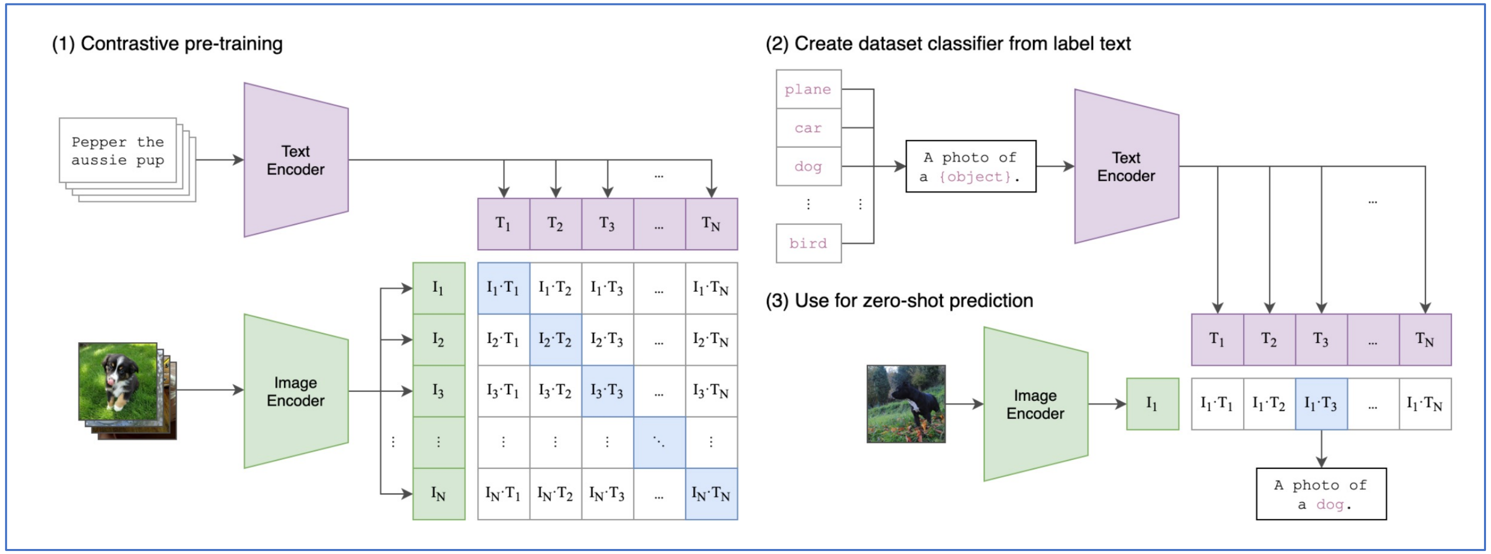
DLCV Hw3 Report

R10942198 林仲偉

Problem 1: Zero-shot image Classification with CLIP (15%)


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
T_1 T_2 T_3 ... T_N
Image Encoder
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
⋮
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

1. Methods analysis (3%)

- Please explain why CLIP could achieve competitive zero-shot performance on a great variety of image classification datasets.

Previous methods (e.g. VGG and ResNet) are fully-supervised method. They need large amount of annotated data to train, and they have less flexibility to do transfer learning.

During training CLIP, N pairs of training data are loaded into model. The image and text coming from same pairs are positive, and those which comes from different pairs are negative. The model maximizes the cosine similarity of positive pairs, and minimize the the cosine similarity of negative pairs. During inference time, it computes the similarity of features from image encoder and text prompt decoder.

CLIP uses enormous amount of data pairs (400 million image and text) from internet to train in a self-supervised way. Also, the number of image class and text prompts are not specified during training. Therefore it can achieve good zero-shot performance.

Rrf: https://arxiv.org/pdf/2103.00020.pdf Fig. 1.

Problem 1: Zero-shot image Classification with CLIP (15%)

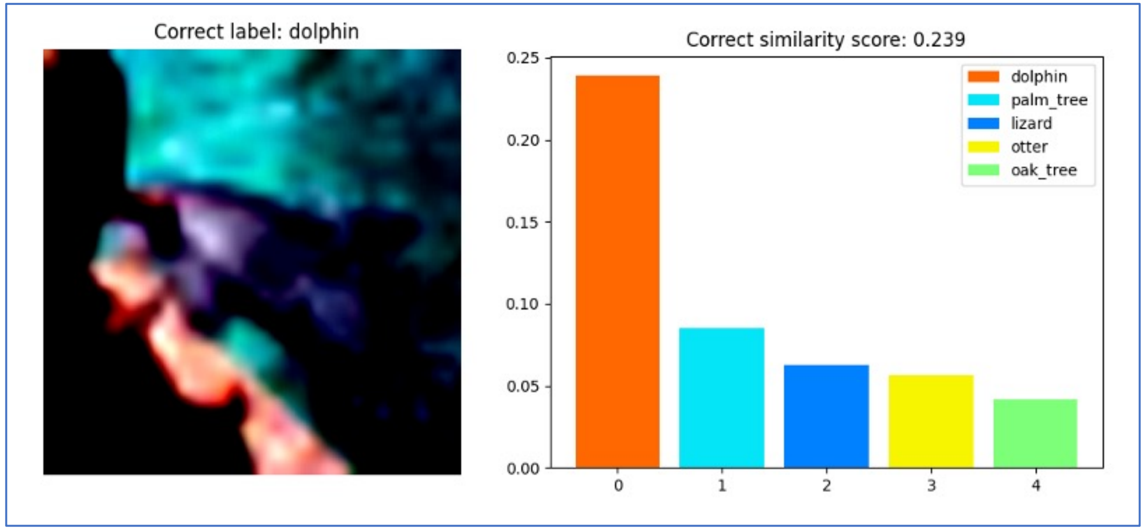
2. Prompt-text analysis (6%)

- Please compare and discuss the performances of your model with the following three prompt templates:
    1) “This is a photo of {object}”
    2) “This is a {object} image.”
    3) “No {object}, no score.”

Prompt Templates	CLIPScore	Discussion
“This is a photo of {object}”	0.609	This prompt template contains another word (photo), so the CLIP score is lower than baseline.
“This is a {object} image.”	0.682	This prompt template implies the data is an image of something. It achieves better performance.
“No {object}, no score.”	0.563	The meaning of this prompt template is not straightforward, so the CLIP score is much lower than baseline.

