Homework #3

Deep Learning for Computer Vision

NTU, Fall 2023

周奕節
r10943131

Problem 1: Zero-shot Image Classification with CLIP

1. Methods analysis (3%)
    - CLIP is trained on not only images but also their captions, and optimized by taking image-text pairs and pushing their output vectors nearer in vector space while separating the vectors of non-pairs.
    - We can design prompt-text by ourselves to make better descriptions of the images

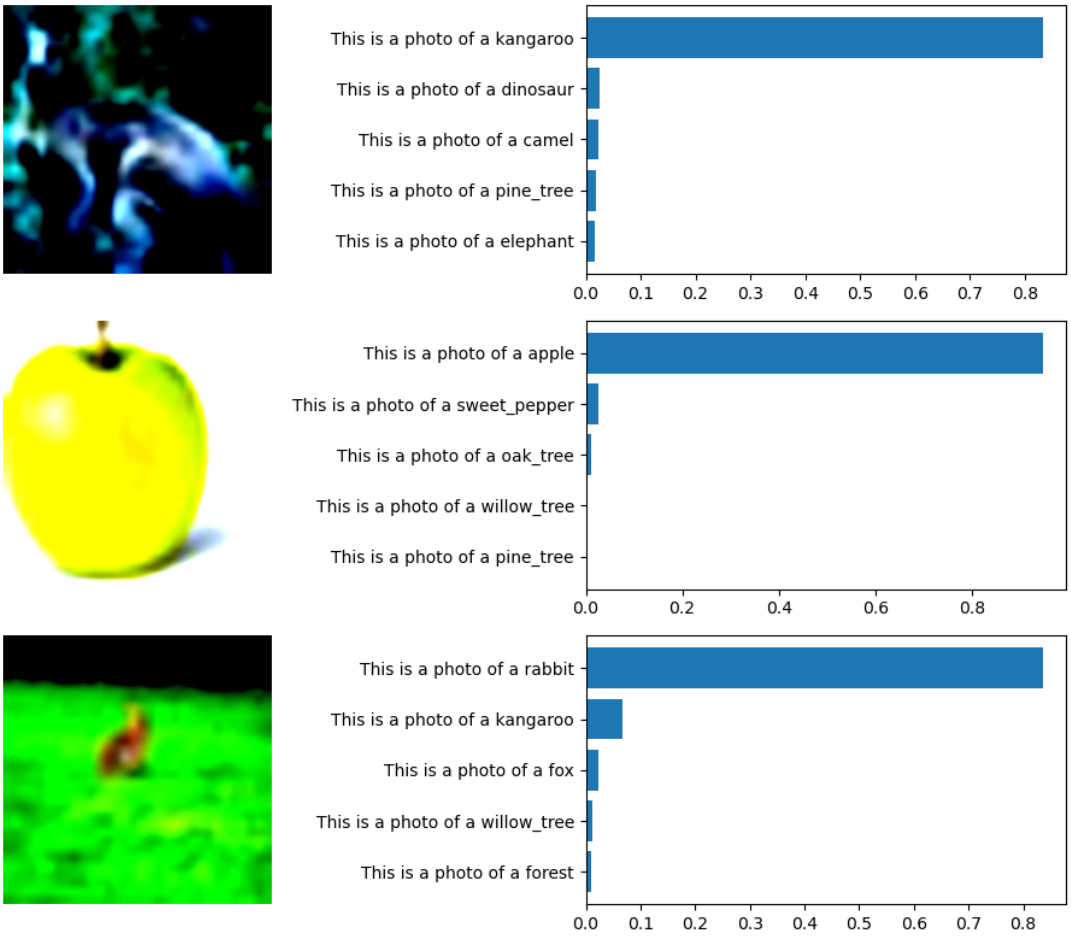
2. Prompt-text analysis (6%)

Prompt	Accuracy
This is a photo if {object}	0.687
This is not a photo of {object}	0.698
No {object} no score	0.550

a. This is a photo if {object}: This is a general and straight-foward promt, which clearly describe the object so can get a descent accuracy
b. This is not a photo of {object}: We can found that adding a negative prompt enhances the score, I tried “No! This is not a photo if {object}” and got an even higher score, thus we can find that negative and strong words have good influence to the performance.
c. No {object} no score: It is an irrelevant sentence with non-sence, thus getting a bad score.

