

## Problem 1

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

## i、 Joint image and text understanding:

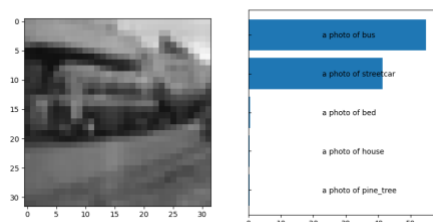
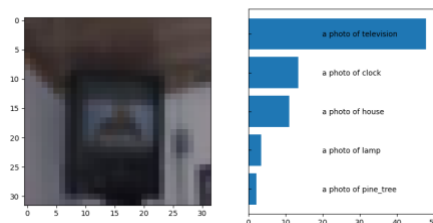
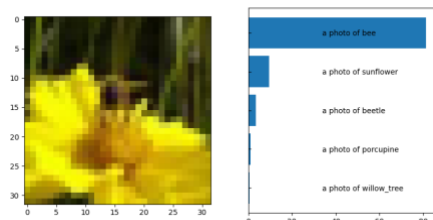
CLIP's training on a massive dataset using contrastive learning approach. It develops a nuanced understanding of images by pairing images with the rich and descriptive nature of the text.

## ii、 Diversity of dataset:

The training data for CLIP is diverse and extensive, covering a wide range of subjects and contexts. This broad exposure allows the model to develop a better understanding of different types of images and their associated textual descriptions.

**2. Prompt-text analysis**

“This is a photo of {object}”	0.74
“This is not a photo of {object}”	0.67
“No {object}, no score.”	0.49

**3. Quantitative analysis**

## Problem 2

Reference: <https://www.kaggle.com/code/shreydan/visiongpt2-image-captioning-pytorch#Training>

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

Setting:

No PEFT, only pretrained vit model: vit\_large\_patch14\_224\_clip\_laion2b

Adding a self-written cross-attention

Only train cross-attention part and output\_reshape parameters.

Score:

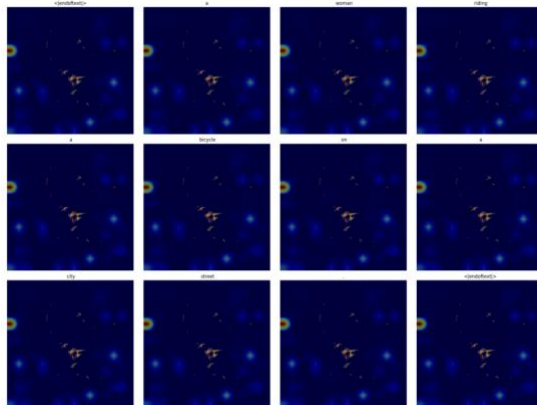
CIDEr: 0.8375

CLIPScore: 0.724

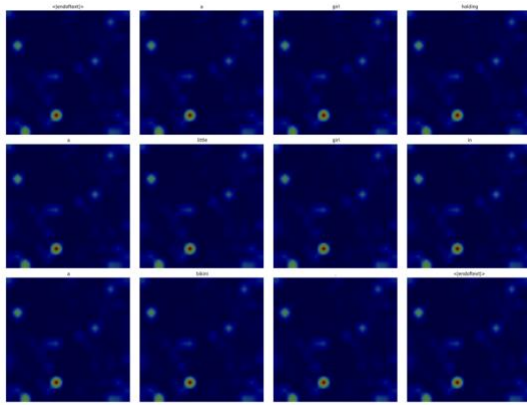
2. Report 3 different attempts of PEFT and their corresponding CIDEr & CLIPScore. (7.5%, each setting for 2.5%)

	CIDEr	CLIPScore
lora	0.67	0.71

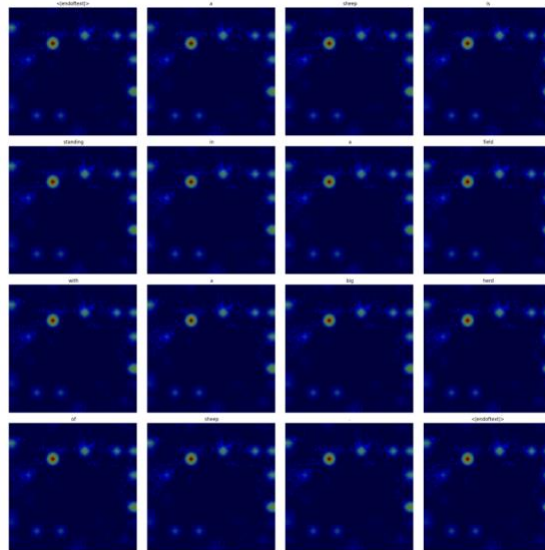
1. TA will give you five test images ([p3\_data/images/]), and please visualize the predicted caption and the corresponding series of attention maps in your report with the following template: (10%, each image for 2%)