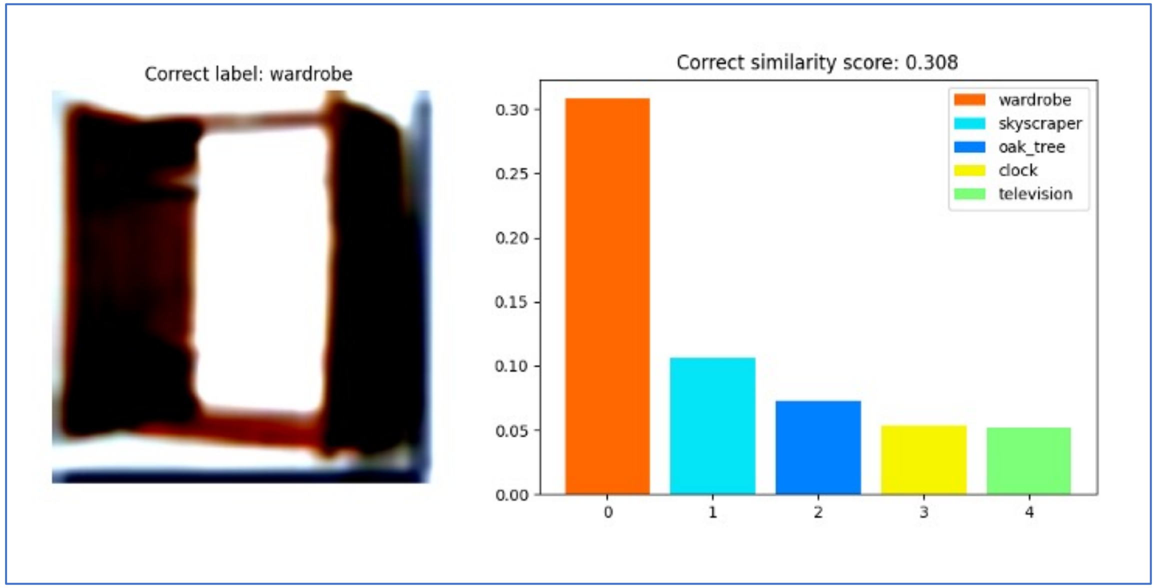
Note: I use the prompt “{object}” to create template as baseline. It achieves clip score=0.635.

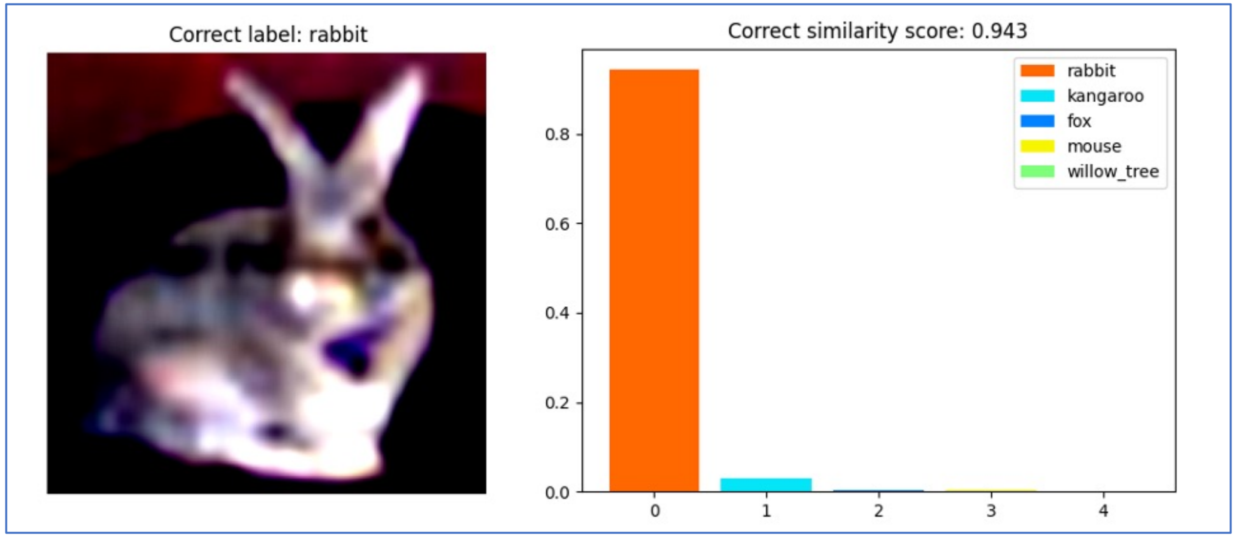
Problem 1: Zero-shot image Classification with CLIP (15%)



3. Quantitative analysis (6%)

- Please sample three images from the validation dataset and then visualize the probability of the top-5 similarity scores





Problem 2: Image Captioning with VL-model (10%)

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

Beam Search (k=5): CIDEr: 0.988 | CLIPScore: 0.715

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%)

Base configuration setting:

Encoder: (freeze all layer in first 6 epochs, then train encoder with 0.1*learning rate in last 6 epochs)

- pre-train `vit_large_patch14_224_clip_laion2b`

Decoder:

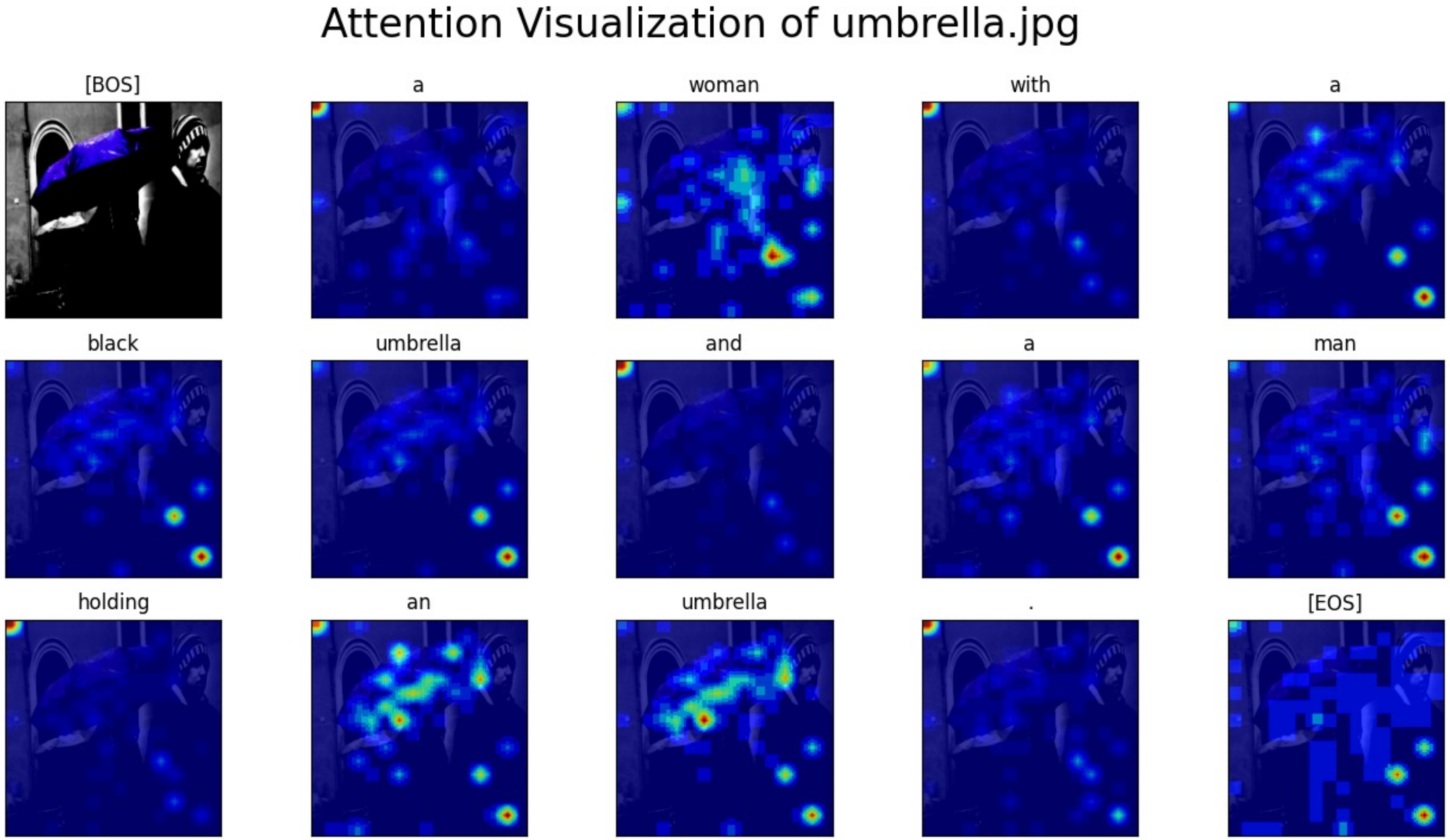
- number of heads = 16
- number of layers = 8
- model dimension = 1024
- feed forward dimension=4096

