DLCV 2022 – HW3

Name : 周宇玄
Student ID : R10525104

Problem 1.:

1.

It's because the CLIP is a model which involved collecting huge custom datasets of labelled images, this approach improve the generalizability of the model.

2.

Please compare and discuss the performances of your model with the following three prompt templates:

Score :

i. *"This is a photo of {object}": 0.6076*

ii. *"This is a {object} image.": 0.682*

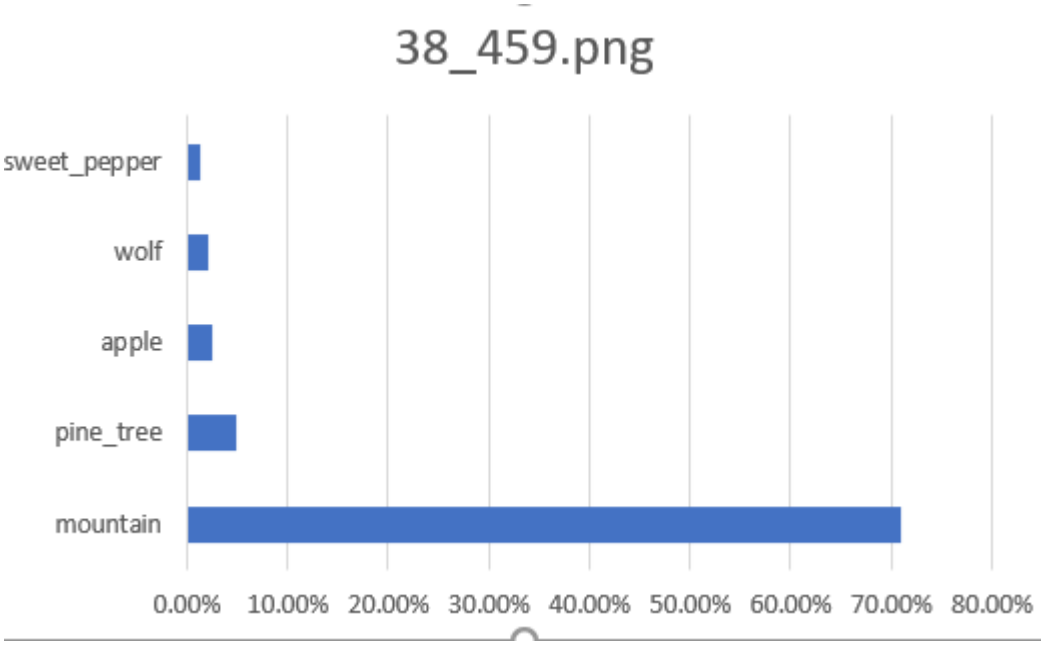
iii. *"No {object}, no score.": 0.5628*

The (ii) "This is a {object} image" get highest score of acc., I think it's because "This is a {object}" is bet more attention than "a photo of {object}", and "No {object}, no score"'s No {object} got lowest score is same reason.

