DLCV HW3

Problem 1

Methods analysis

Previous models, such as ResNet and VGG, are trained given a set of image and label pair. In this setting, we will face performance degradation as transfering to another distribution. And, it takes a lot of effort to collect large label data. In CLIP, they use image-caption pair which is easier to collect on online. They use large image-caption pair to learn image representation. By constrative learning, they aligh image and text distribution. Hence, they can get high quality image representation. So, CLIP is more general than previous setting and achieve better zeroshot performance.

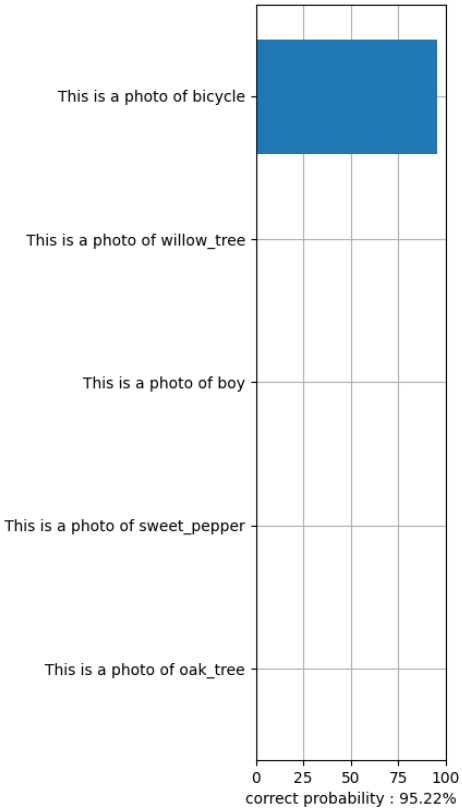
Prompt-text analysis

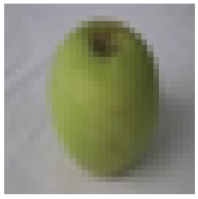
Belowing is the performance of different prompt. We can see the more simple the sentence is the higher accuracy the model get. The sentence use double negative have worst performance. Maybe it is due to the text representation is unnormal so it is hard to match correct image in this prompt.

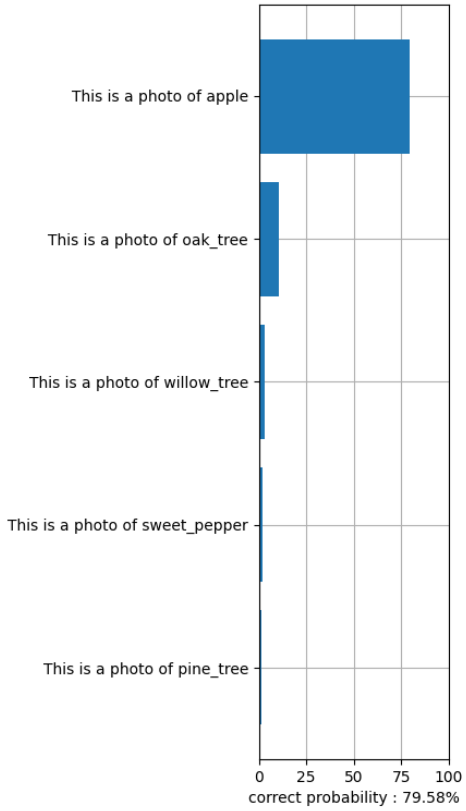
Prompt	Acc
This is a photo of {object}	60.8%
This is a {object} image	67.7%
No {object}, no score	55.0%

Quantitative analysis


This is a photo of house
This is a photo of willow_tree
This is a photo of pine_tree
This is a photo of palm_tree
This is a photo of oak_tree
0
25
50
75
100
correct probability : 60.72%


This is a photo of bicycle
This is a photo of willow_tree
This is a photo of boy
This is a photo of sweet_pepper
This is a photo of oak_tree
0
25
50
75
100
correct probability : 95.22%





Problem 2

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

model setting
transformer encoder : vit_large_patch14_224_clip_laion2b
transformer decoder layer : 6
attention head : 8
feadforward layer dimension : 2048
optimizer : Adam
learning rate : 1e-5
loss function : cross entropy loss
epoch : 20
batch size : 32
caption maximal size : 55
model performance