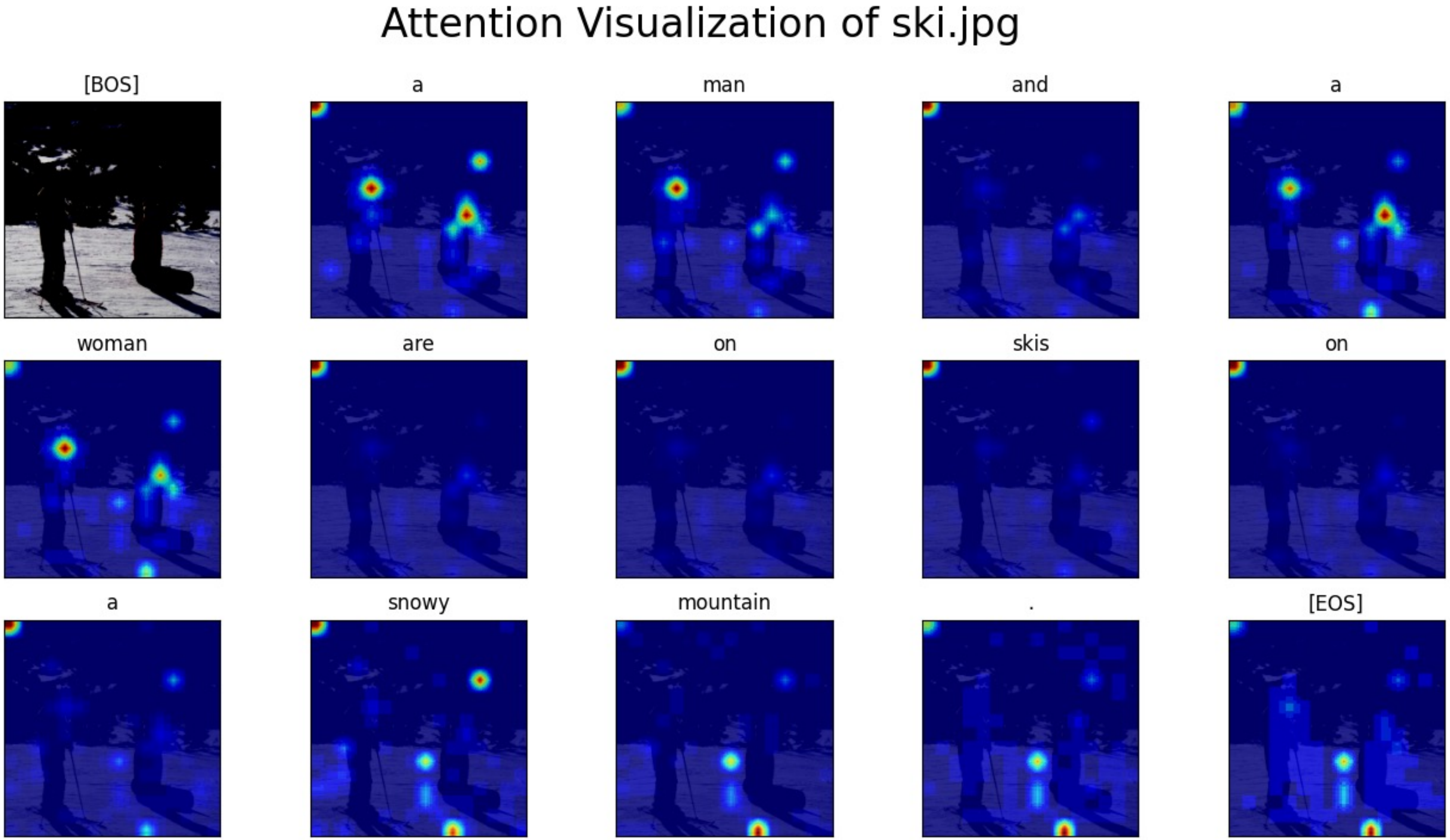
Decode strategy: Beam Search

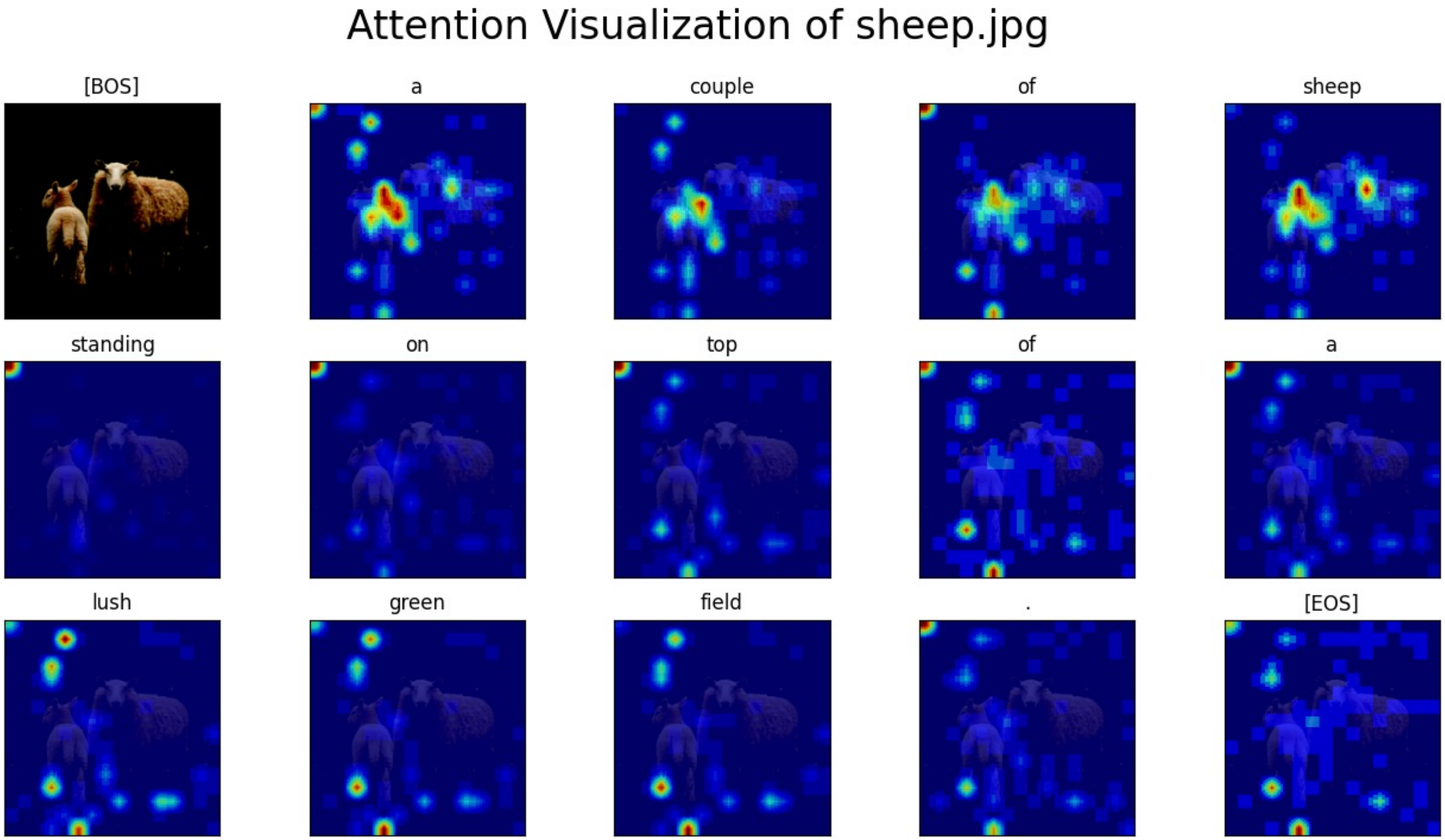
Different attempts	CIDEr	CLIPScore
Encoder is fixed during training	0.789	0.707
Pre-train Encoder: vit_base_patch16_224	0.604	0.681
Decode strategy: Greedy Search	0.896	0.724