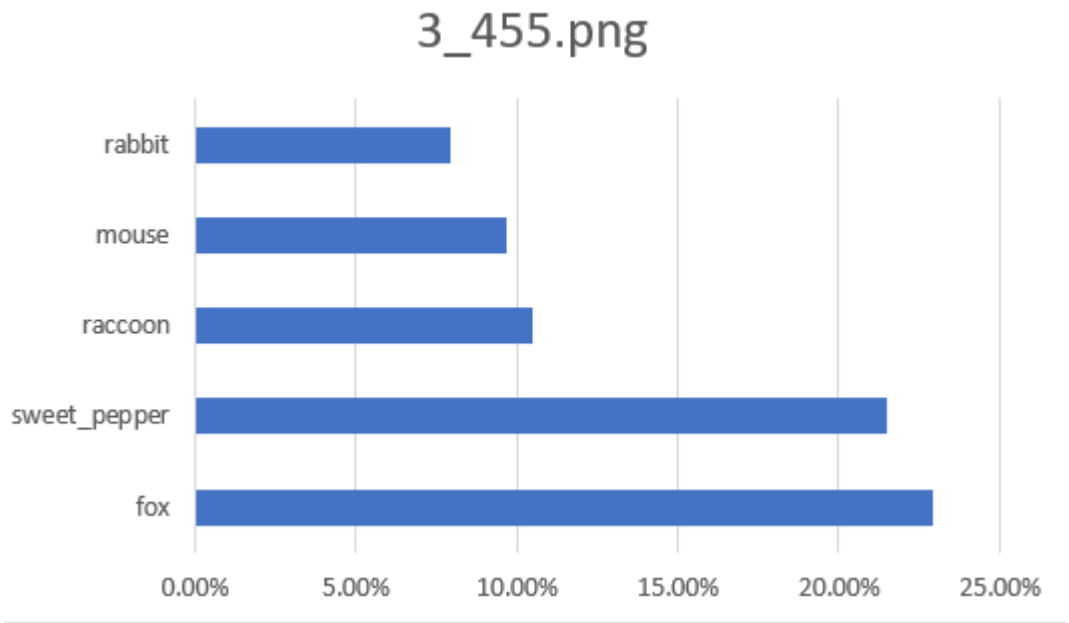
At last submit, I choose ii "This is a {object} images." be my template because it has highest score.

3.

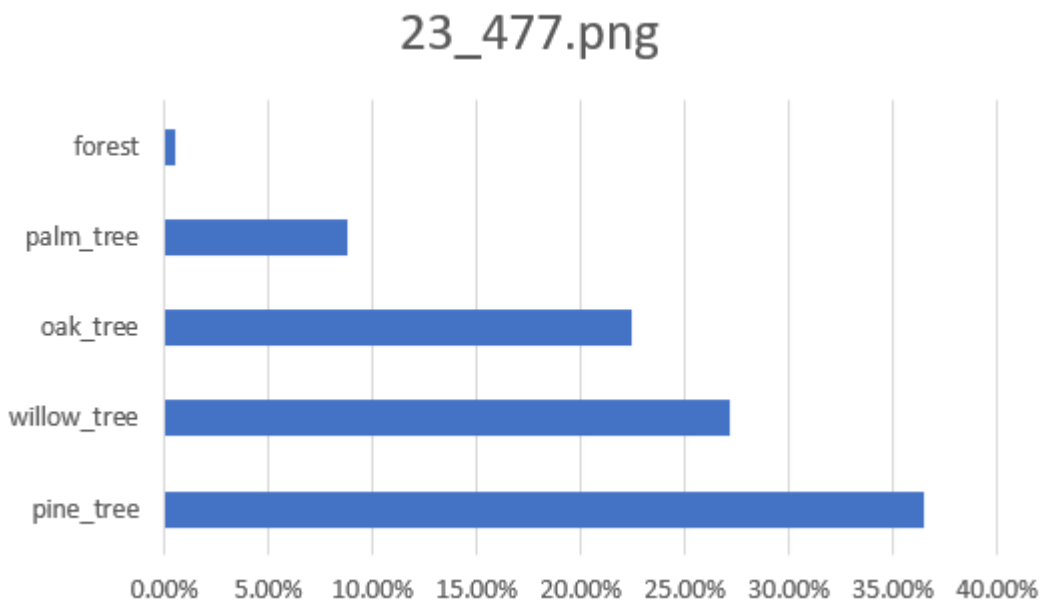
38_459.png :


38_459.png
sweet_pepper
wolf
apple
pine_tree
mountain
0.00%
10.00%
20.00%
30.00%
40.00%
50.00%
60.00%
70.00%
80.00%

3_455.png


3_455.png
rabbit
mouse
raccoon
sweet_pepper
fox
0.00%
5.00%
10.00%
15.00%
20.00%
25.00%

23_477.png


23_477.png
forest
palm_tree
oak_tree
willow_tree
pine_tree
0.00%
5.00%
10.00%
15.00%
20.00%
25.00%
30.00%
35.00%
40.00%

Problem 2.:

1.