CIDEr	CLIPScore
89.2664	73.0220

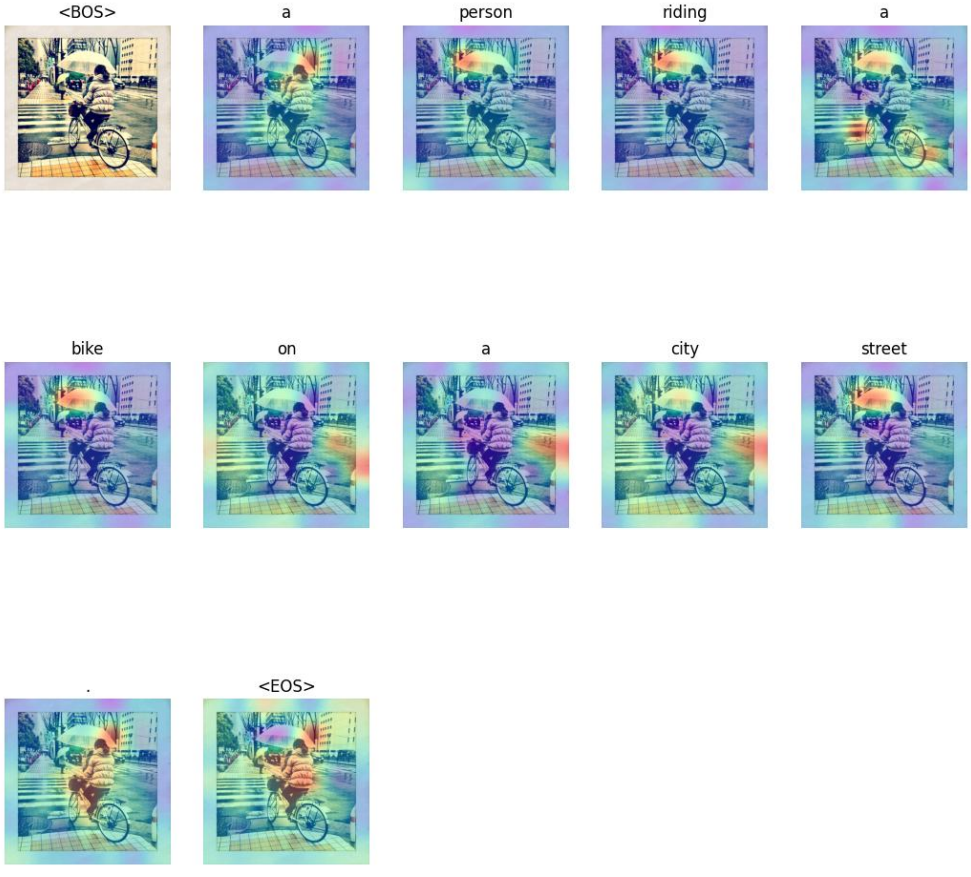
Report other 3 different attempts and their corresponding CIDEr & CLIPScore

setting	CIDEr	CLIPScore
clip base encoder	81.1227	72.2692
without freeze encoder	85.4257	71.7047
without pretraine encoder	15.4632	46.5601