Problem 3: Visualization of Attention in Image Captioning (20%)

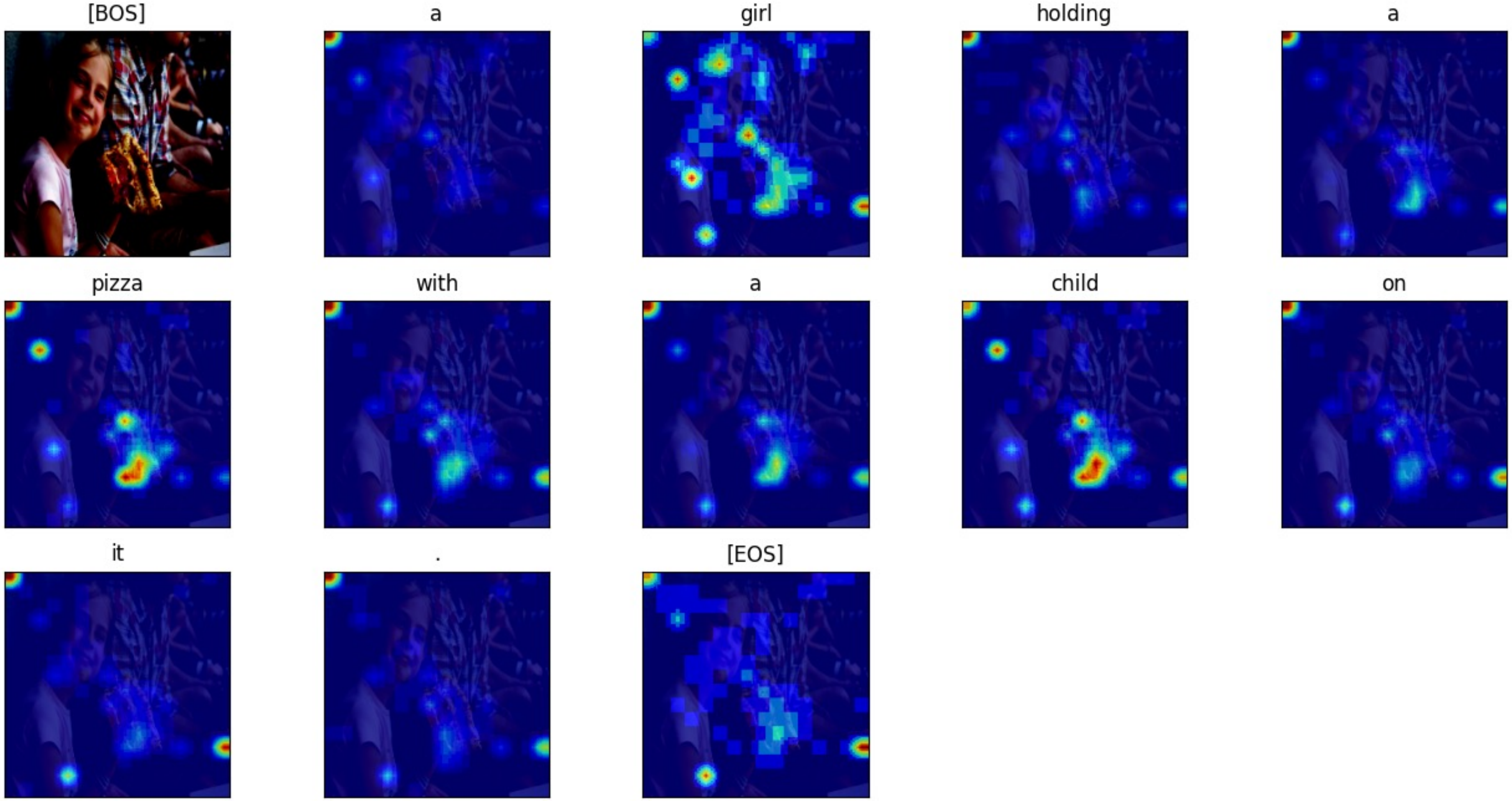
1. TA will give you five test images ([p3_data/images/]), and please visualize the predicted caption and the corresponding series of attention maps in your report: (10%, each image for 2%)



