DLCV HW3 Report

tags: DLCV

Course	Student ID	Name	Date
2022 Fall NTU DLCV	R10943109	Shiuan-Yun Ding	2022-11-19

HackMD Link: https://hackmd.io/@mirkat1206/HkqJbqqSo

Problem 1: Zero-Shot Image Classification with CLIP (30%)

(3%) Methods Analysis

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

- CLIP consists of one visual encoder and one text encoder.
- CLIP learned visual concepts with natural language supervision.
- CLIP is trained with 400 million image-text pairs.
- Because of 3 above reasons, CLIP can achive strong zero-shot performances.

(6%) Prompt-Text Analysis

Prompt Text Templates

1. "This is a photo of {object}"
2. "This is a {object} image"
3. "No {object}, no score"

Experimental Results

Metrix	Baseline (10%)	Prompt 1	Prompt 2	Prompt 3
Accuracy	60%	60.82%	67.81%	54.94%

(6%) Quantitative Analysis

Image Goden Label	Prompt 1: "This is a photo of {object}"

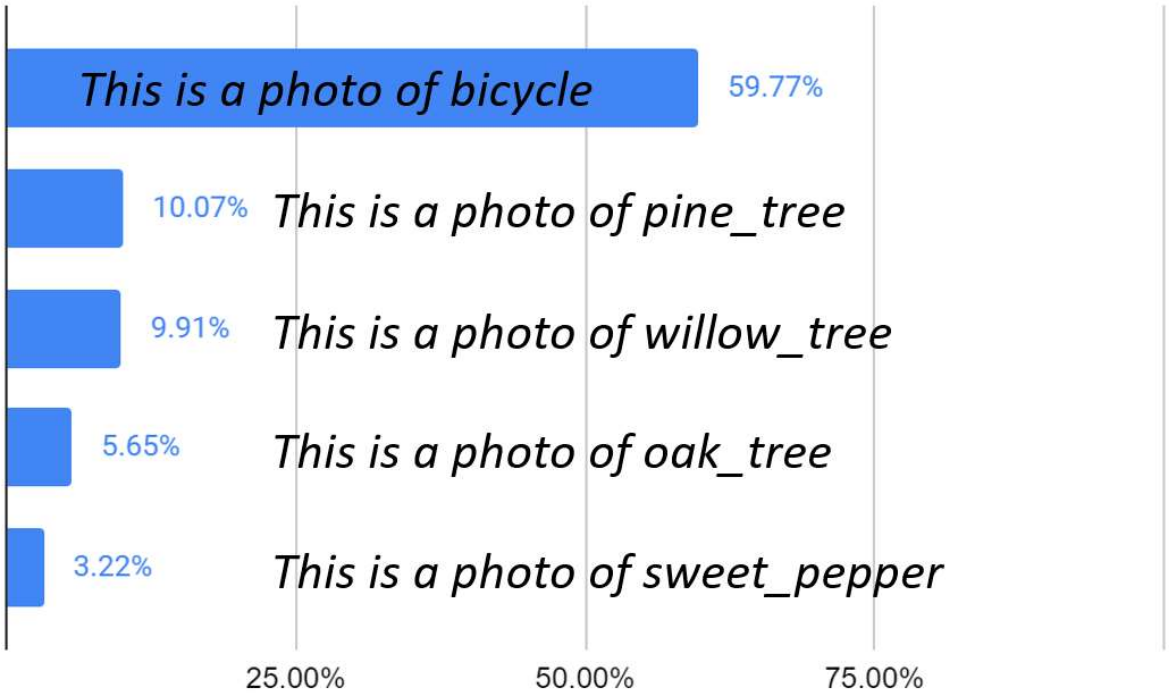
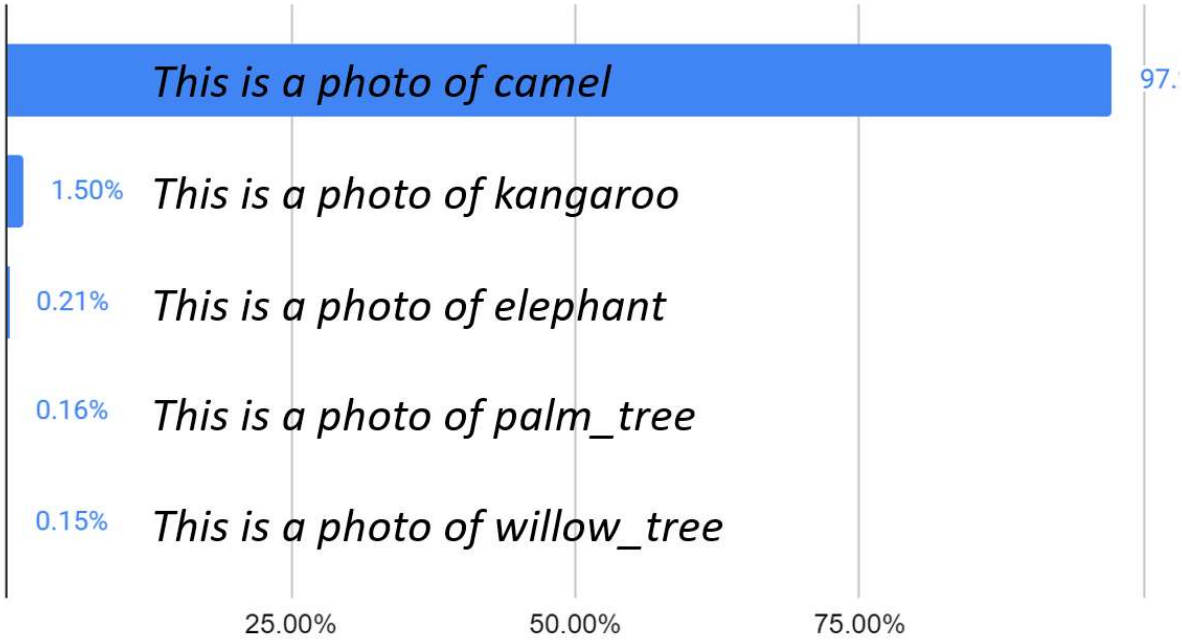
Image Goden Label	Prompt 1: "This is a photo of {object}"
 bicycle	
 camel	