3. Quantitative analysis (6%)

- For the second and third images, it predicted correctly, however for the first image I can’t recognize what object is it.



Problem 2: PEFT on Vision and Language Model for Image Captioning

Evaluation metrics report (10%)

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

	CIDEr	CLIPScore
Best Setting	0.84	0.72

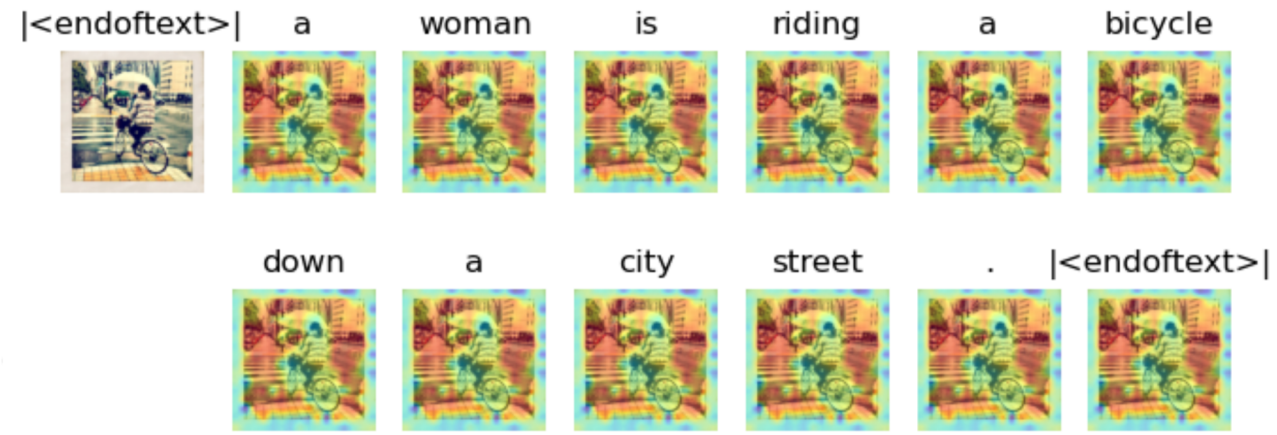
- Report 3 different attempts of PEFT and their corresponding CIDEr & CLIPScore.

	CIDEr	CLIPScore
Adapter	0.71	0.69
Prefix Tuning	0.48	0.69
Lora	0.84	0.72

Problem 3:Visualization of Attention in Image Captioning

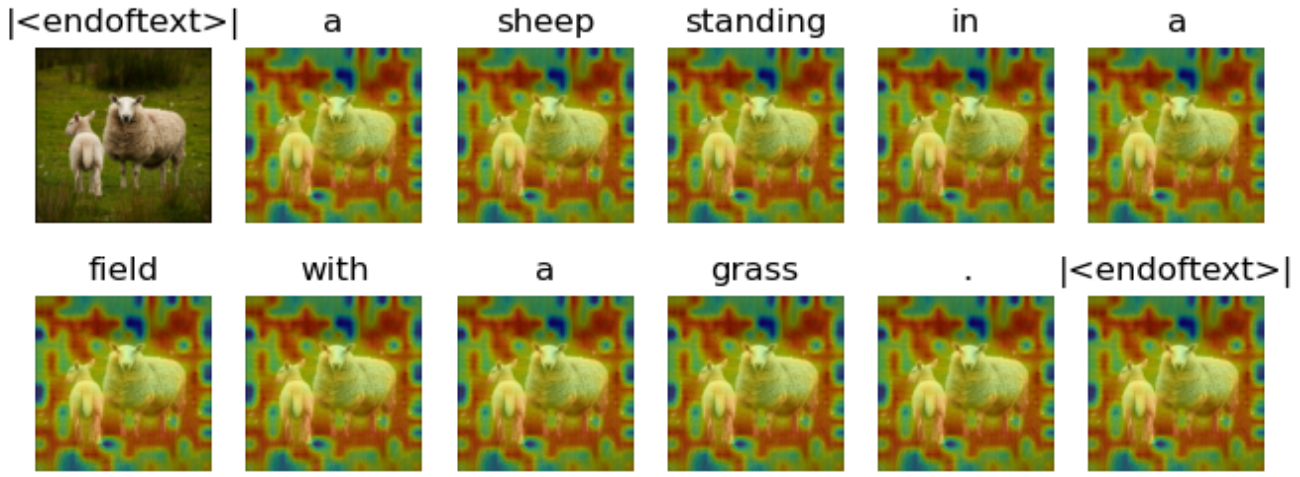
1. Visualize the **predicted caption** and the corresponding series of **attention maps**

