DLCV HW3

r11922a05 資工AI碩一 林聖硯

Problem 1: Zero-shot image classification with CLIP

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.

The core idea of CLIP model is to learn visual representation from **natural language supervision**. That is, CLIP models will need to learn to recognize **a wide variety of visual concepts in images and associate with their names** rather than direcrly learning **how to recognize a certain kind of image**. As a result, CLIP models can then be applied to nearly arbitrary visual classification tasks.

2. Prompt-text analysis: Please compare and discuss the performances of your model with the following three prompt templates:

   (a) “This is a photo of {object}”

   (b) “This is a {object} image.”

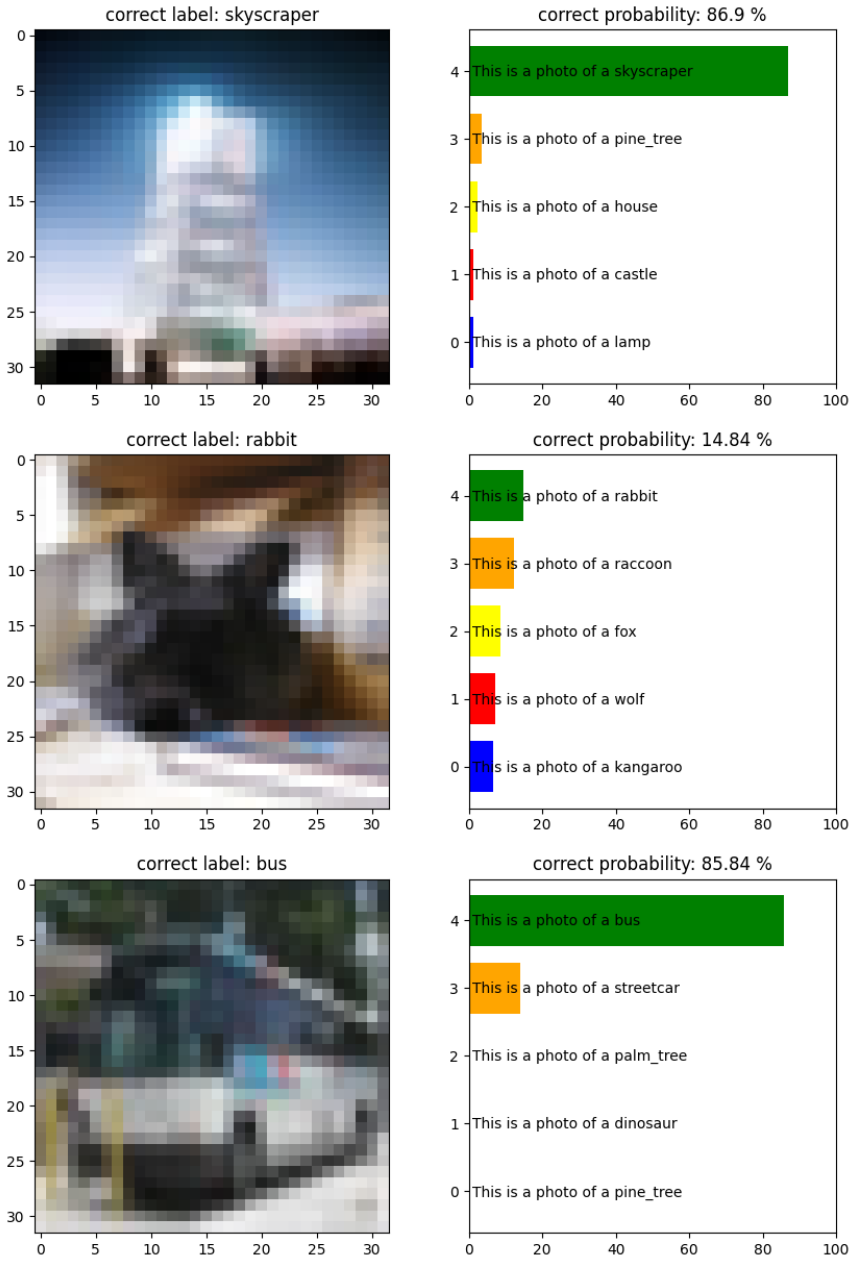
   (c) “No {object}, no score.”