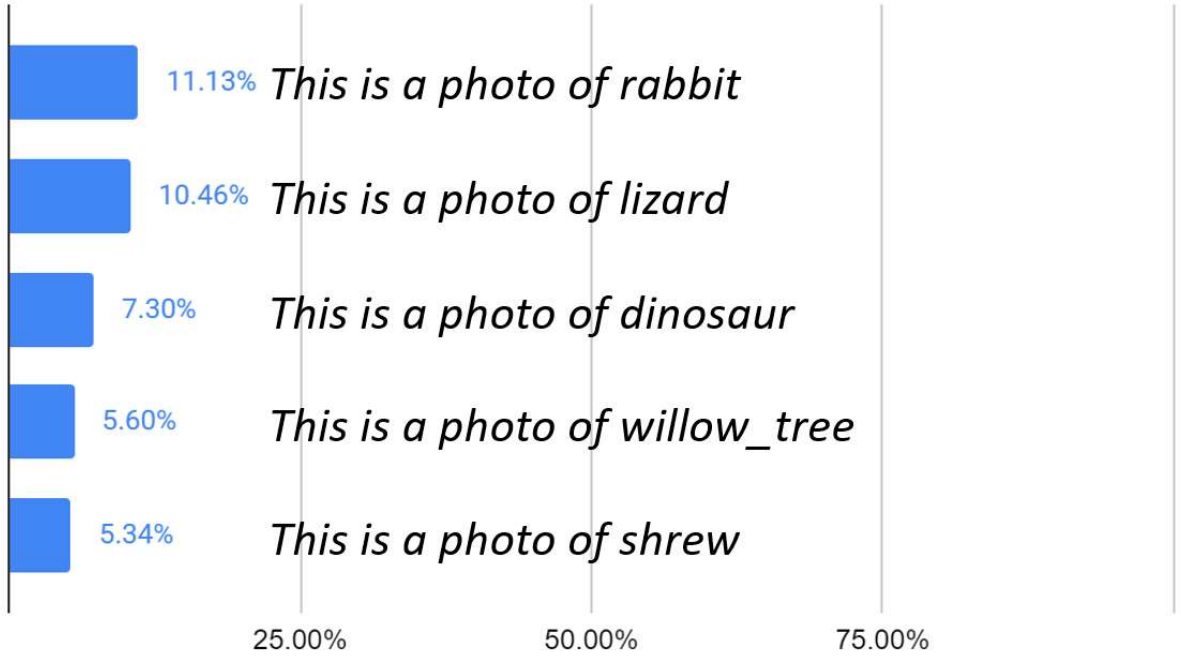
Image Goden Label	Prompt 1: "This is a photo of {object}"
 shrew	

