Student ID: R10625016
Name: 許致銓
Department: 森林所

Problem1: Zero-shot image classification with CLIP

1. Methods analysis (3%)

- **Previous methods (e.g. VGG and ResNet) are good at one task and one task only, and require 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.**

  I think the main difference between CLIP and VGG, ResNet, and other previous methods is the attention block. For image features, CLIP calculates the attention scores for each patch and token pair. In this way, CLIP can learn better about how the target object and token correspond on a larger scale, which is similar to so-called in-context learning. Moreover, the pretrained model is so important for the success of zero-shot. Without a proper pretrained dataset, I don't think CLIP can be better than VGG or ResNet.

2. Prompt-text analysis (6%)

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

  1. ***"This is a photo of {object}"***
  2. ***"This is not a photo of {object}"***
  3. ***"No {object}, no score."***

  I used ViT-L/14 as my backbone. Compared to my prompt "It is a real {object}", I think the prompt as natural language also matters. "Real" is a confirmed word for an absolute tone, and the model can have a better performance with the prompt.

Prompt	Accuracy
1	0.68
2	0.70

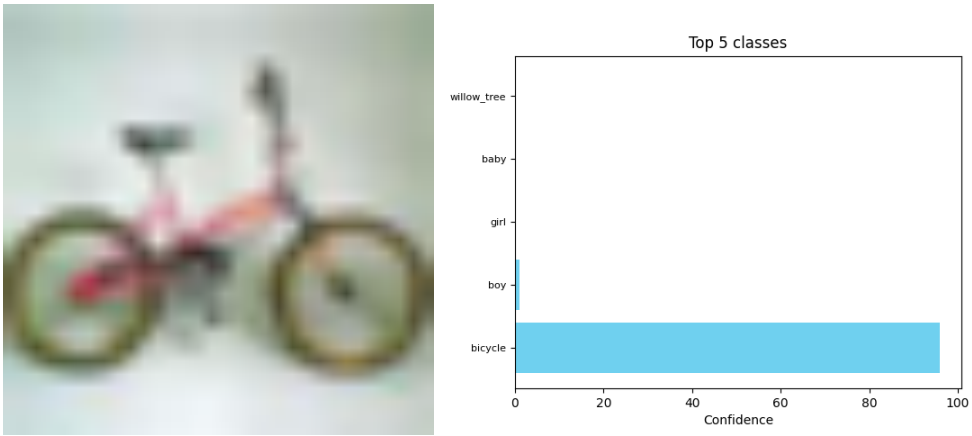
3	0.46

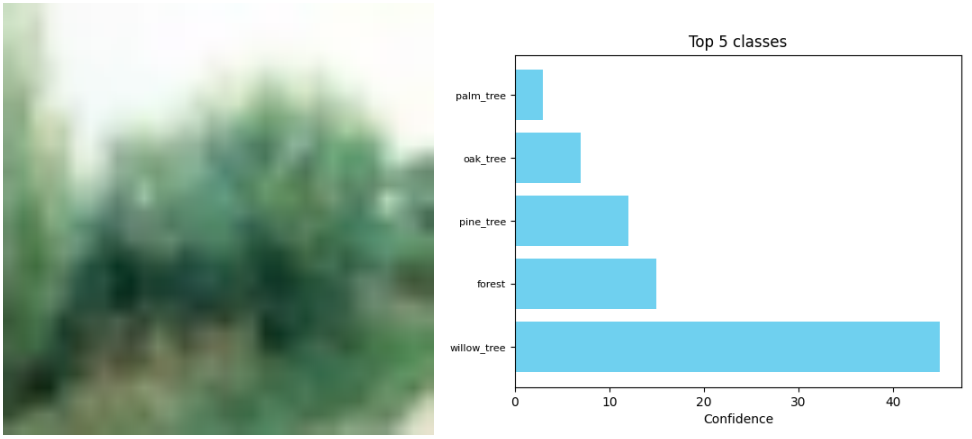
3. Quantitative analysis (6%)