I tried the folloing model-prompt combinations.

val accuracy	Prompt - (a)	Prompt - (b)	Prompt - (c)
Model - RN50	0.51	0.3824	0.2468
Model - RN101	0.5992	0.5344	0.3264
Model - RN50x4	0.5	0.4628	0.2744
Model - RN50x16	0.5976	0.5288	0.3788
Model - ViT-B/32	0.6948	0.6816	0.5628
Model - ViT-B/16	0.7556	0.6168	0.4872

Disuccsion: From the experiment results, we could conclude that prompt has significant impact on the model performance. The best prompt in this task is prompt (a). In general, the performance of ViT-based models is much better than resnet-based models.

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



Problem 2: Image Captioning with VL-model

1. Report your best setting and its corresponding CIDEr & CLIPScore on the validation data. (TA will reproduce this result) (2.5%)

Best settings

- Encoder: vit_large_patch14_224_clip_laion2b + freeze encoder
- Decoder (structure reference): number of head = 8, number of hidden dimension = 1024, number of decoder layer = 6
- optimizer: AdamW, lr = 2e-5, epoch = 30

Performance (The sentences look normal but I just don't know why the scores are so low and I've already tried my best T__T)

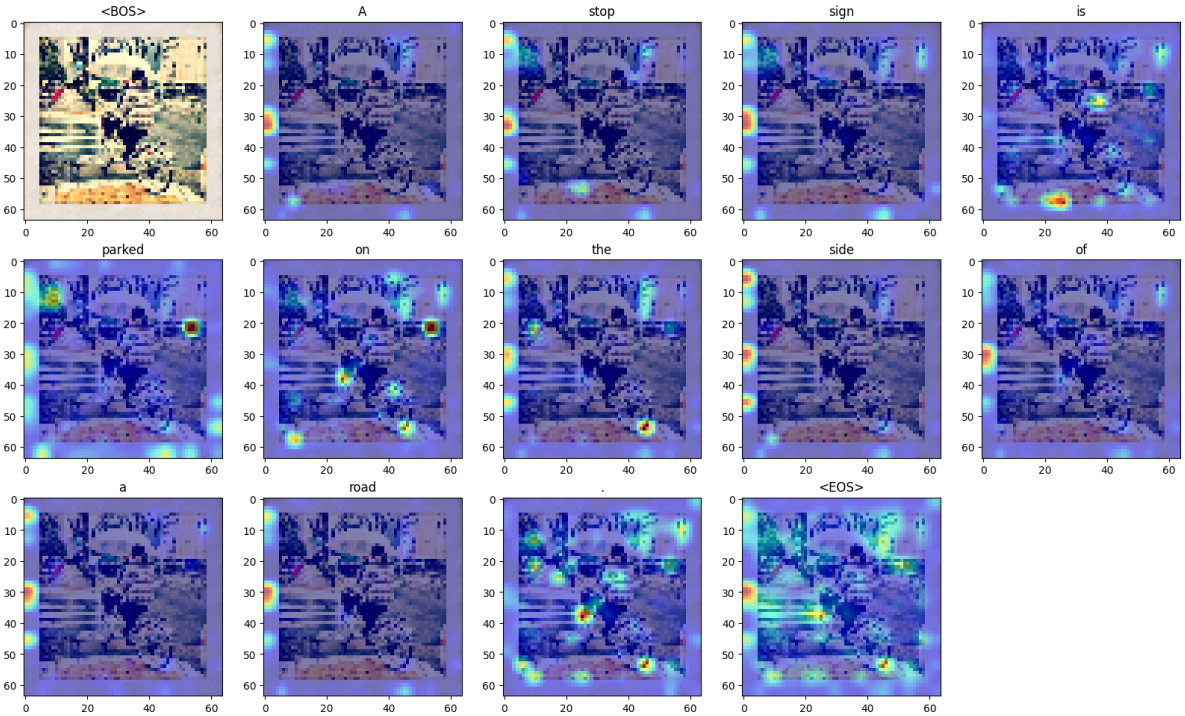
- CIDEr: 0.3616
- CLIPScore: 0.5740

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%)