ClDEr :1.04

CLIPScore: 0.72

My best setting is same with 2.'s First type model, it's same with the torch.nn.doules.transformer, add a CNN architecture combine with it encoder, and no pretrained encoder, which is provided on HW3-intro

2.

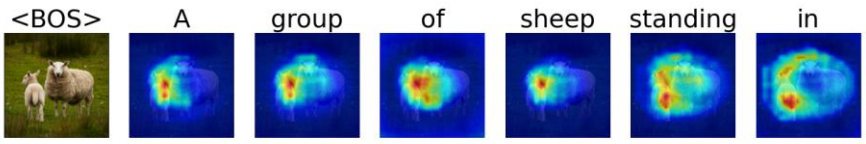
The transform architecture is basically same with the torch.nn.doules.transformer, add a CNN architecture combine with it encoder, and no pretrained encoder, which is provided on HW3-intro, the different of three attempts is the first one I use original architecture just like intro provided, second I add encoder and decoder each two layers, third I use second architecture but use dropout 0.2(original is 0.1), their CIDEr & CLIPScore is show on below table:

	First	Second	Third
CIDEr	1.04	0.84	0.82
CLIPScore	0.72	0.66	0.65

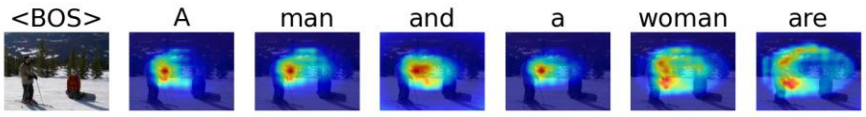
We can see that the original architecture has best CIDEr and CLIPScore, I think the reason is after I add two layers on both encoder and decoder, make the model become too deep and didn't choose good parameter for it.

Problem 3.:

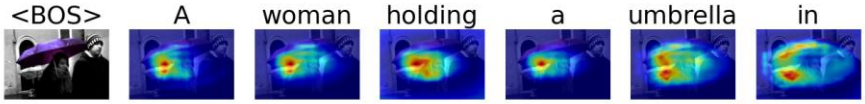
1.

