DLCV hw3
B09901062 黃宥翔

Problem1

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.

Ans: **Because CLIP is trained on a wide variety of images with a wide variety of natural language supervision that's abundantly available on the internet. The network can be instructed in natural language to perform a great variety of classification benchmarks, without directly optimizing for the benchmark's performance(similar to zero-shot). The key is "not directly optimizing for the benchmark"**

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}": 0.5784
    - "This is a {object} image.": 0.6832
    - "No {object}, no score.": 0.562

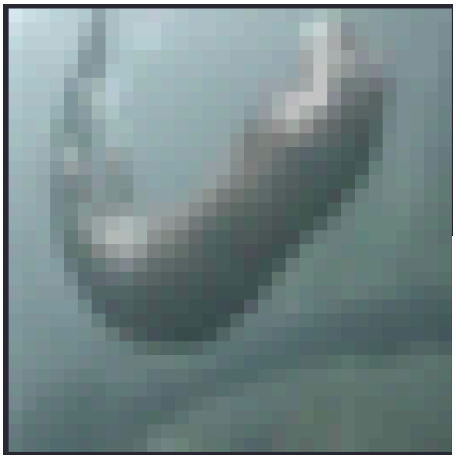
3. Quantitative analysis (6%)

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

```
Top predictions of ('10_462.png',):

     willow_tree: 32.47%
        oak_tree: 13.53%
    sweet_pepper: 11.94%
       pine_tree: 11.22%
       palm_tree: 4.82%
```

Ans: baby



```
Top predictions of ('47_484.png',):

         dolphin: 75.44%
           worm: 7.35%
        elephant: 3.42%
           otter: 3.11%
            baby: 1.75%
```

Ans: otter

```
Top predictions of ('40_493.png',):

           couch: 62.50%
      willow_tree: 8.86%
           chair: 5.55%
       pine_tree: 5.13%
        oak_tree: 3.26%
```

Ans: couch

Problem2

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

   Ans: Best setting: encoder using *vit_large_r50_s32_384*

   *Score: CIDEr: 0.934278617992512 | CLIPScore: 0.7123361569399804*

   Decoder layers: 6 / Head: 12

   Max length padding: 55

   Decoder feature size: 768

   Decoder feed forward dim: 2048

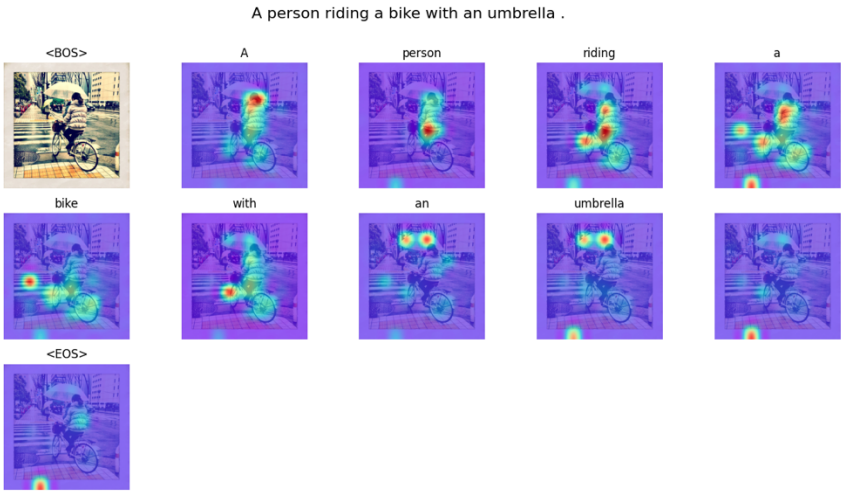
   Train around 30 epoches

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%)

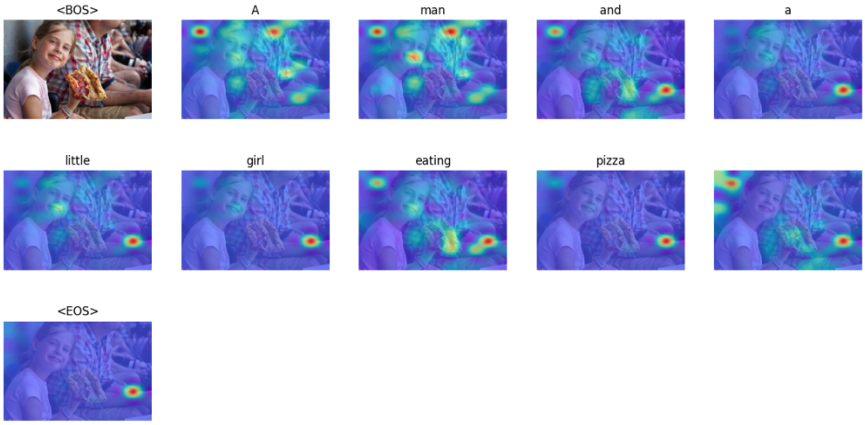
    1. Smaller head / Smaller feature size / Smaller encoder model / inference without beam search : [ CLIP score: 0.6699 | CIDEr score 0.6940 ]
    2. Bigger head / Bigger feature size / Smaller encoder model / inference without beam search : [ CLIP score: 0.6517 | CIDEr score 0.7119 ]
    3. Bigger head / Bigger feature size / Bigger encoder model / inference without beam search : [ CLIP: 0.6884 | current best CIDEr: 0.8657]