Image Goden Label	Prompt 2: "This is a {object} image"
 bicycle	

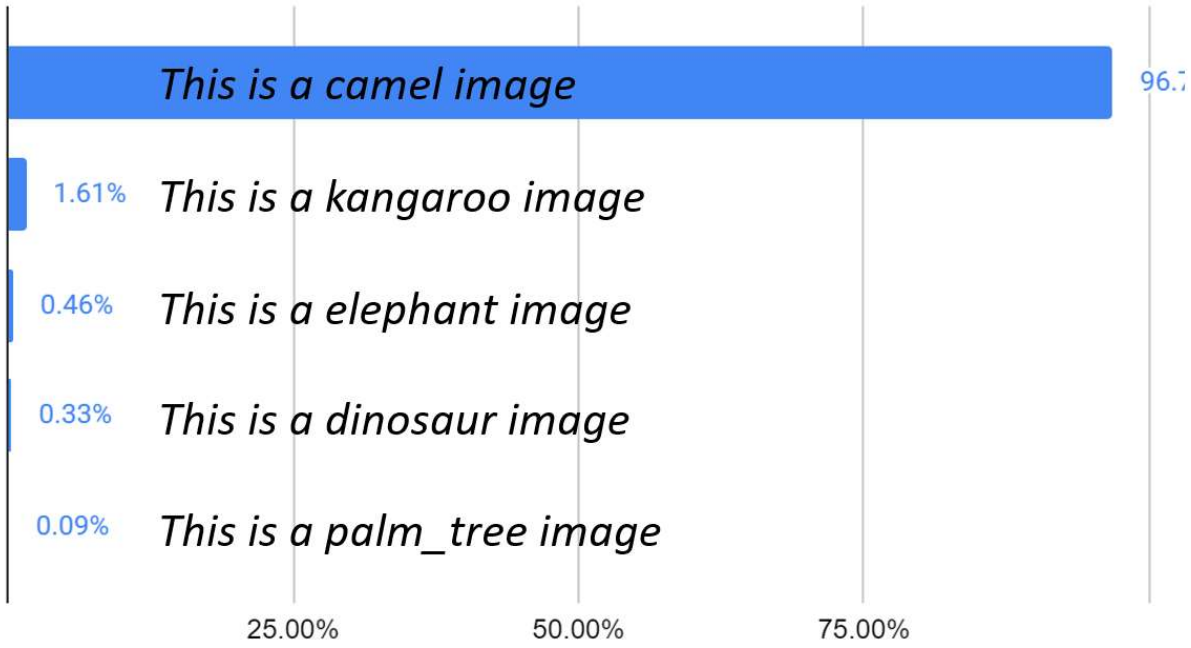
Image Goden Label	Prompt 2: "This is a {object} image"
 camel	
 shrew	

Image Goden Label	Prompt 3: "No {object}, no score"