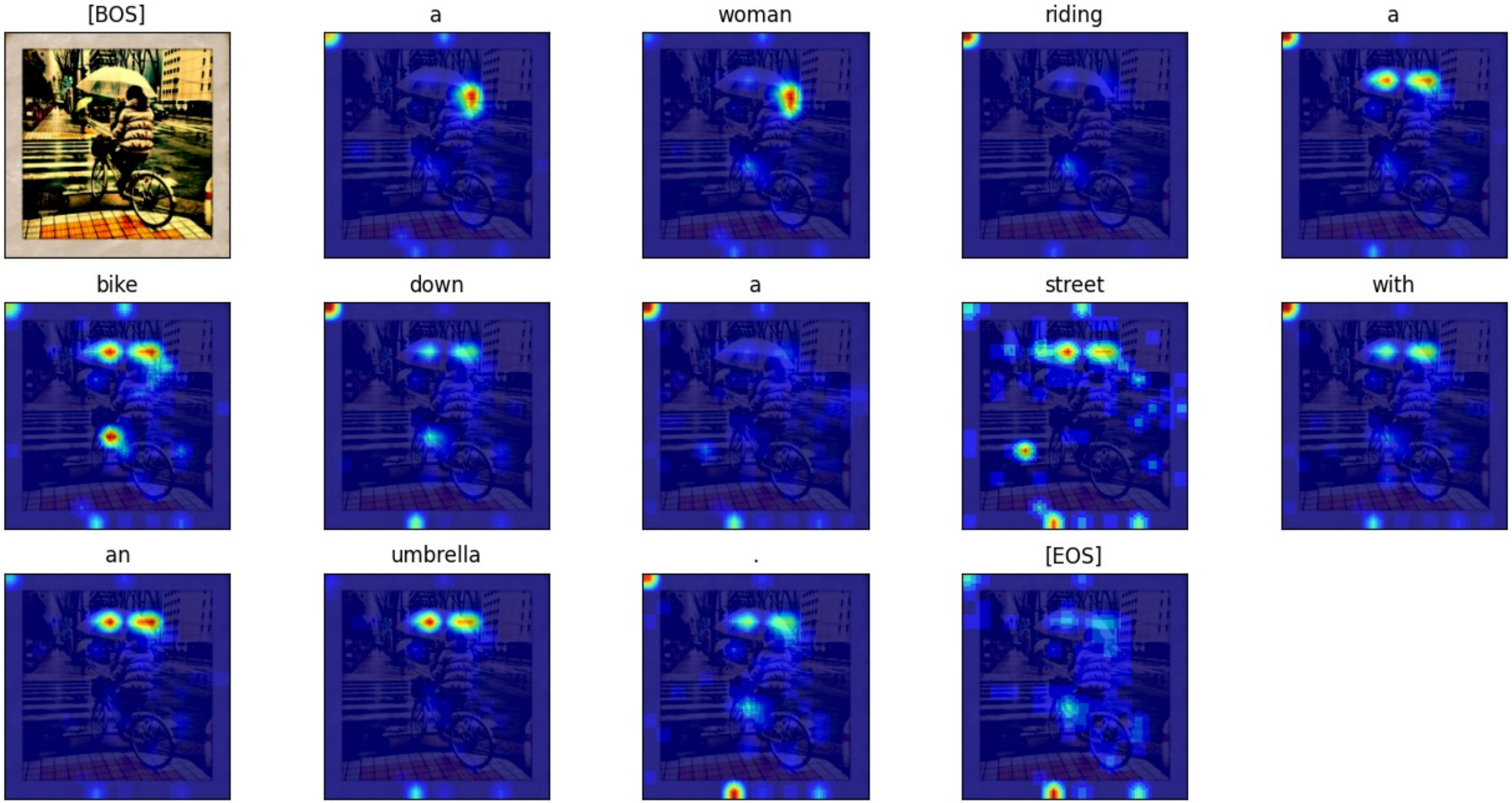

Attention Visualization of ski.jpg
[BOS]
a
man
and
a
woman
are
on
skis
on
a
snowy
mountain
.
[EOS]


Attention Visualization of sheep.jpg
[BOS]
a
couple
of
sheep
standing
on
top
of
a
lush
green
field
.
[EOS]