Problem3

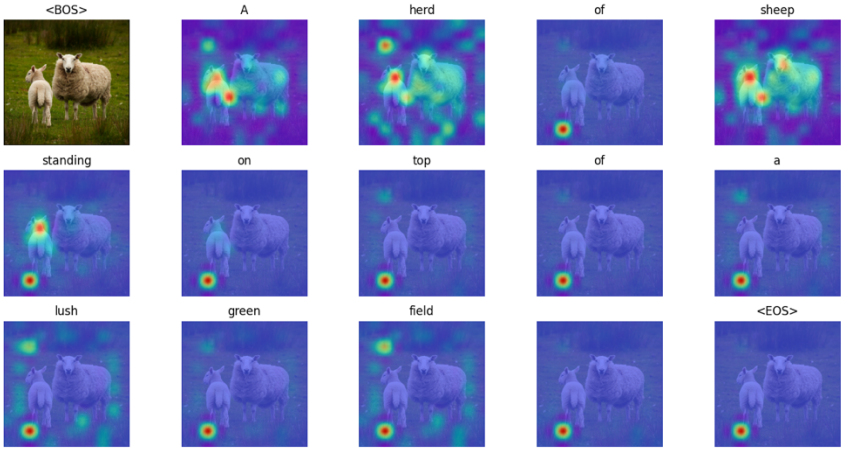
1. TA will give you five test images ([p3_data/images/]), and please visualize the **predicted caption** and the corresponding series of **attention maps** in your report with the following template: (10%, each image for 2%)



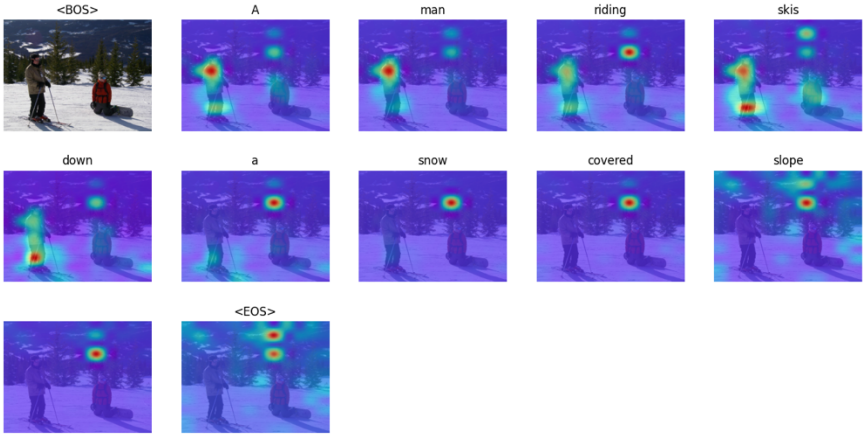
A man and a little girl eating pizza .



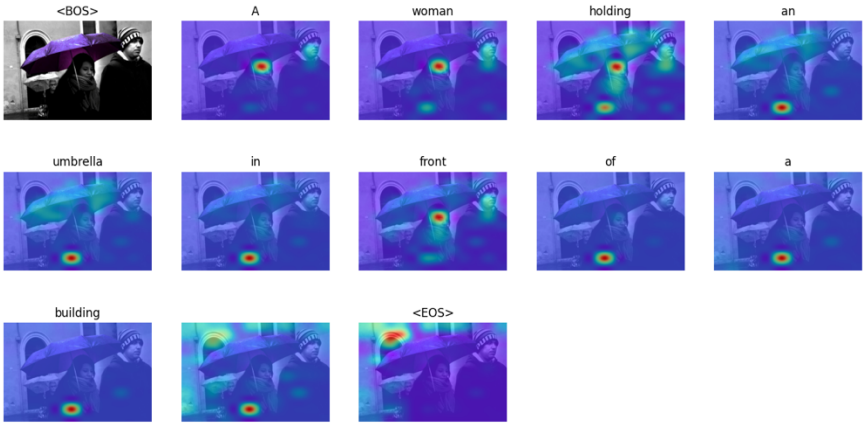

A herd of sheep standing on top of a lush green field .



A man riding skis down a snow covered slope .




A woman holding an umbrella in front of a building .