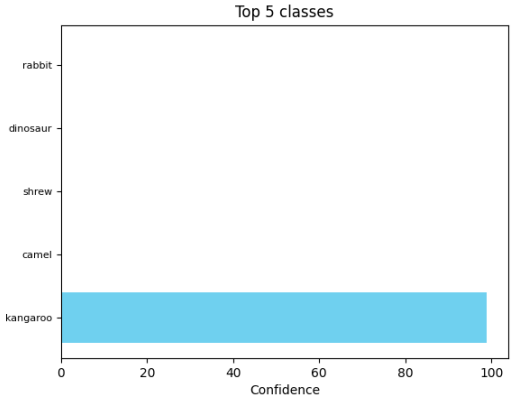
- **Please sample three images from the validation dataset and then visualize the probability of the top-5 similarity scores as the following example:**

  I asked the TA for advice about the format of the figure. He said it is OK to plot as below.

  “It is a real {object}”







- **Reference**
  1. https://towardsdatascience.com/all-you-need-to-know-about-in-context-learning-55bde1180610

Problem2: PEFT on Vision and Language Model for Image Captioning

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

I used “vit_large_patch14_clip_336.openai_ft_in12k_in1k” as my pretrained encoder. Greedy is the auto-regressive method I designed.

Model Configuration:

```
• Optimizer: AdamW
• 'n_epochs': 10
• 'batch_size': 32
• 'lr': 5e-4
• 'weight_decay': 1e-2
• 'max_tokens': 60
```

Method	Best CIDEr	Best CLIP
Prefix-Tuning	0.9265	0.7368

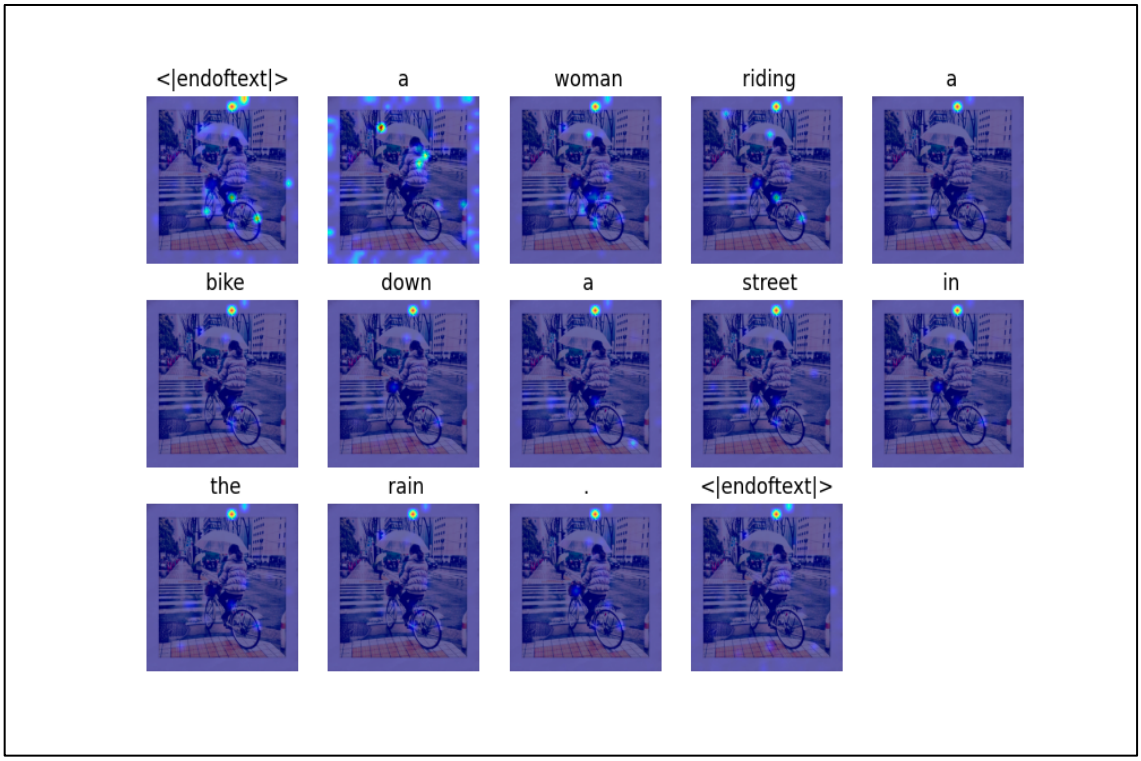
2. Report 3 different attempts of PEFT and their corresponding CIDEr & CLIPScore. (7.5%, each setting for 2.5%)