Attention Visualization of girl.jpg

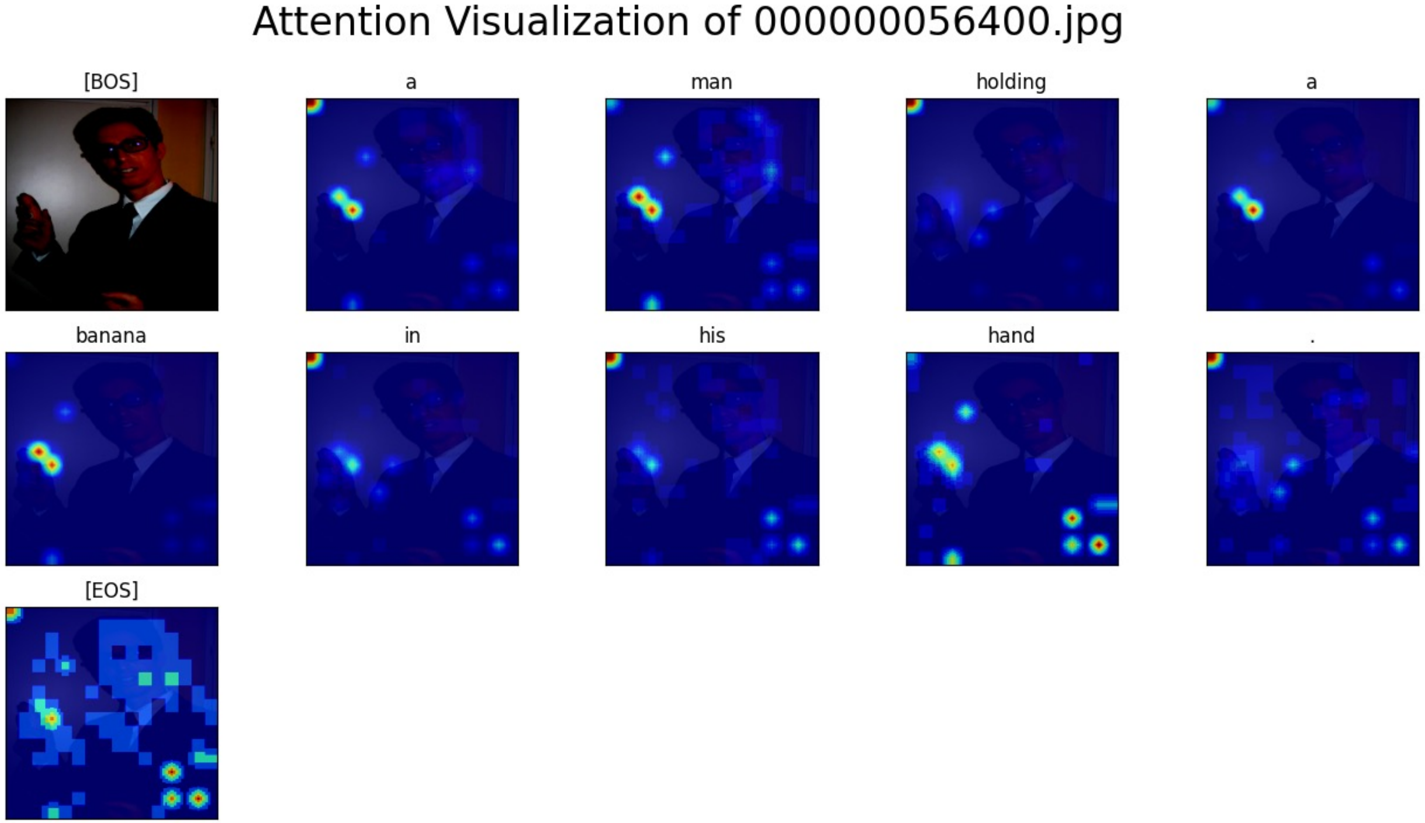


Attention Visualization of bike.jpg



Problem 3: Visualization of Attention in Image Captioning (20%)

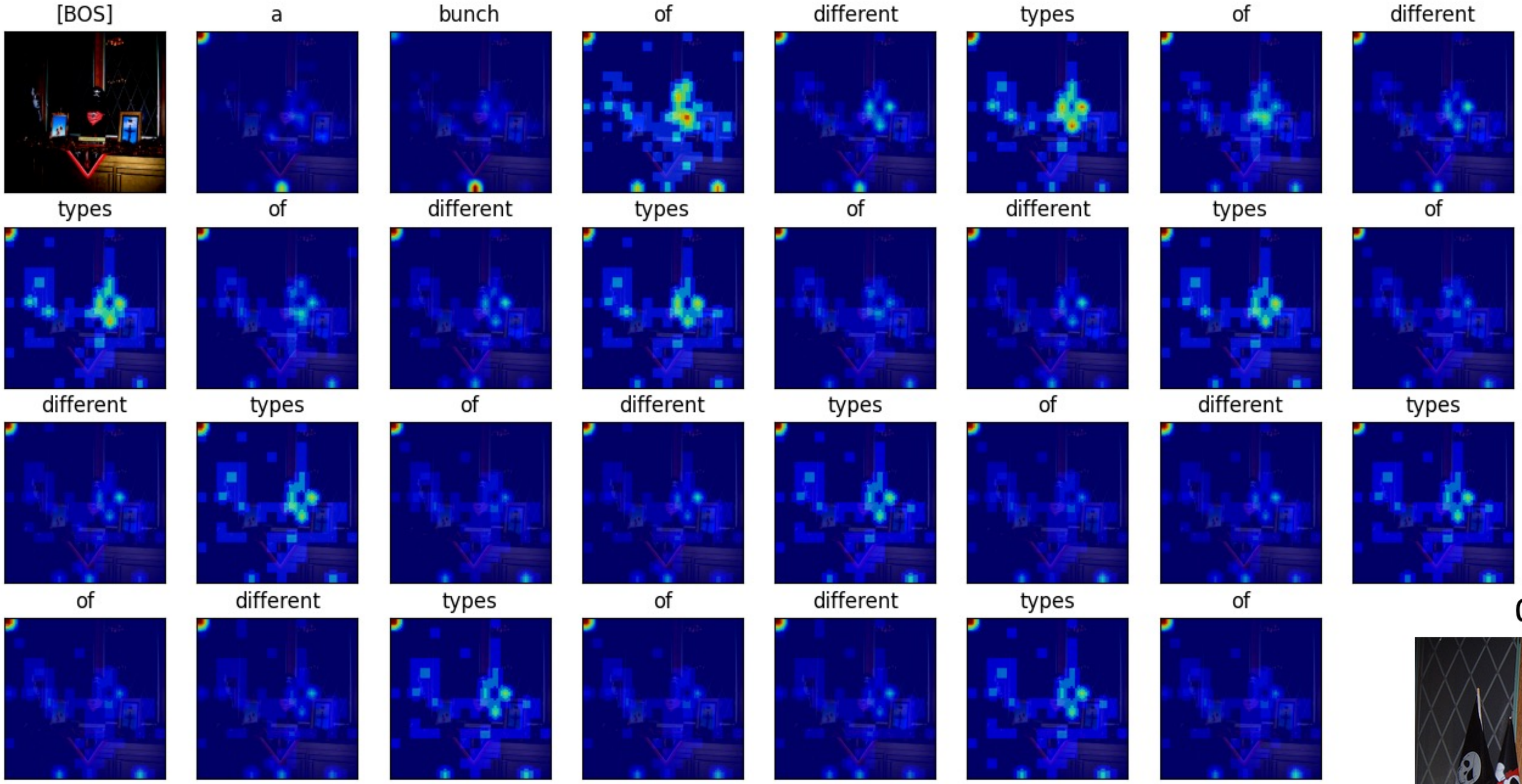
2. According to CIDEr, you need to visualize (i) (ii) in the validation dataset of problem 2. (5%)
    i. Top-1 and last-1 image-caption pairs
    ii. Its corresponding CIDEr score



000000056400.jpg

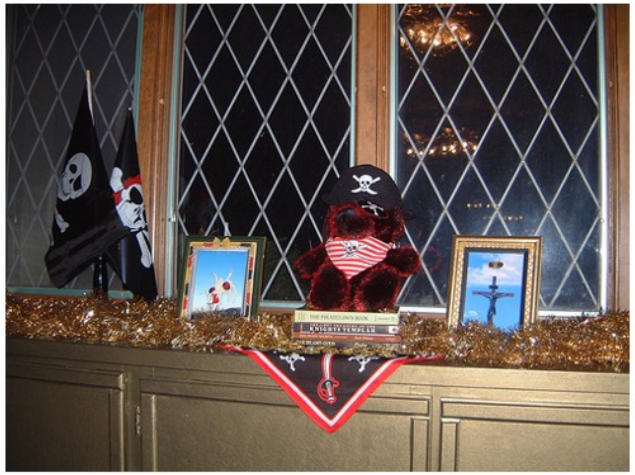
- Image: 000000056400.jpg
- Caption: a man holding a banana in his hand .
- CIDEr score: 2.897173532247371