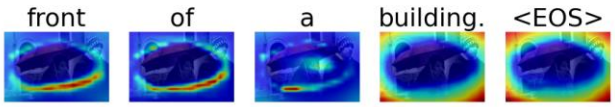

<BOS> A group of sheep standing in

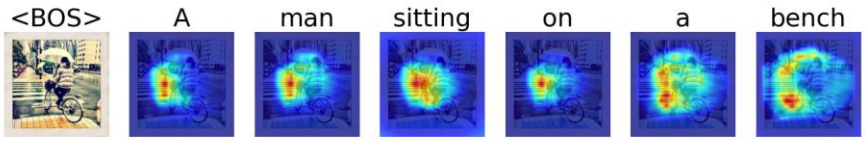

a field. <EOS>

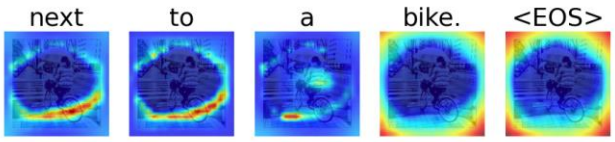

<BOS> A man and a woman are

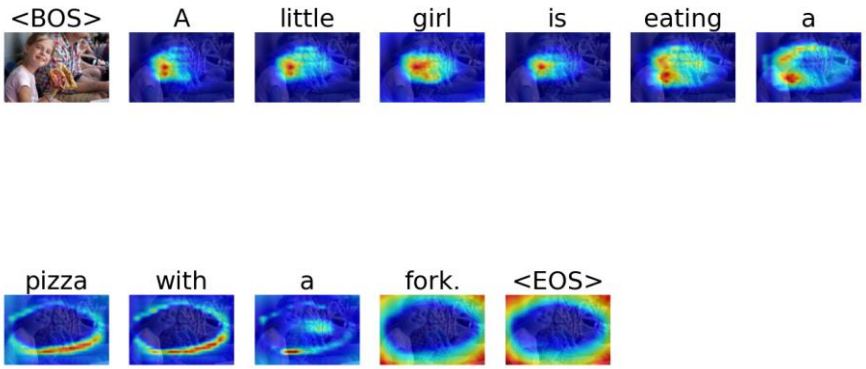

standing in the snow. <EOS>


<BOS>
A
woman
holding
a
umbrella
in


front
of
a
building.
<EOS>


<BOS>
A
man
sitting
on
a
bench

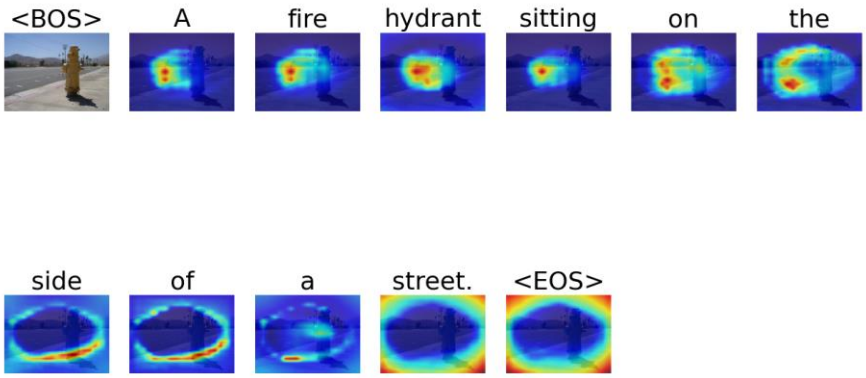

next
to
a
bike.
<EOS>



2.

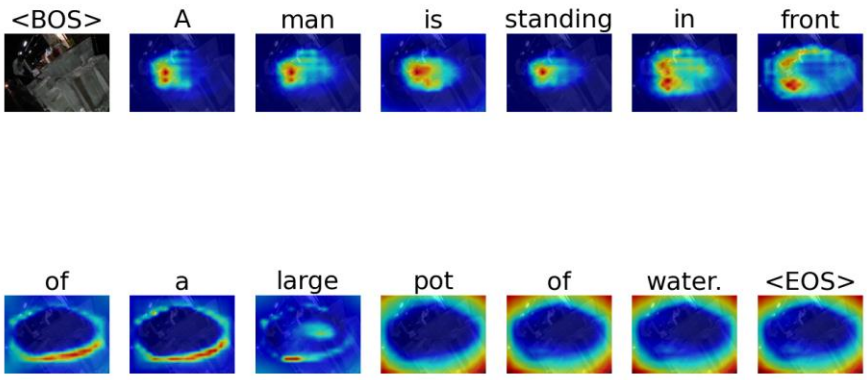
Top-1 CLIPscore: 0.999

Top-1 image : 000000392315



Last-1 CLIPScore : 0.373

Last-1 image : 461413605



3.

I think the attended region reflect the corresponding word in the caption and it make caption reasonable, for example, Last-1 image's generated caption is "A man is standing in a large pot of water", "A man is standing" parts's attended region is on the man, and "in front" is around the man, "of a large pot of water" is basically specific on the background, I think this shows attended region reflect the corresponding word in the caption, and make caption reasonable for this attended region even if it is not true.

Reference :

Hw3_intro

P1 :

https://github.com/openai/CLIP

P2 :

https://github.com/openai/CLIP
https://github.com/saahiluppal/catr
https://github.com/huggingface/tokenizers
https://github.com/bckim92/language-evaluation