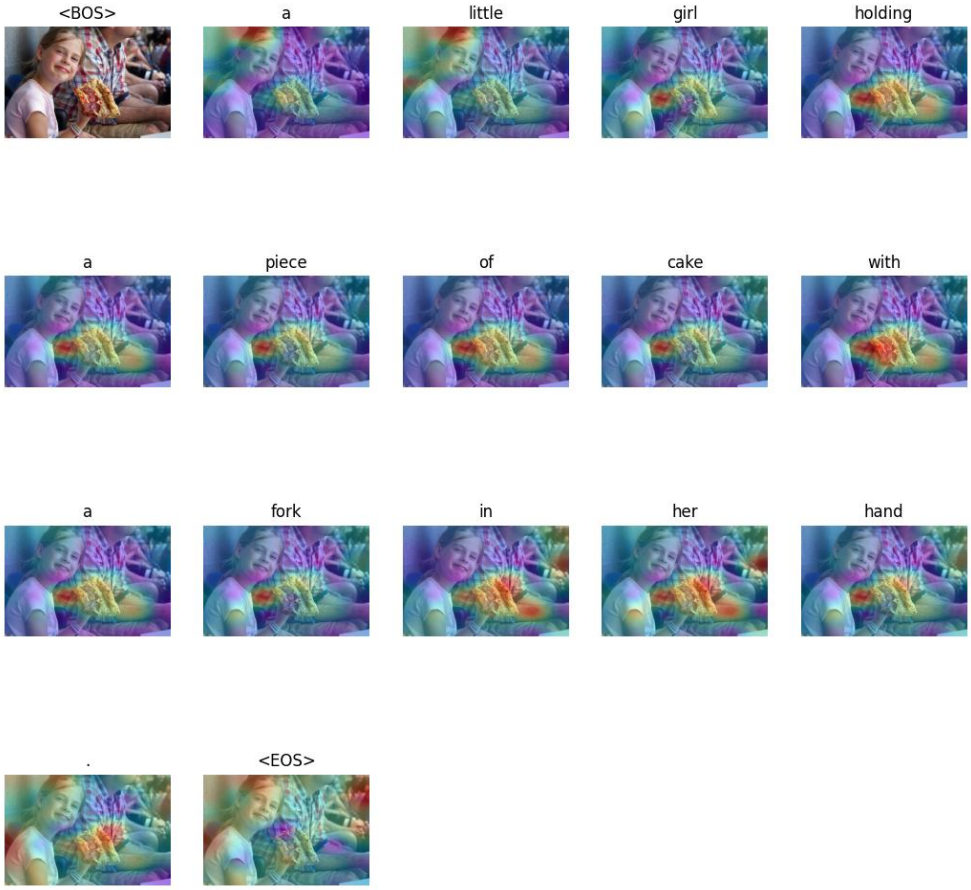
Problem 3

Please visualize the predicted caption and the corresponding series of attention maps in your report