Method	Best CIDEr	Best CLIP

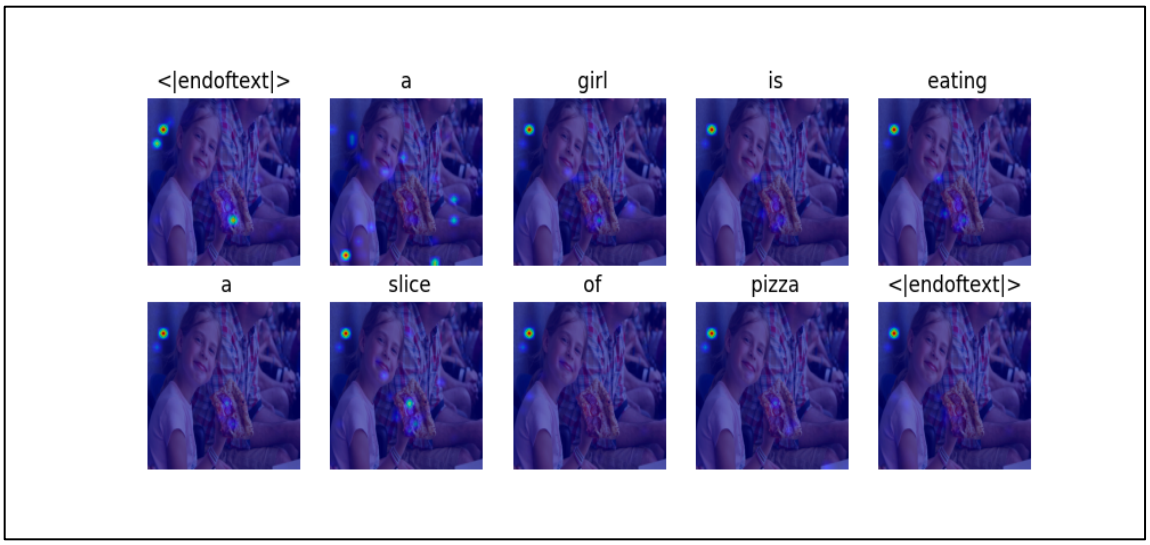
Adapter	**0.9215**	**0.7388**
Prefix-Tuning	**0.9265**	**0.7368**
LORA	**0.6283**	**0.6798**