Attention Visualization of 000000294973.jpg



- Image: 000000294973.jpg
- Caption: a bunch of different types of different types of different types of different types of different types of different types of different types of different types of different types of
- CIDEr score: 0.00017149937733775526

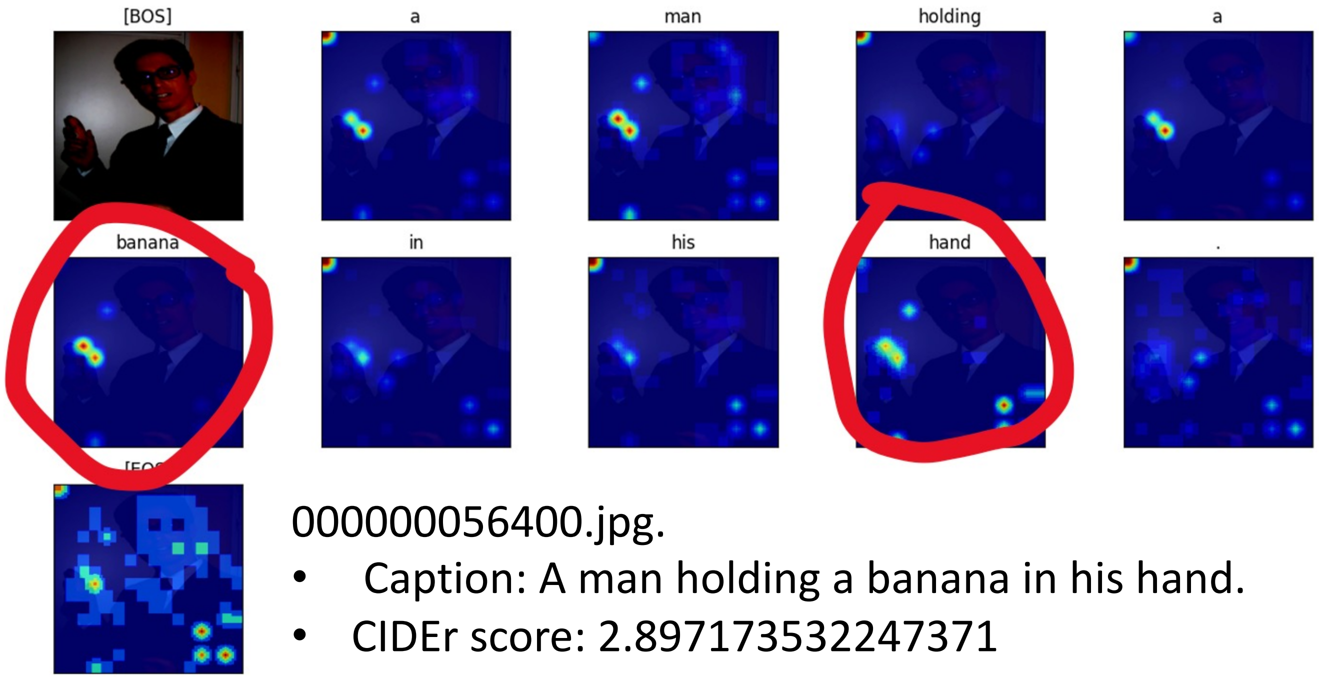
000000294973.jpg



Problem 3: Visualization of Attention in Image Captioning (20%)

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%)

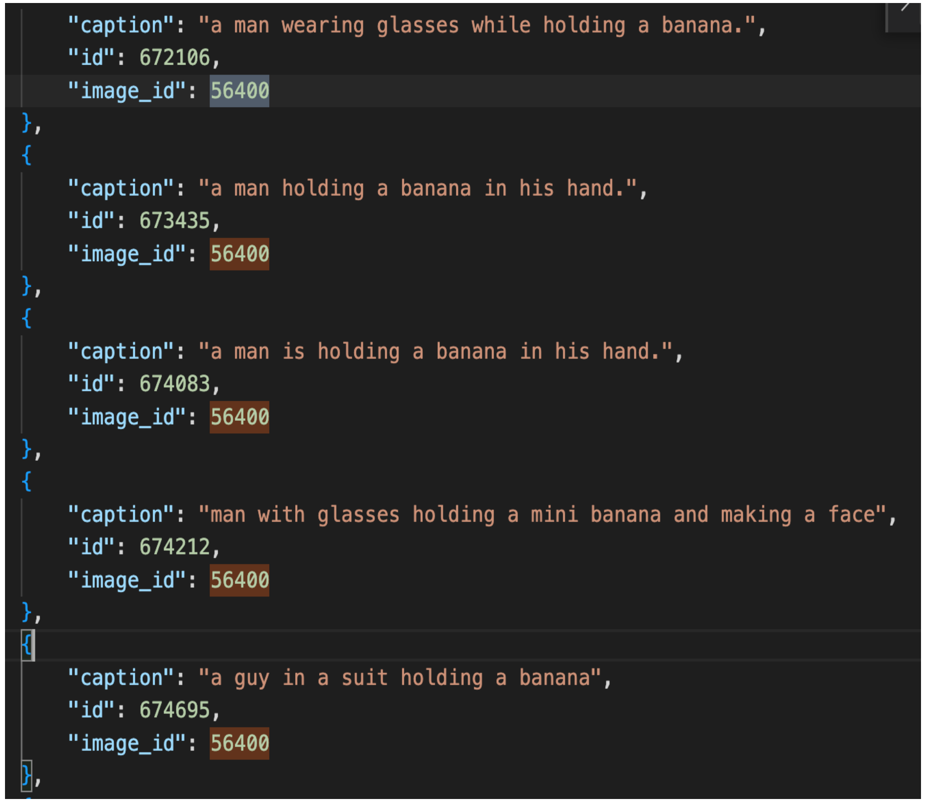
Attention Visualization of 000000056400.jpg