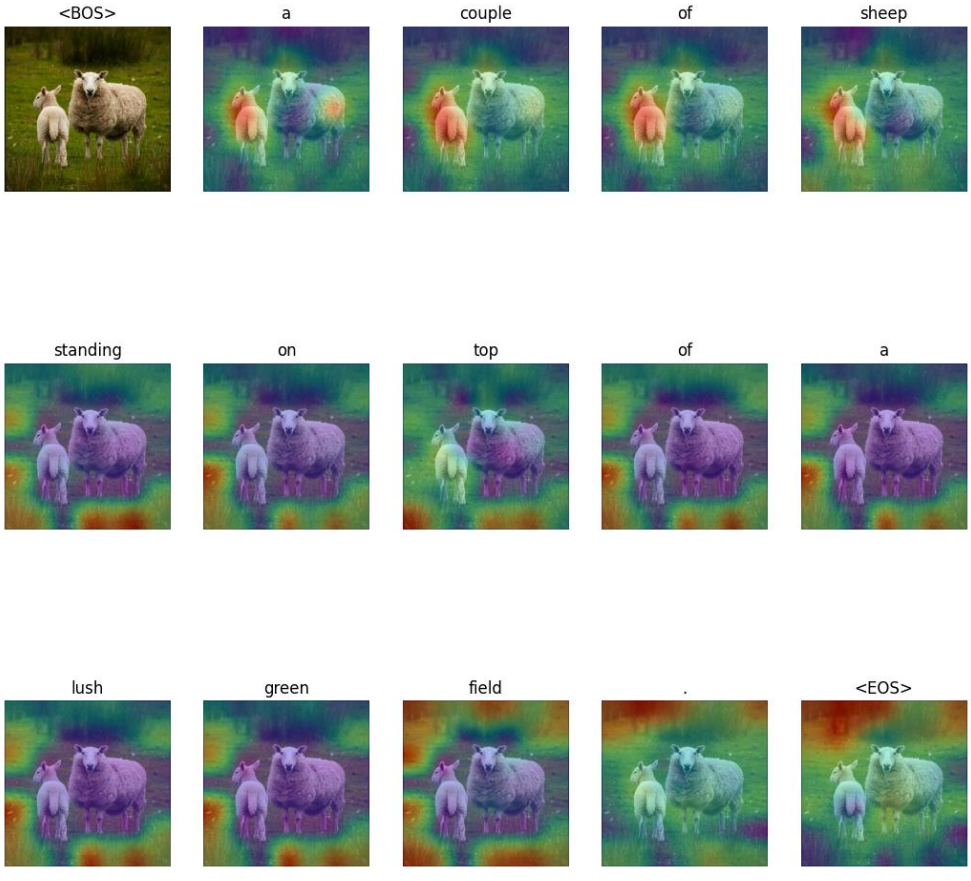
Bike



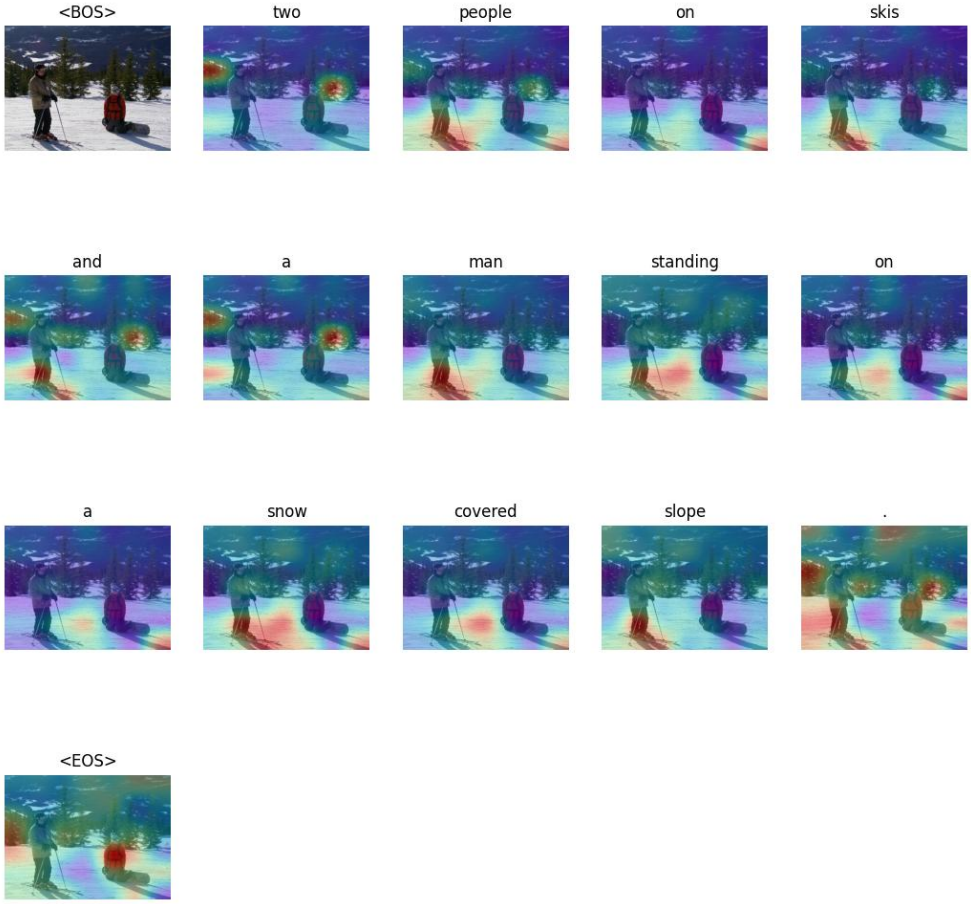
Girl


<BOS>
a
little
girl
holding
a
piece
of
cake
with
a
fork
in
her
hand
.
<EOS>

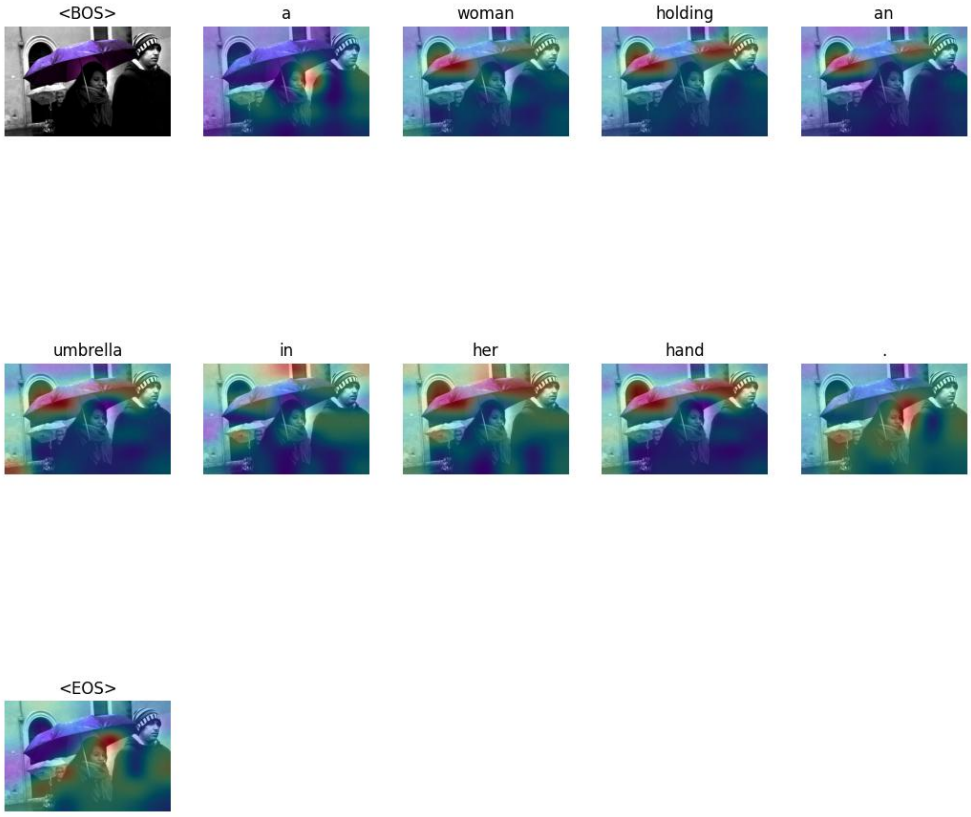
Sheep


<BOS>
a
couple
of
sheep
standing
on
top
of
a
lush
green
field
.
<EOS>

Ski



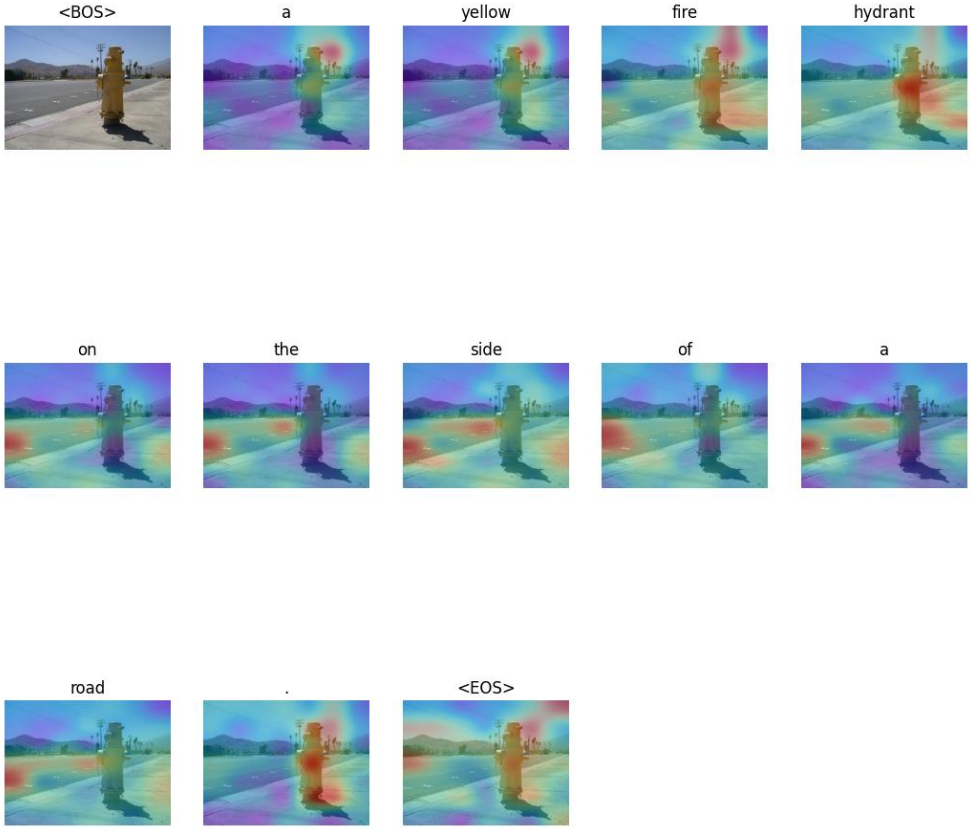
Umbrella



Visualize validation dataset of problem 2

top-1

CLIPScore : 98.44



last-1 :

CLIPScore : 36.77

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?

In top-1 image, it is reasonable. For example, in yellow and fire hydrant word the attention is high in the hydrant itself and in side and road the attention is high in road background.
In last-1 image, it is unreasonable. First, the decode caption is not like a normal sentence. Second, the decode word doesn't exist in the image. Hence, the word is not related to attention map.

Reference

Transformer :

https://pytorch.org/docs/stable/modules/torch/nn/modules/transformer.html#Transformer

(https://pytorch.org/docs/stable/modules/torch/nn/modules/transformer.html#Transformer)

Timm :

https://rwightman.github.io/pytorch-image-models/

(https://rwightman.github.io/pytorch-image-models/)

Visualize :

https://github.com/jeonsworld/ViT-pytorch

(https://github.com/jeonsworld/ViT-pytorch)