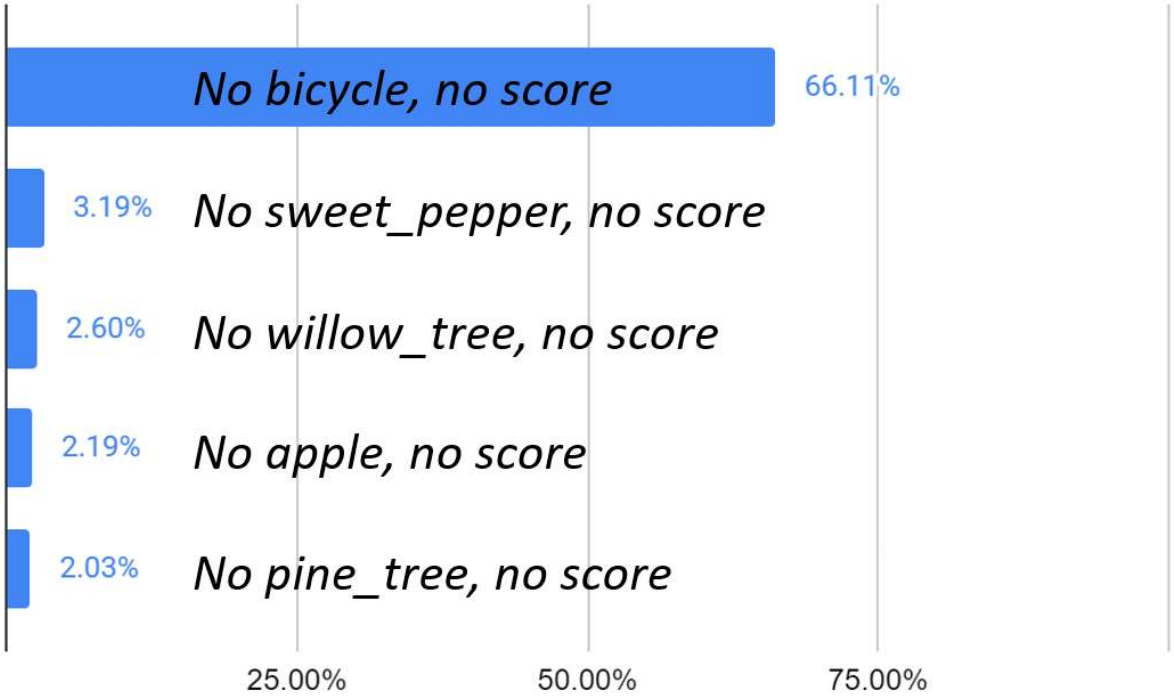
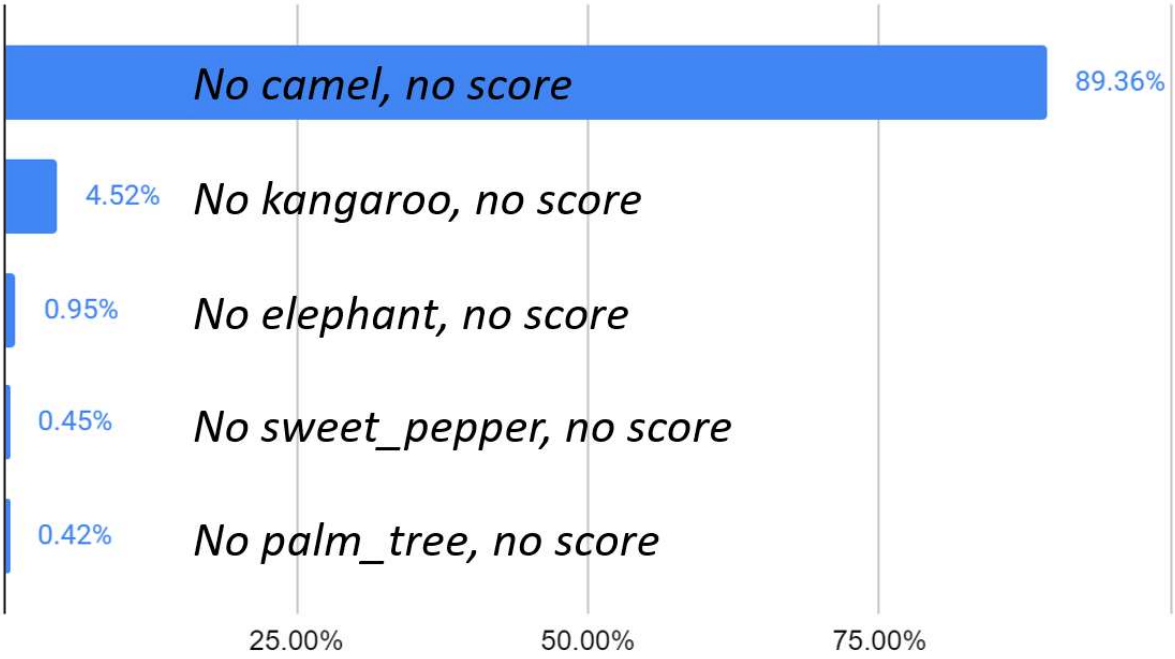