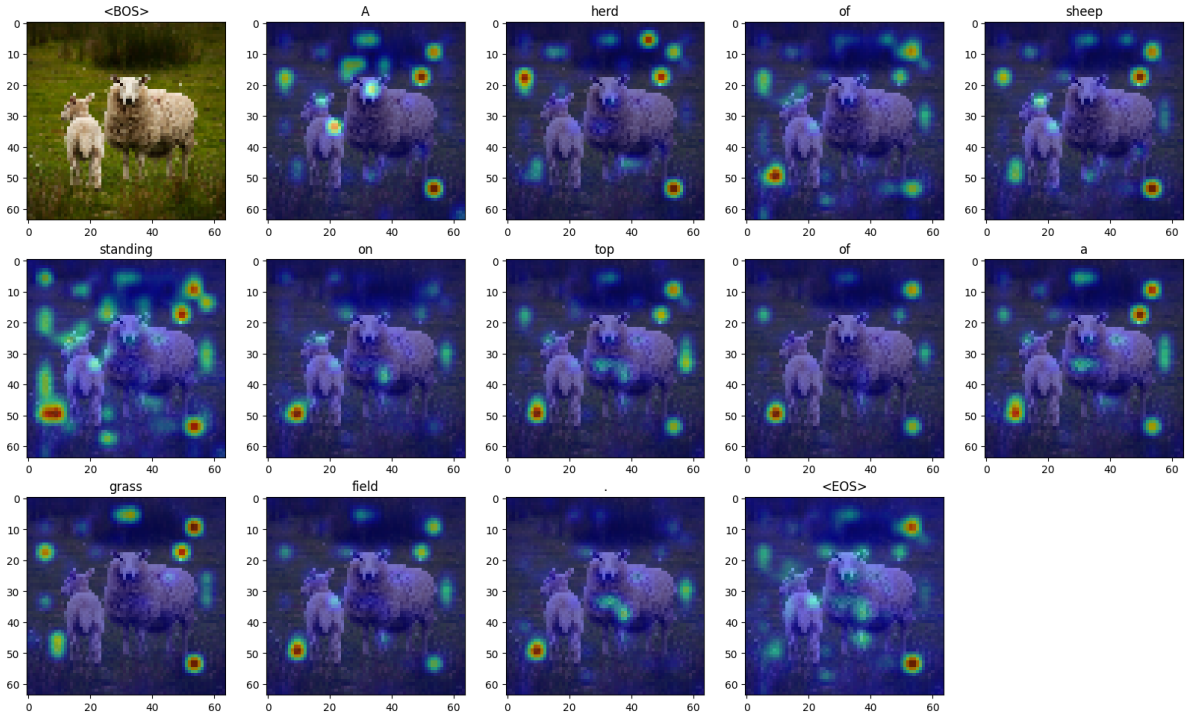
Attempts	CIDEr	CLIPScore
Change encoder to vit _base_patch16_224 in best settings	0.1354	0.4757
Unfreeze encoder in best settings	0.2807	0.3182
Change number of decoder layer to 12 in best settings	0.32593	0.5037