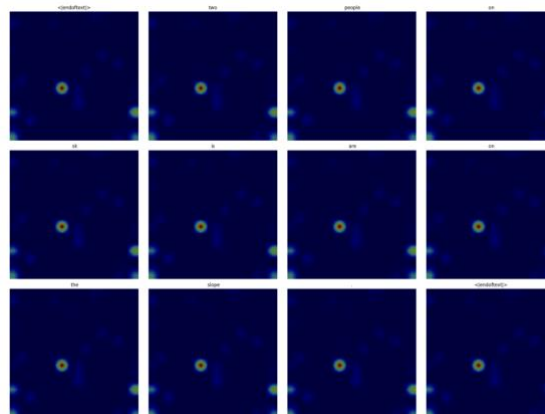
bike.jpg



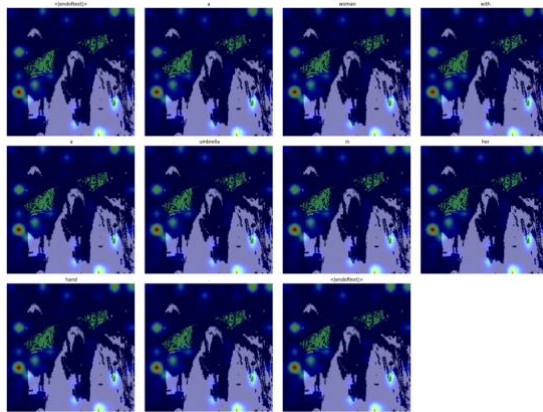
girl.jpg



sheep.jpg




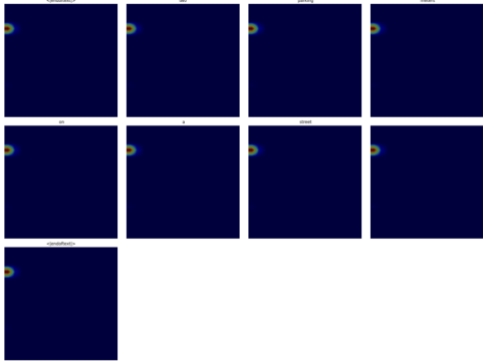

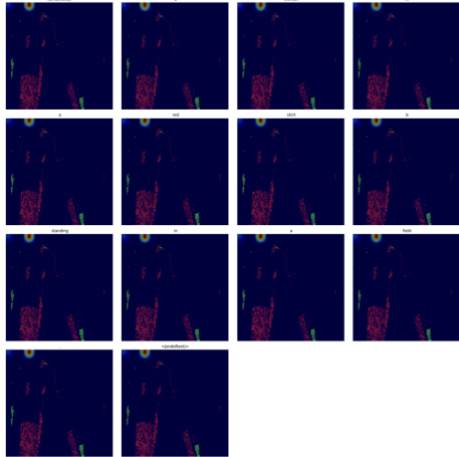
ski.jpg



umbrella.jpg

2. According to CLIPScore, you need to:

- i、 visualize top-1 and last-1 image-caption pairs
- ii、 report its corresponding CLIPScore

<p><b>Top-1</b></p> <p><b>Image:</b>000000013414.jpg</p> <p><b>CLIPScore:</b> 0.95</p> <p><b>Caption:</b> two parking meters on a street</p>  	<p><b>Last-1</b></p> <p><b>Image:</b> 4927180699.jpg</p> <p><b>CLIPScore:</b> 0.35</p> <p><b>Caption:</b> a woman in a red shirt is standing in a field .</p>  
--	--

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 caption works, but the detail in the maps showing where words focus isn't great. This might be because we only used a small number of pictures (about 50,000) to train the system, making it hard to get really accurate maps.