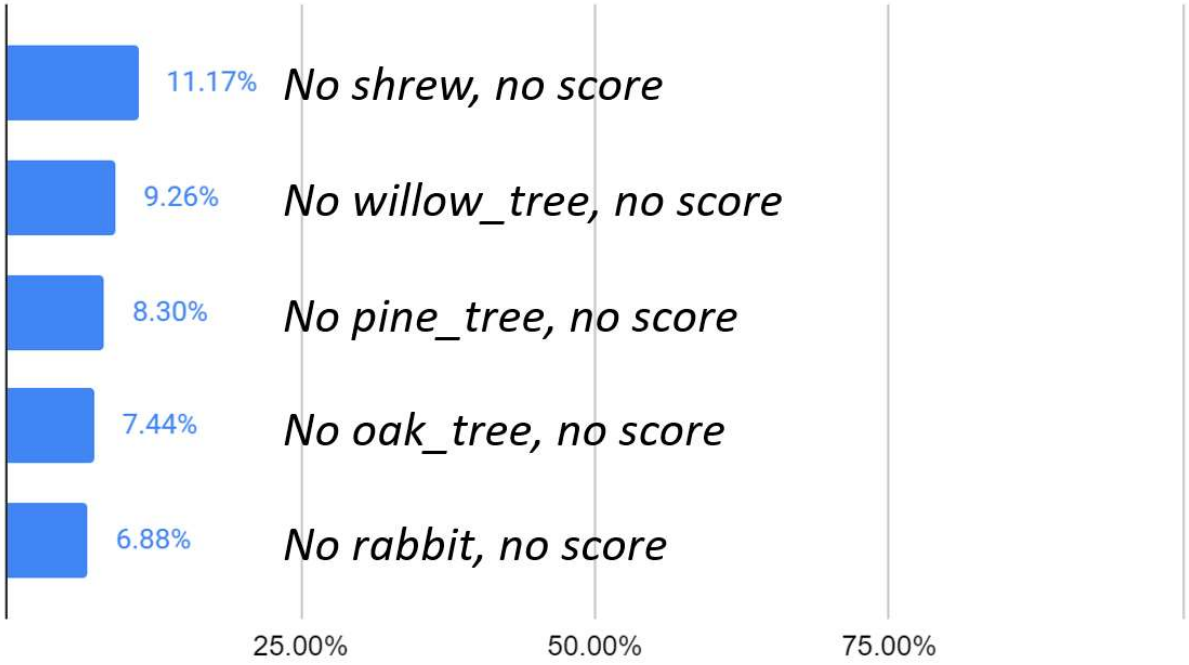
Image Goden Label	Prompt 3: "No {object}, no score"
 bicycle	
 camel	