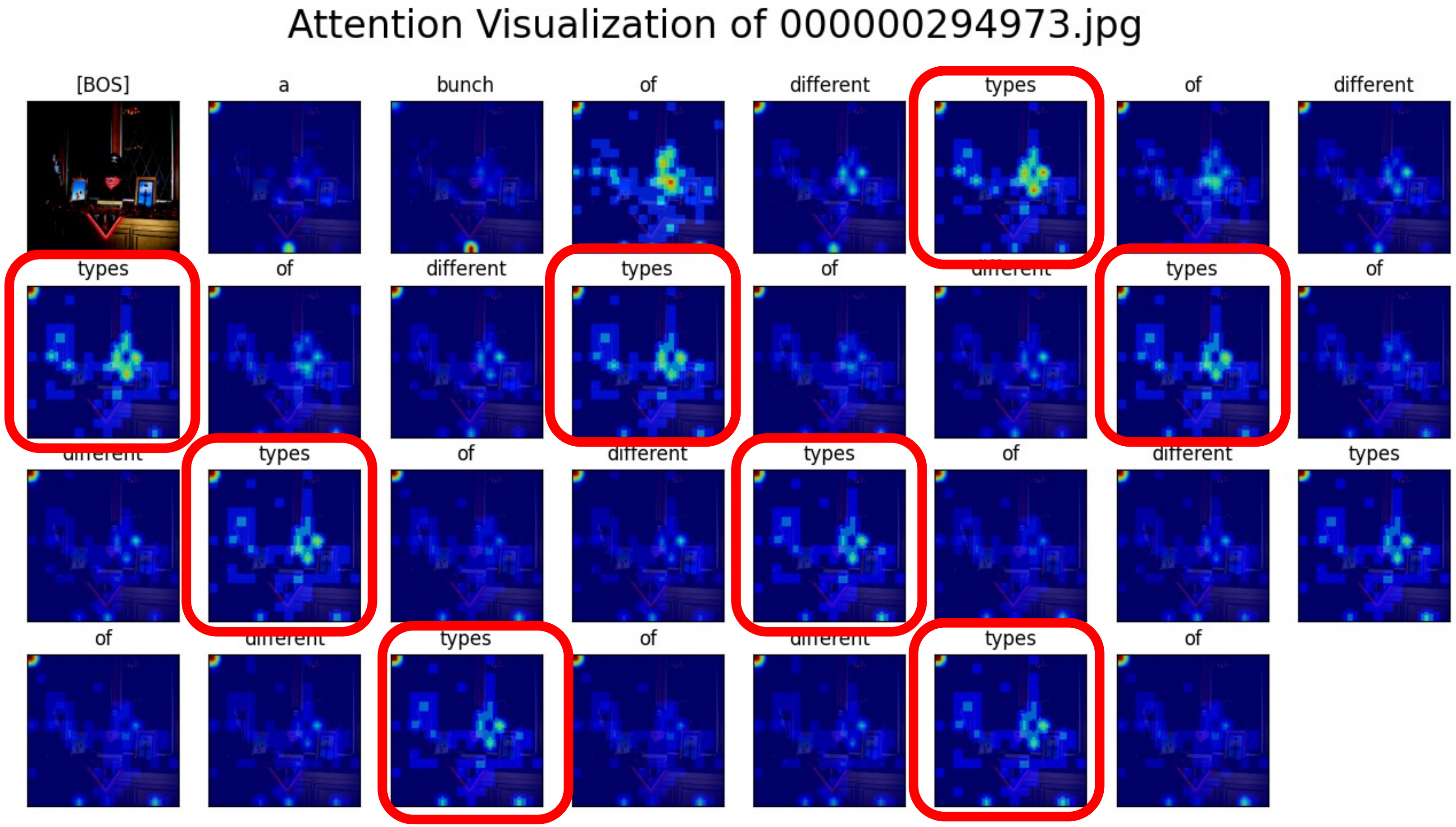
000000056400.jpg.

- Caption: A man holding a banana in his hand.
- CIDEr score: 2.897173532247371
- The attention region briefly reflect the corresponding key word (ex: banana, hand).
- Also, all of the ground truth captions look consistent. The words ”man, hand, holding, banana” show up frequently when describing the image, so it’s a reasonable caption.

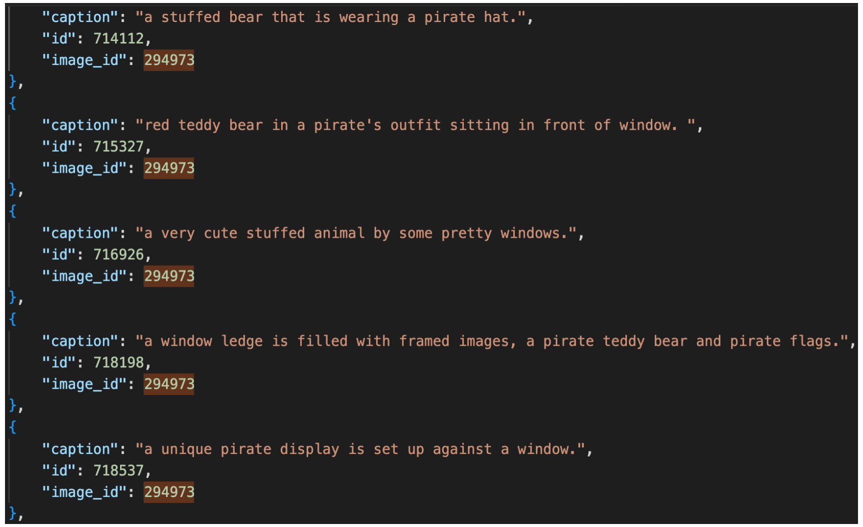
The ground truth captions



```
    "caption": "a man wearing glasses while holding a banana.",
    "id": 672106,
    "image_id": 56400
},
{
    "caption": "a man holding a banana in his hand.",
    "id": 673435,
    "image_id": 56400
},
{
    "caption": "a man is holding a banana in his hand.",
    "id": 674083,
    "image_id": 56400
},
{
    "caption": "man with glasses holding a mini banana and making a face",
    "id": 674212,
    "image_id": 56400
},
{
    "caption": "a guy in a suit holding a banana",
    "id": 674695,
    "image_id": 56400
},
```



The ground truth captions



```
    "caption": "a stuffed bear that is wearing a pirate hat.",
    "id": 714112,
    "image_id": 294973
},
{
    "caption": "red teddy bear in a pirate's outfit sitting in front of window. ",
    "id": 715327,
    "image_id": 294973
},
{
    "caption": "a very cute stuffed animal by some pretty windows.",
    "id": 716926,
    "image_id": 294973
},
{
    "caption": "a window ledge is filled with framed images, a pirate teddy bear and pirate flags.",
    "id": 718198,
    "image_id": 294973
},
{
    "caption": "a unique pirate display is set up against a window.",
    "id": 718537,
    "image_id": 294973
},
```

000000294973.jpg.

- Caption: a bunch of different types of different types of different types of different types of different types of different types of different types of different types of different types of
- CIDEr score: 0.00017149937733775526
- It’s NOT a reasonable caption (不停跳針). However, the ground truth captions are not very consistent. The description are focusing on different items in the picture. That’s why it has lower CIDEr score.
- However, same words (ex: types) are having the same attention, which implies that the attention region still reflect the corresponding word in the caption.