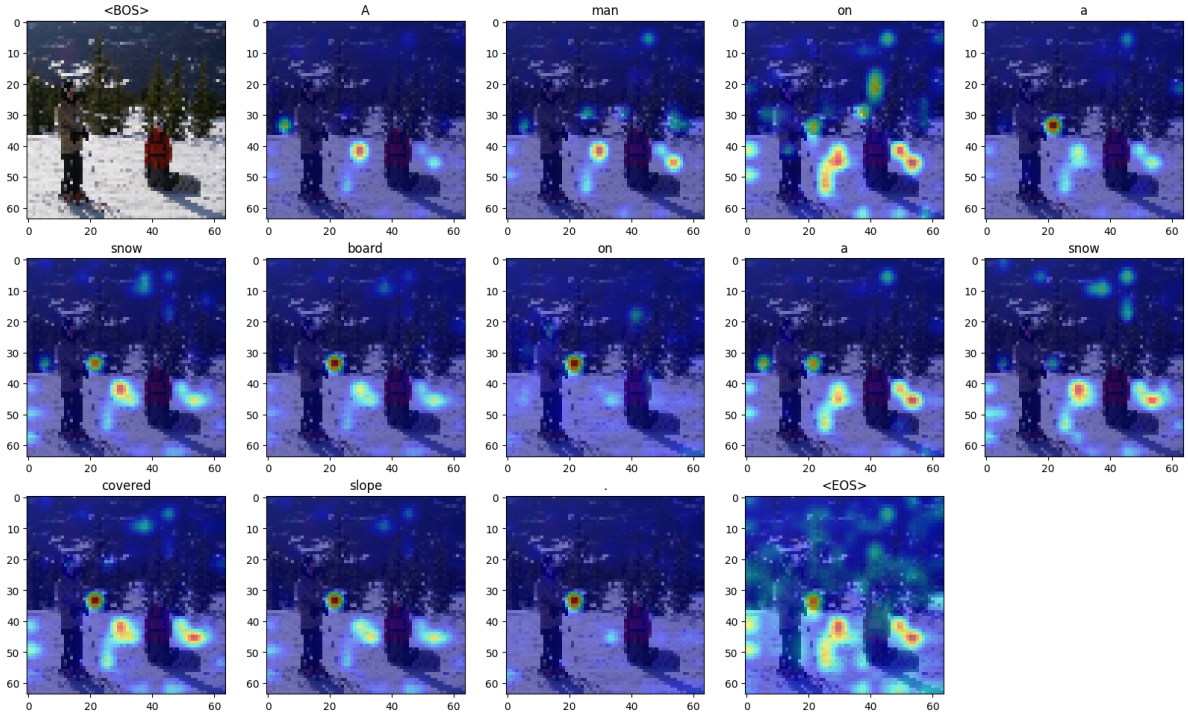
Problem 3: Visualization of Attention in Image Captioning

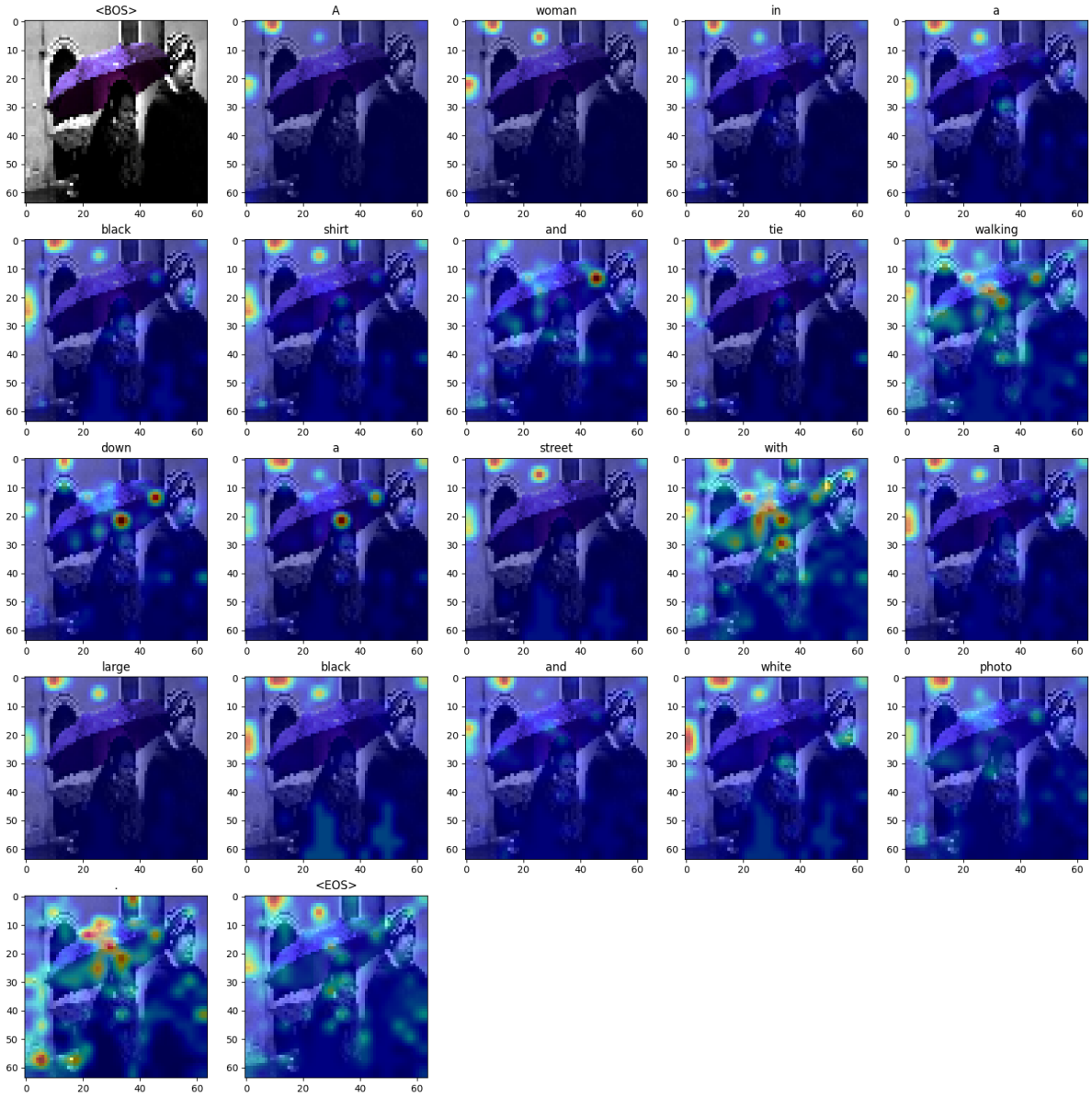
1. TA will give you five test images ([p3_data/images/]), and please visualize the **predicted caption** and the corresponding series of **attention maps** in your report with the following template: (10%, each image for 2%)


