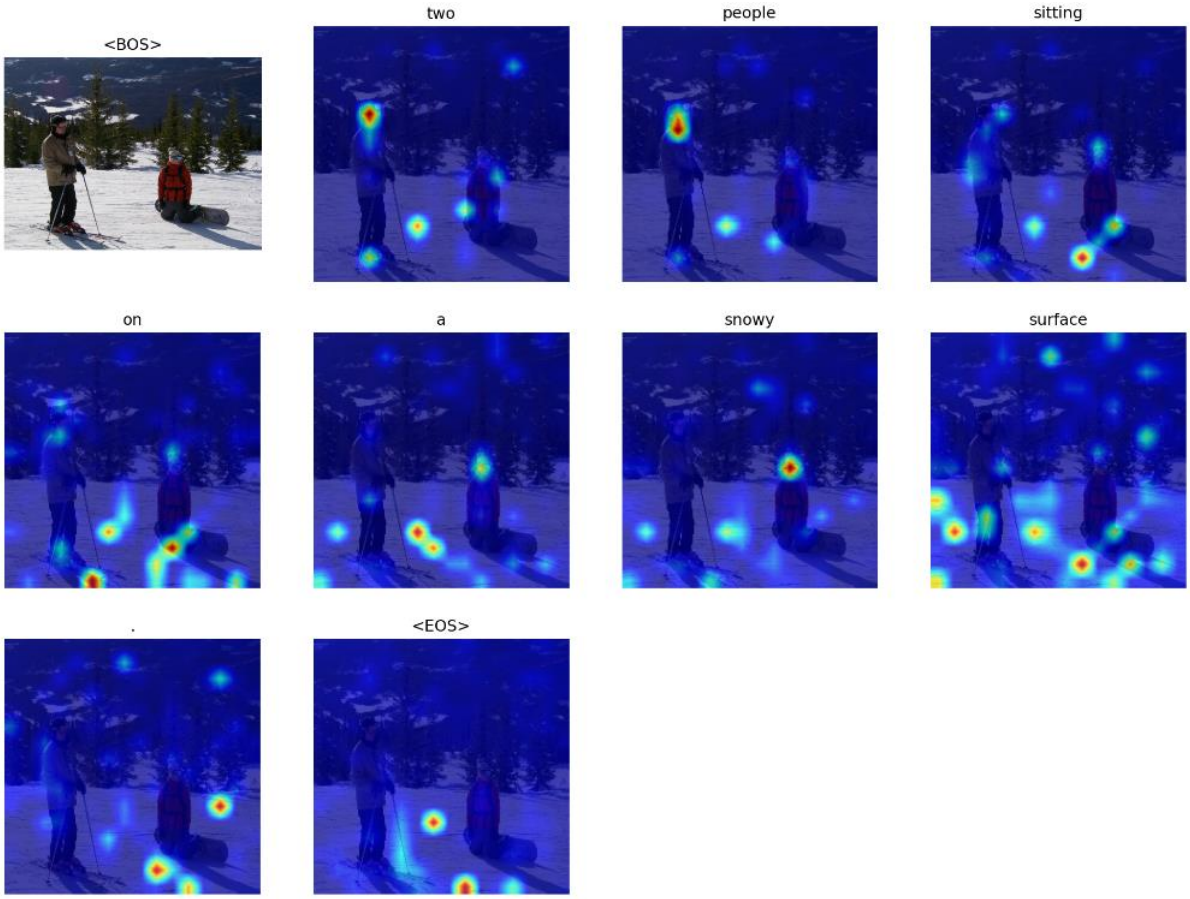
- girl.jpg



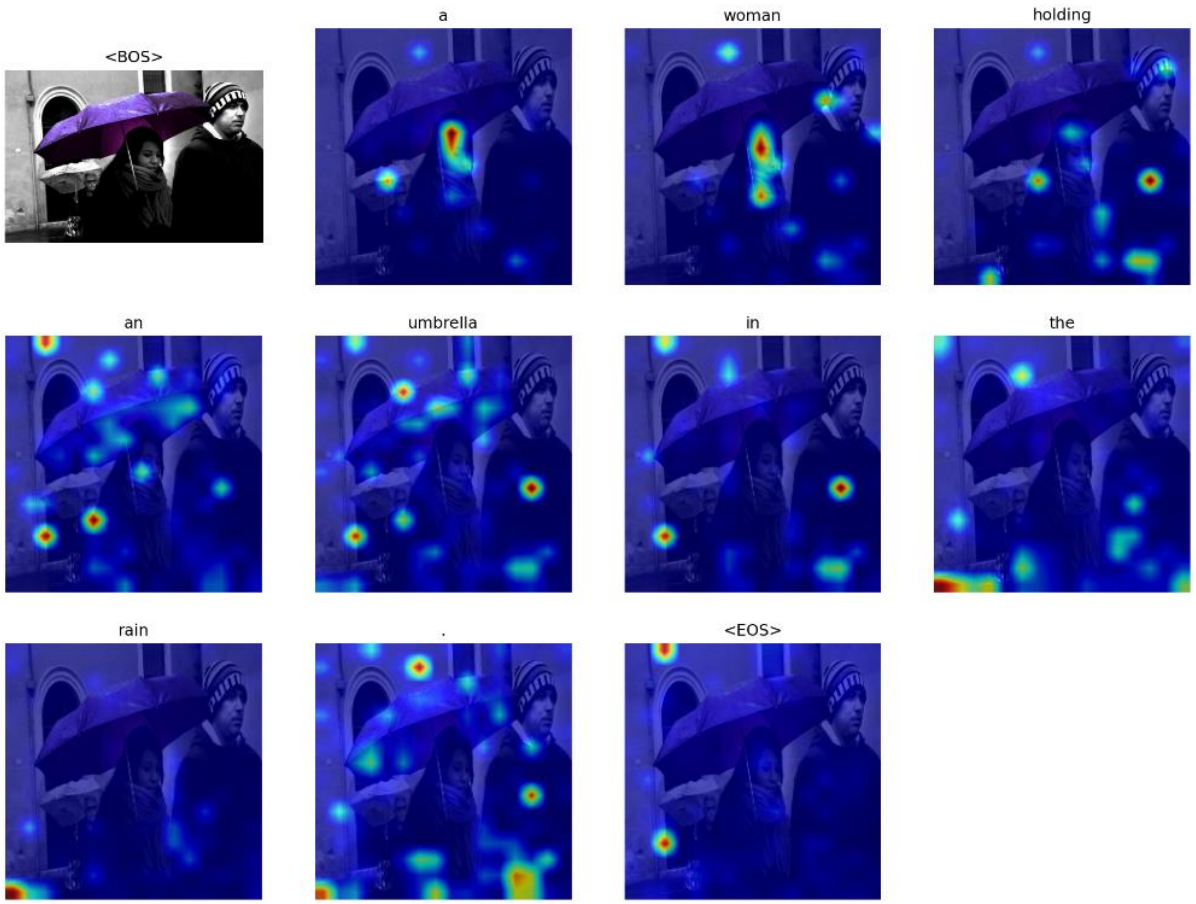
- sheep.jpg



- ski.jpg



- umbrella.jpg