bike.jpg

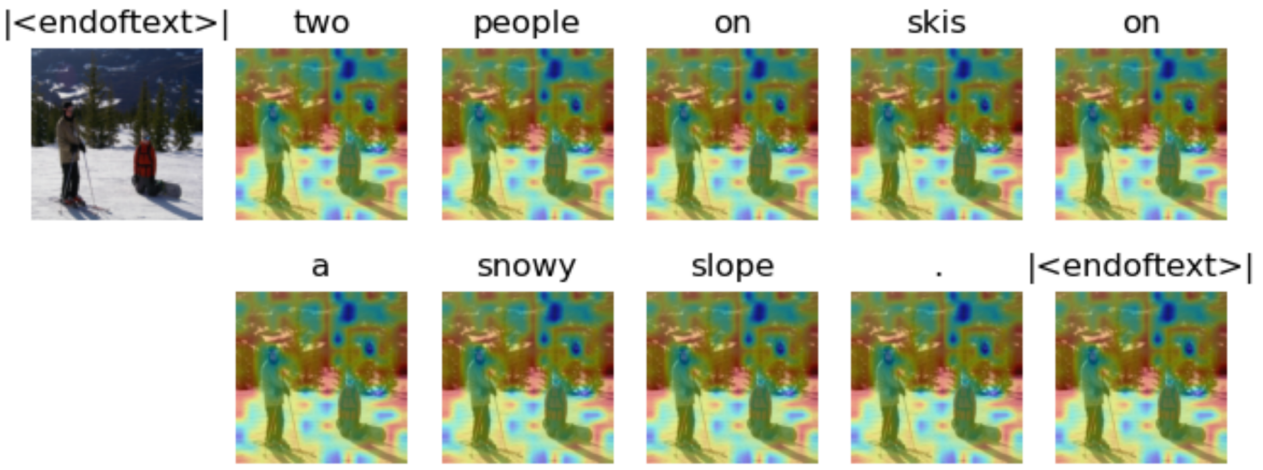


girl.jpg

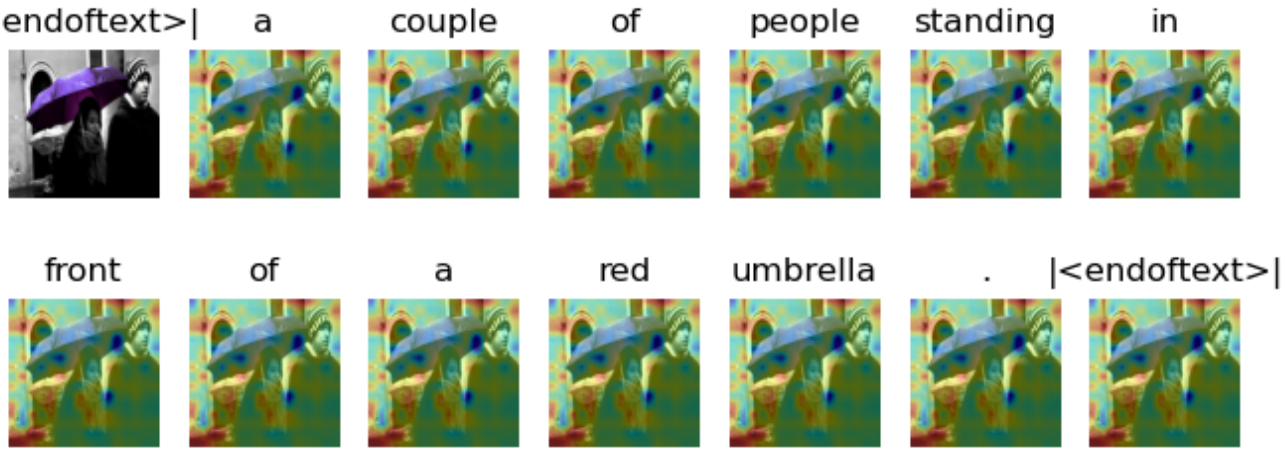
sheep.jpg



ski.jpg



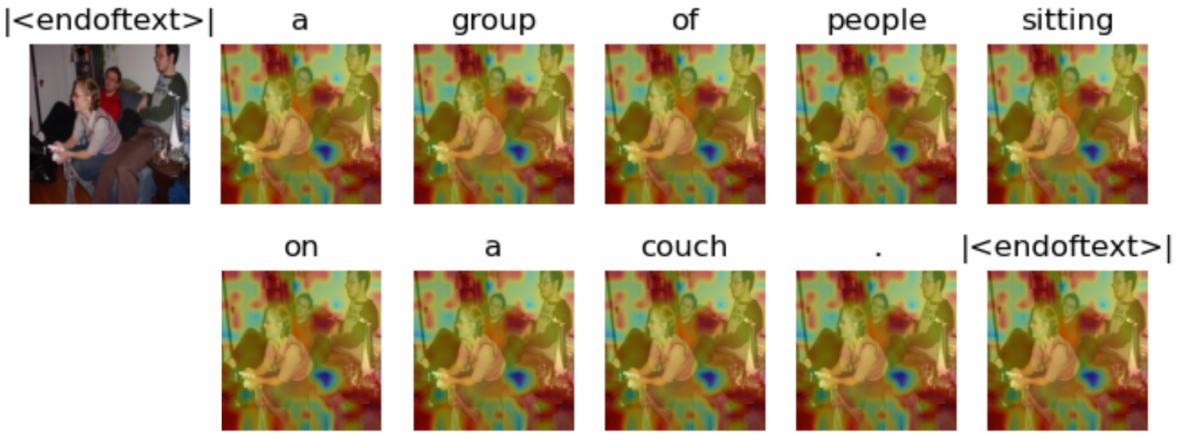
umbrella.jpg



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

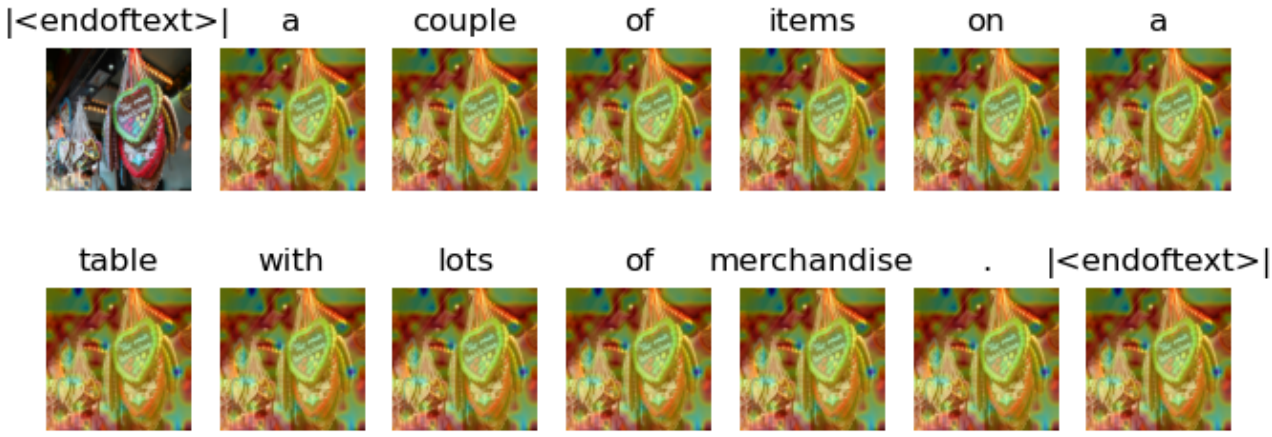
Top-1: 000000034708.jpg

CLIPScore: 0.983



Last-1: 000000562675.jpg

CLIPScore: 0.405



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%)

The captions are reasonable since it can tell most of the objects in the image. For example, bike.jpg: woman, bicycle, city street. Also it can tell the verb in the image, like riding, holding, standing…

For the region, because I chose a giagantic encoder model, the output dimension is 1664, when applying to decoder we need to add a linear transformation to turn it to 768, thus the effect of visualization isn't that obvious.Unfortunately, the difference between Top-1 and Last-1 image aren't obvious either.