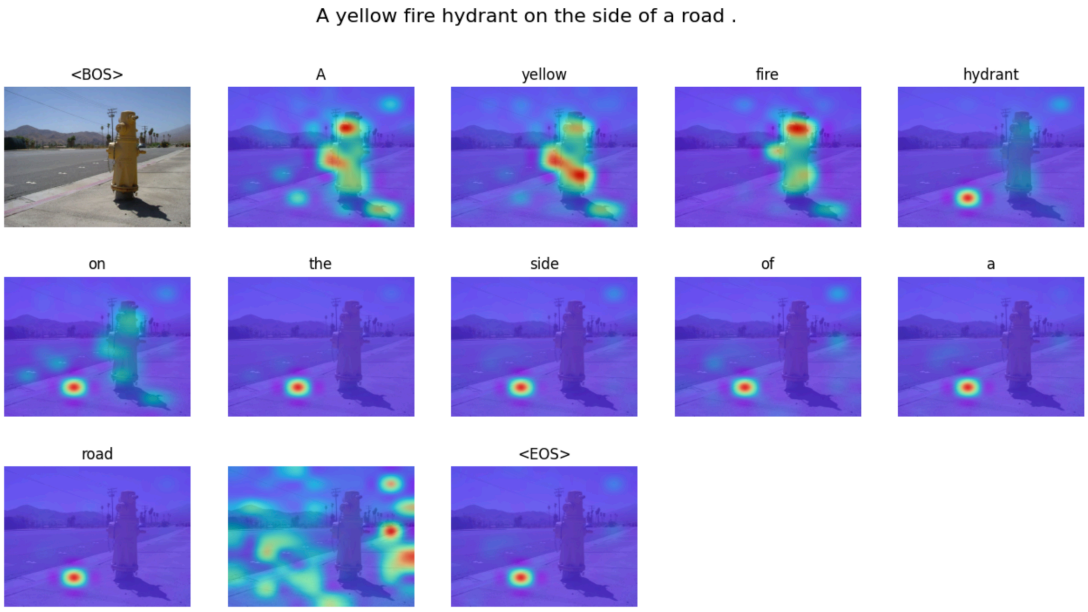


2. According to **CLIPScore**, you need to visualize:
    1. top-1 and last-1 image-caption pairs
    2. its corresponding CLIPScore

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

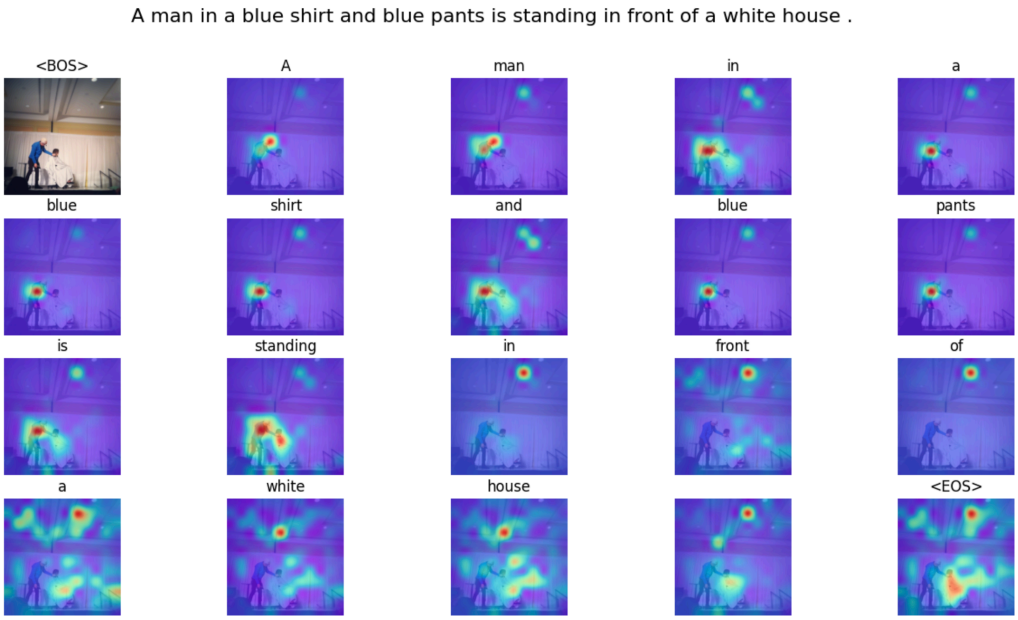
3. Analyze the predicted captions and the attention maps for each word according to the previous question. Is the caption reasonable? Does the attended region reflect the corresponding word in the caption? (5%)

- Best: 000000392315.jpg
    - CLIP score: 0.984375




I think the caption is very reasonable but the attention map of the last part of the sentence is a little unprecise. At the word “road”, the attention map only shows focus on a dot.

- Worst: 000000302838.jpg
  - CLIP score: 0.3037109375




The caption is mostly wrong. It isn't a man in front of a white house. It's a man putting on a show. Although the white house part is incorrect, the attention map pretty much recognize the important part of the image