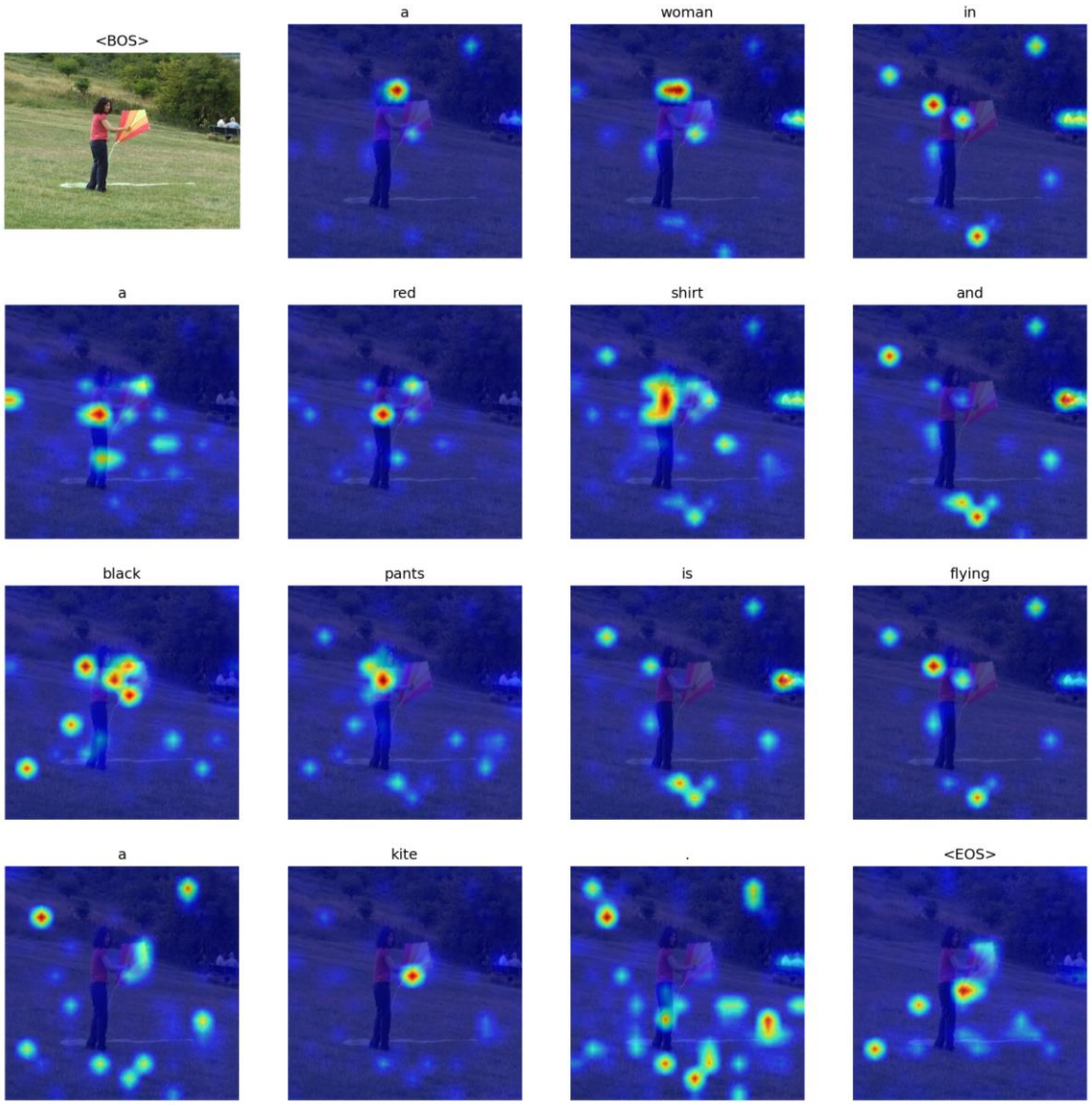
2. According to **CLIPScore**, you need to visualize:
   i. top-1 and last-1 image-caption pairs

   [Top-1]

   Image: 000000179758.jpg

   CLIPScore: 0.9800

   Predicted Caption: a woman in a red shirt and black pants is flying a kite.

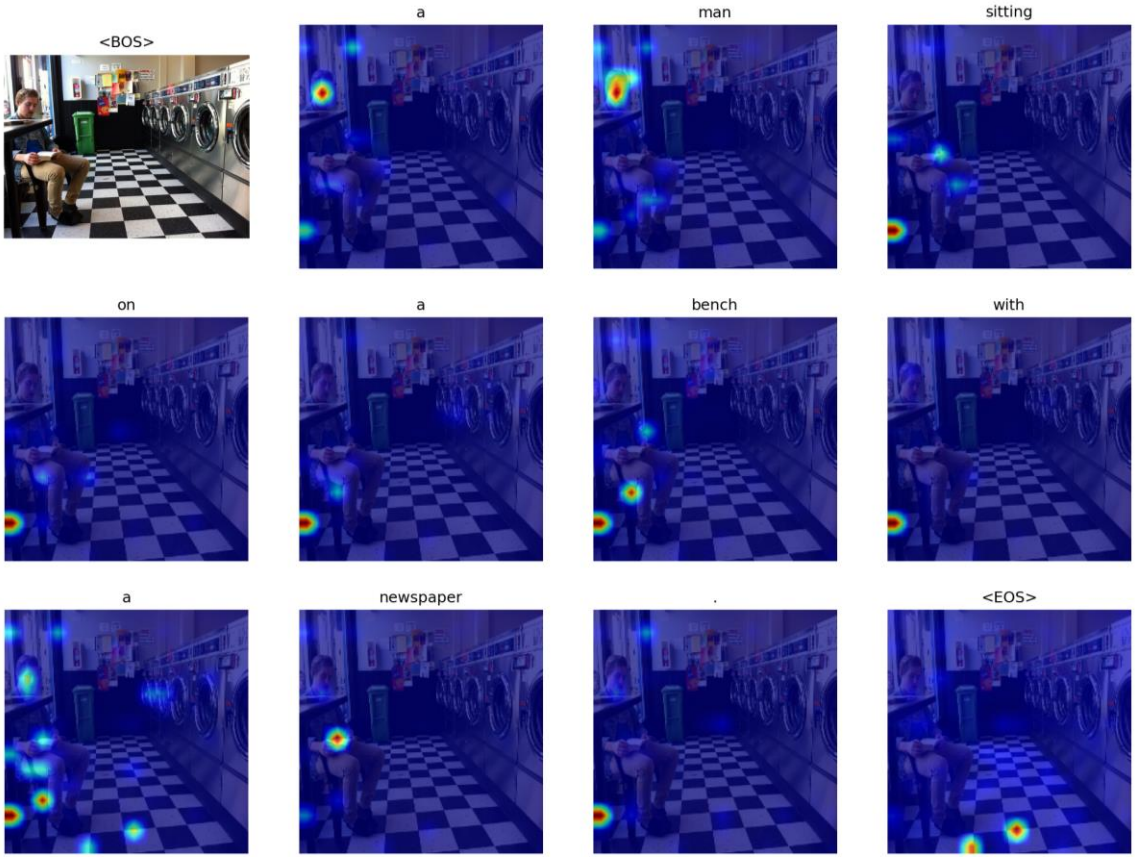


[Last-1]

Image: 6320721815.jpg

CLIPScore: 0.3949

Predicted Caption: a man sitting on a bench with a newspaper.



ii. its corresponding CLIPScore

[Top-1]

- Image: 000000179758.jpg
- CLIPScore: 0.9800

[Last-1]

- Image: 6320721815.jpg
- CLIPScore: 0.3949

in the validation dataset of problem 2. (5%)

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? (5%)
   For top-1, the caption is reasonable. I think the attended region reflect the words, you can see ‘woman,’ ‘red,’ ‘shirt’ and ‘kite’ are precisely shown with the attended region.
   For last-1, I think the caption is reasonable, it just misunderstand book with the newspaper, but in the other part it is correct. I also think the attended region reflect the word, because ‘man,’ ‘sitting’, ‘bench’ and ‘newspaper’ are correspond to the image.

Refernces:

- Part1:

- https://github.com/openai/CLIP
- For report question1: CALIP: Zero-Shot Enhancement of CLIP with Parameter-free Attention https://arxiv.org/pdf/2209.14169.pdf

- Part2 & 3:
  - https://github.com/saahiluppal/catr/blob/master/main.py
  - https://github.com/zarzouram/image_captioning_with_transformers