<BOS>
A
stop
sign
is
parked
on
the
side
of
a
road
.
<EOS>


<BOS>
A
woman
in
a
red
shirt
is
holding
a
hot
dog
.
<EOS>


<BOS>
A
herd
of
sheep
standing
on
top
of
a
grass
field
.
<EOS>

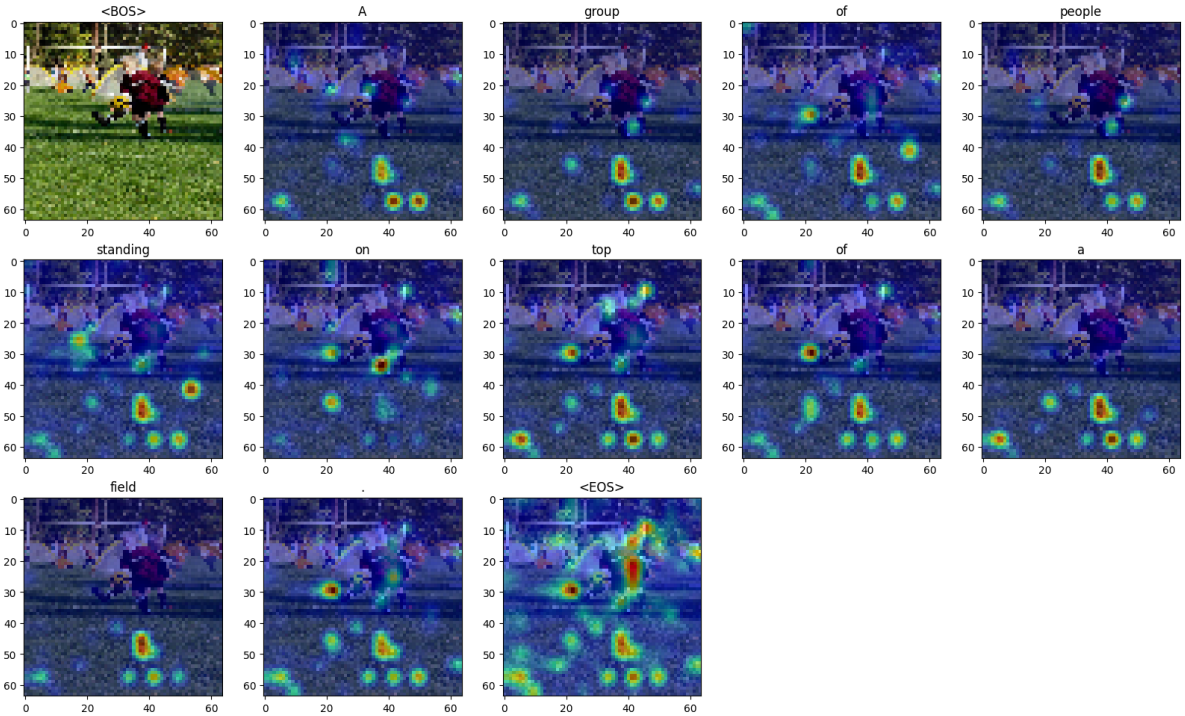

<BOS>
A
man
on
a
snow
board
on
a
snow
covered
slope
.
<EOS>



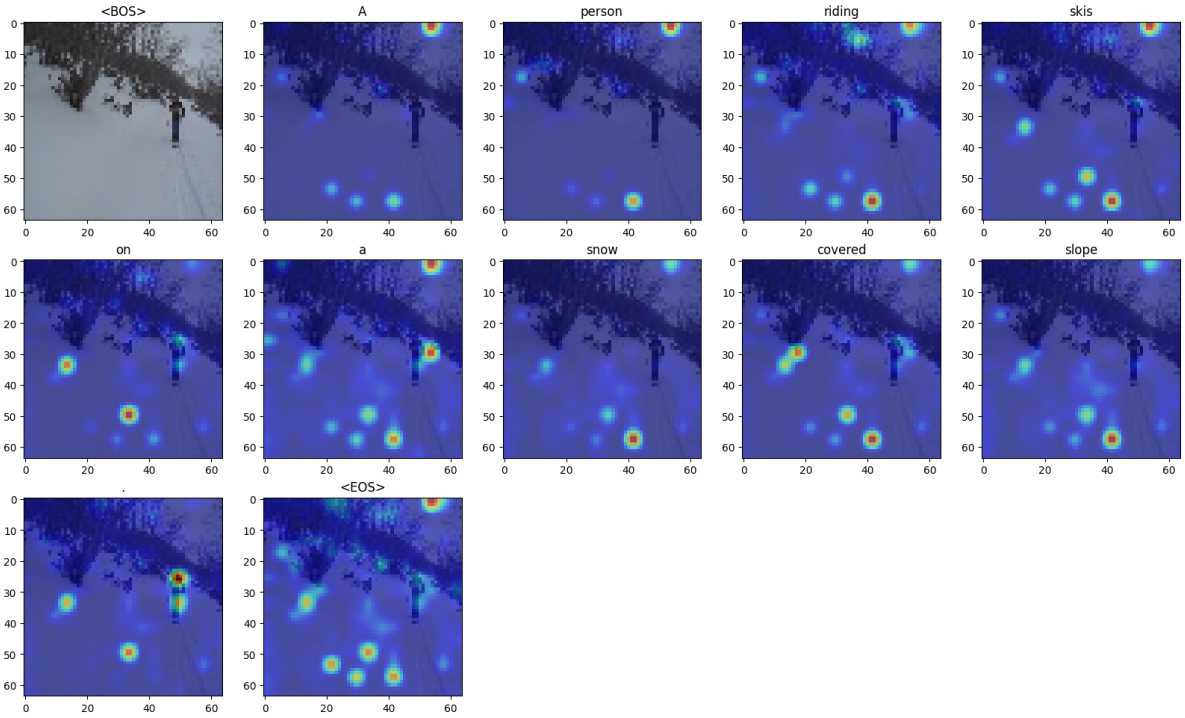
2. According to CLIPScore, you need to visualize:
   a. top-1 and last-1 image-caption pairs
   b. its corresponding CLIPScore

   in the validation dataset of problem 2.

   000000000368.jpg



000000472295.jpg



Img name	CLIP score

000000000368.jpg	0.41778
000000472295.jpg	0.87402

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%)

In both images, the attention maps seem not reasonable and the attended region doesn’t reflect the corresponding word in the caption even if the CLIP score is really high in ‘000000428508.jpg’. However, I believe that the problem is mainly due to the lack of predicting ability of my model. (I’ve also tried to use the attention weight from the first layer to the last layer, but the result seems worse)