3. **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



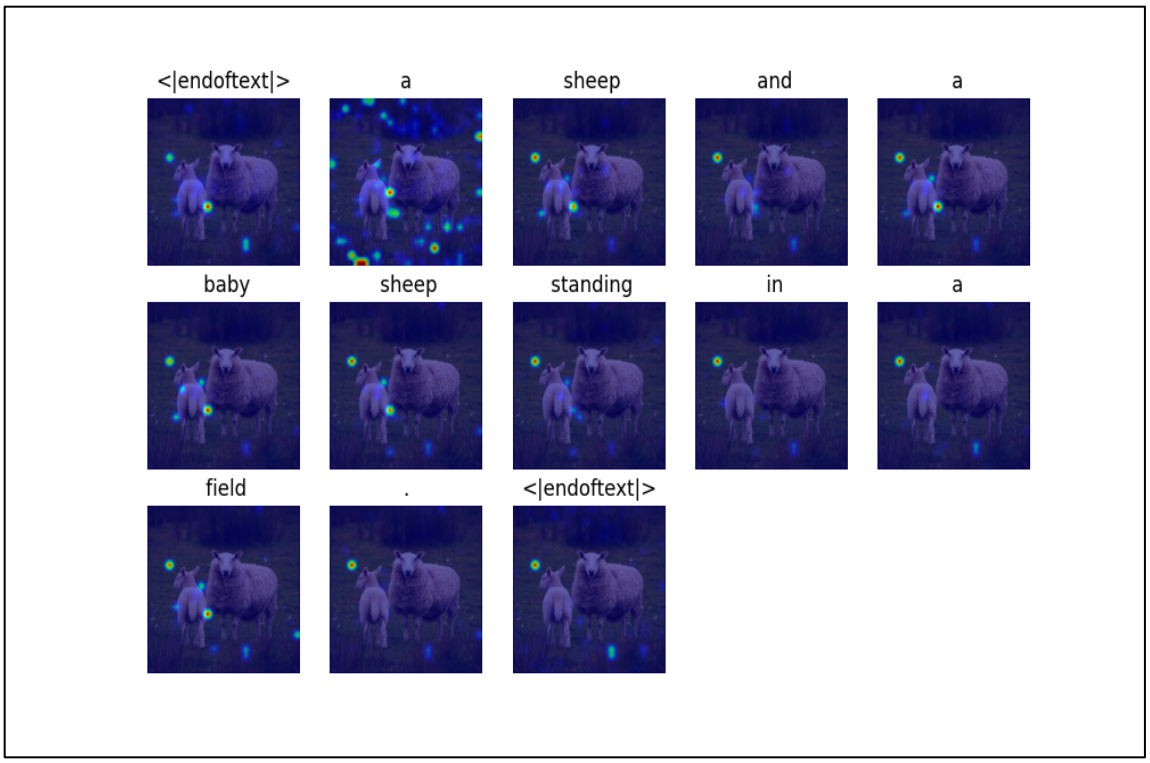
Girl


<|endoftext|>
a
girl
is
eating
a
slice
of
pizza
<|endoftext|>

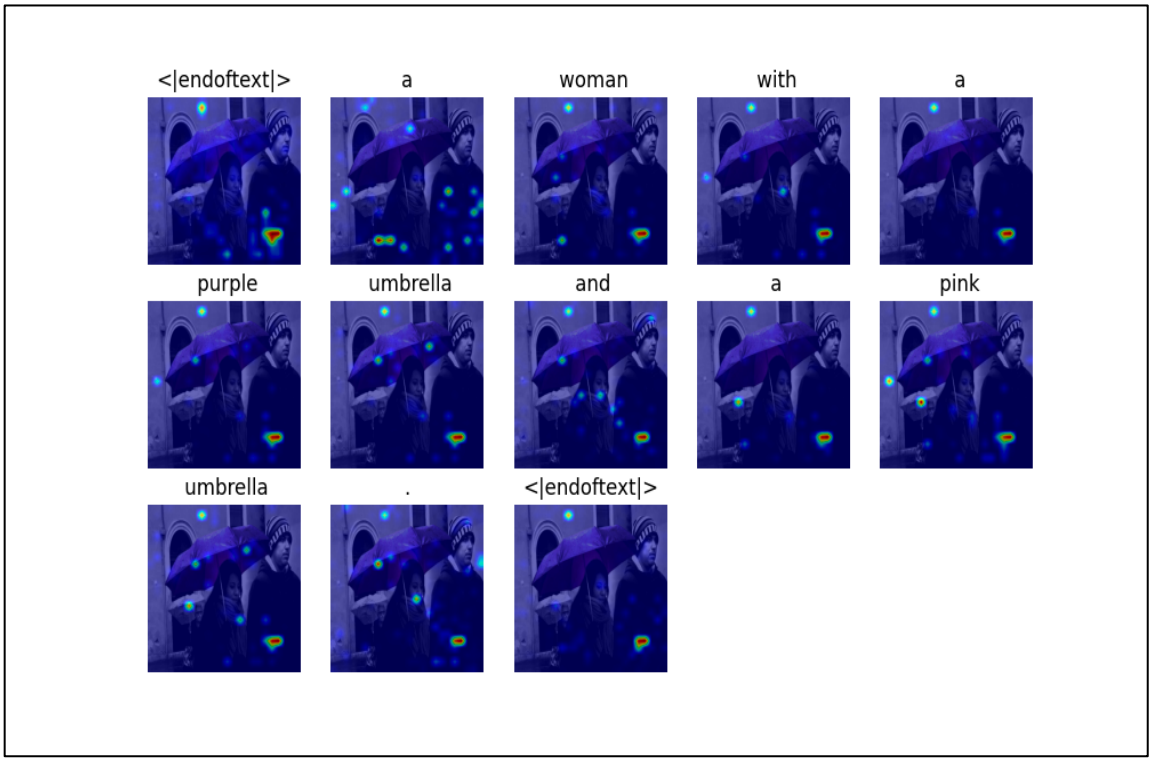
Ski


<|endoftext|>
a
man
and
a
woman
on
sk
is
standing
on
a
snow
covered
slope
.
<|endoftext|>

Sheep


<|endoftext|>
a
sheep
and
a
baby
sheep
standing
in
a
field
.
<|endoftext|>

Umbrella


<|endoftext|>
a
woman
with
a
purple
umbrella
and
a
pink
umbrella
.
<|endoftext|>

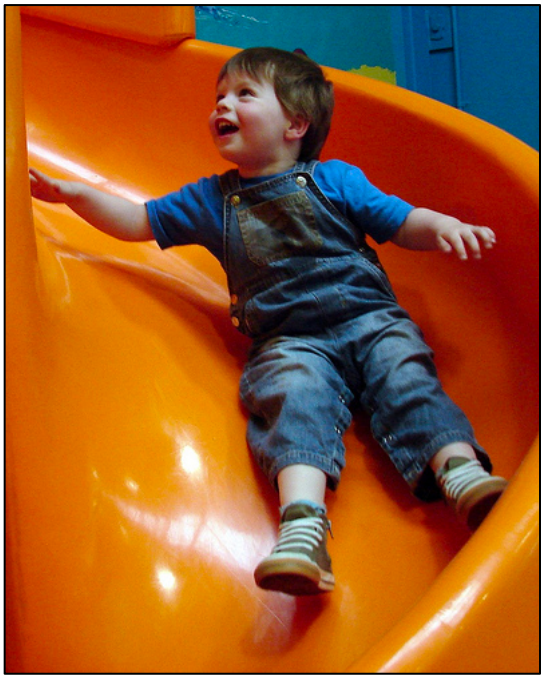
4. **According to CLIPScore, you need to:**
    1. **visualize top-1 and last-1 image-caption pairs**
    2. **report its corresponding CLIPScore**

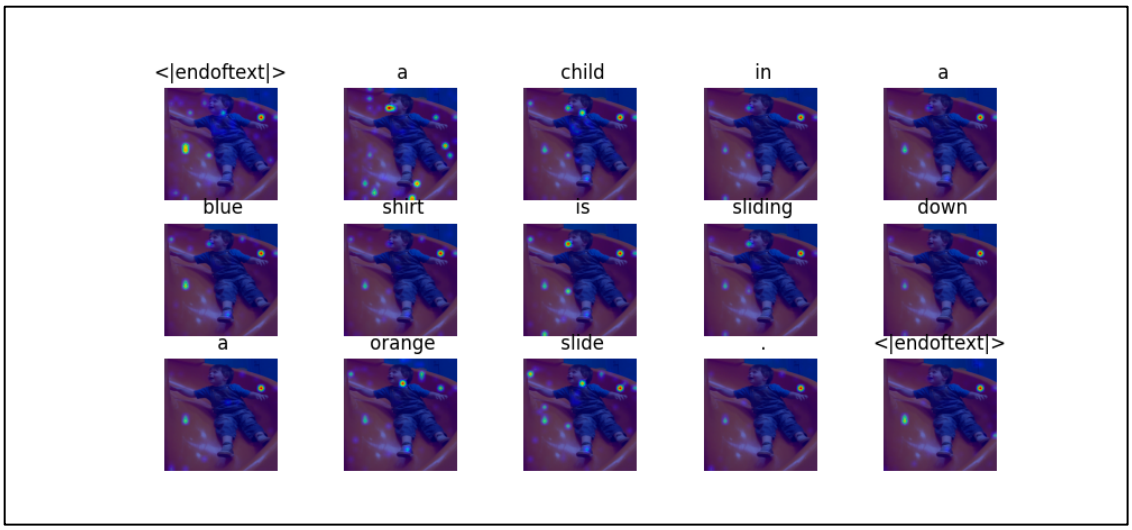
in the validation dataset of problem 2. (5%)

Top1: 2988244398.jpg

- Captions: “a child in a blue shirt is sliding down a orange slide .”
- CLIP Score: 1.005

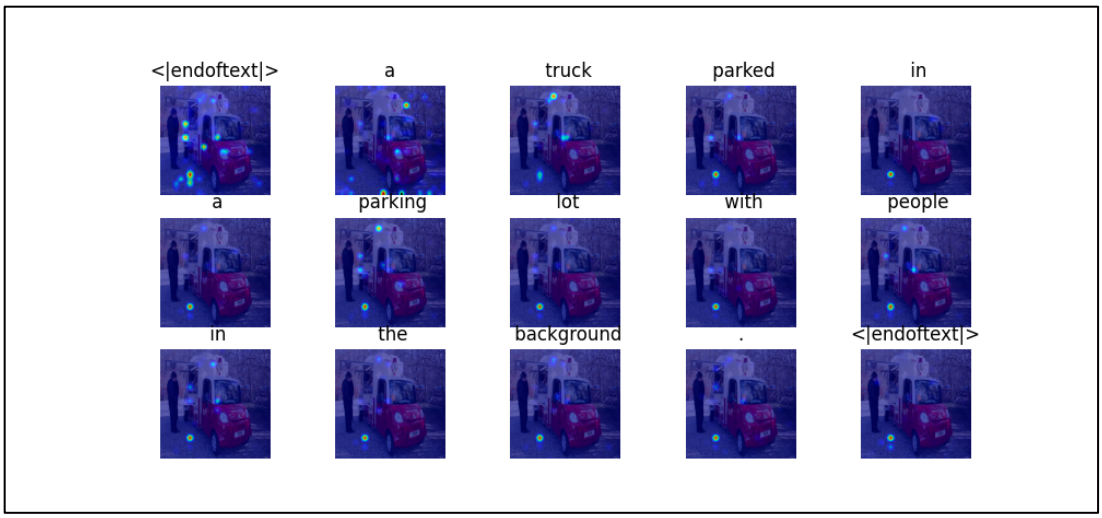
  (It is weird for the score > 1, but I checked the code and it was just my best result. I think that it resulted from the implementation of 2.5 * max(cos_sim, 0))





Last 1: 000000353836.jpg

- Captions: “a trunk parked in a parking lot with people in the background.”
- CLIP Score: 0.449



5. 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%)

I don't think it is reasonable at all. My CLIP score is high, but my attention map has a bad performance. As you can see, the above heat map doesn't correspond to the captions except for the word "child". In my opinion, it doesn't make any sense. I asked the TA for suggestions, and they told me that the main problem stemmed from the small dataset. Besides, different layers may affect the performance of the attention map. Based on this, the explainable AI is interesting extended from the above questions.

- Reference

1. https://paperswithcode.com/sota/image-classification-on-imagenet
2. https://github.com/mlfoundations/open_clip
3. https://github.com/bckim92/language-evaluation
4. https://huggingface.co/docs/hub/timm
5. https://github.com/lucidrains/bidirectional-cross-attention/blob/main/bidirectional_cross_attention/bidirectional_cross_attention.py
6. https://github.com/XiangLi1999/PrefixTuning/blob/cleaned/transformers/src/transformers/modeling_bart.py
7. https://pypi.org/project/loralib/