Image Goden Label	Prompt 3: "No {object}, no score"
 shrew	

Reference

- https://github.com/openai/CLIP
- https://aclanthology.org/2022.acl-long.421.pdf

Problem 2: Image Captioning with Vision and Language Model (50%)

Matrix	CIDEr	CLIPScore
Simple Baseline (13%)	0.72	0.67
Strong Baseline (7%)	0.87	0.70
My Result	0.5671	0.6251

Implementation Details

Encoder	Decoder
vit_huge_patch14_224_clip_laion2b	https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

Optimizer	Learning Rate	Criterion	Epochs
Adam	4e-4	CrossEntropyLoss	17

(2.5%) BestCIDEr & CLIPScore

CIDEr	CLIPScore
0.5671	0.6251

(7.5%) Three Different Attempts

1. Encoder with vit_base_patch16_224

CIDEr	CLIPScore	Epochs
0.3164	0.5343	20

2. Encoder with vit_large_patch16_224_in21k

CIDEr	CLIPScore	Epochs
0.0002	0.5143	3

3. Encoder with vit_huge_patch14_224_clip_laion2b

CIDEr	CLIPScore	Epochs
0.5671	0.6251	17

Reference

- https://github.com/rwightman/pytorch-image-models/blob/main/timm/models/vision_transformer.py
- https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning
- https://github.com/rammyram/image_captioning
- https://blog.csdn.net/qq_37541097/article/details/113247318

Problem 3: Visualization of Attention in Image Captioning (20%)

(10%) Five Test Images

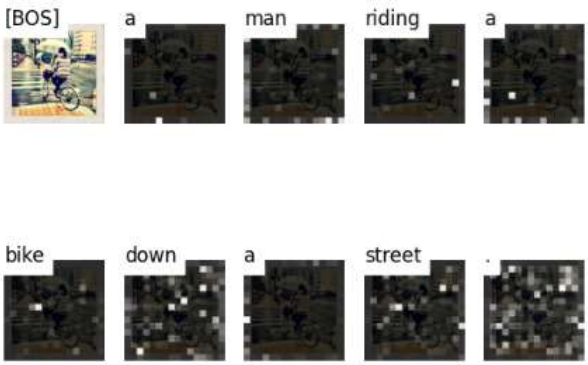
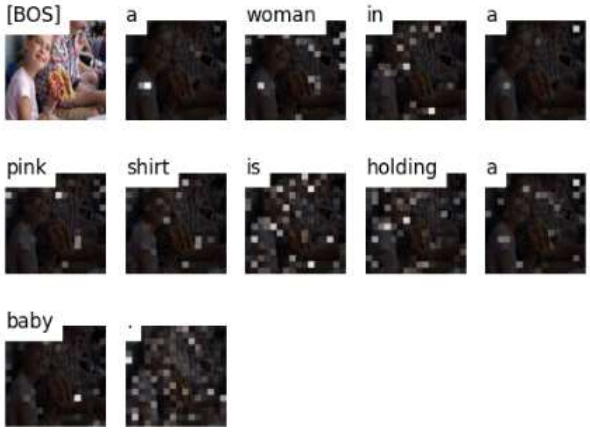
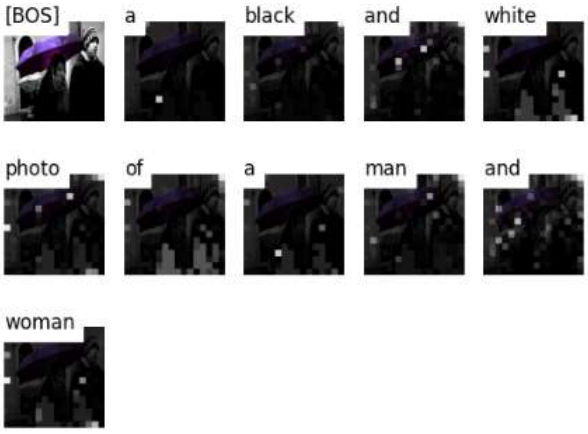
Image	Attention Maps
bike	

Image	Attention Maps
girl	
sheep	
ski	
Umbrella	

(5%) Top-1 and Last-1 Image-Caption Pairs

Top-1

Image 000000056306	
CLIPScore	0.9997
Attension Maps	

Last-1

Image 000000392315	
CLIPScore	0.2191

Attension Maps	

(5%) Q & A

1. Is the caption reasonable?
   - Not really. In my opinion, both sentences are having the same quality: the last-1 image caption does capture the man in the picture, but it mistakes the color; the top-1 does capture the road, but it mistakes the fire hydrant as a yellow bus.
   - I think the reason that the top-1 has the very high 0.9997 scoer is because the CLIPScore is based on clip and also I use "vit_huge_patch14_224_clip_laion2b" as encoder. I think clip has mistakens the white hydrant as a yellow bus, and so does my encoder. Therefore, the high score is there.
2. Does the attended region reflect the corresponding word in the caption?
   - Not at all. In fact, all my attention region are messed up. It may be because of my poor transformer-decoder model.

Reference

- https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning