DLCV HW3 Report

NTUEE
B10901187
Hung Kai Chung

Nov. 2023

Problem 1: Zero-shot Image Classification with CLIP

1. Methods analysis

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.

Previously the **ResNet** and **VGG** are good at one task due to training on this specific dataset, but in new task this method can't do very well. In contrast, **CLIP** (Contrastive Language–Image Pre-training) not directly optimizing for the benchmark, it uses another supervision: the text paired with images found across the internet. In Figure 1, CLIP has image and text encoder, trying to get text image pair similarity high. CLIP learns both visual & language concepts instead of trying to get the label of the specific task, hence, it achieve competitive zero-shot performance on a variety of image classification datasets.

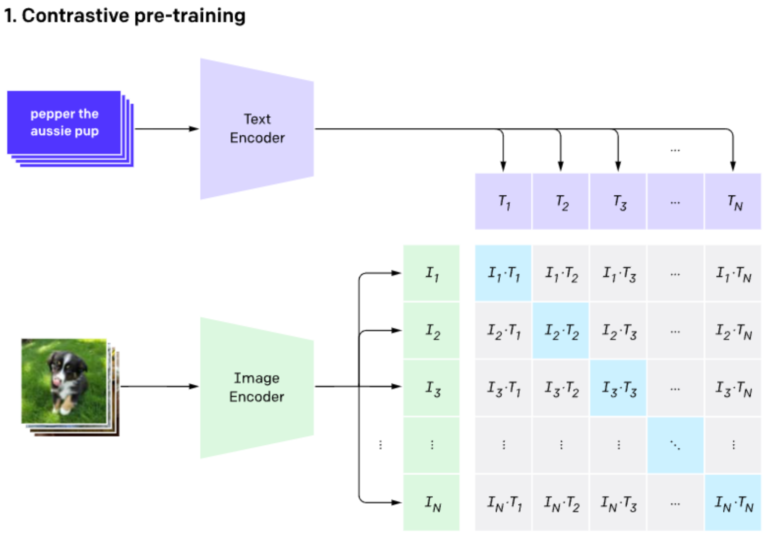
1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
$I_2{\cdot}T_1$ $I_2{\cdot}T_2$ $I_2{\cdot}T_3$ ... $I_2{\cdot}T_N$
$I_3{\cdot}T_1$ $I_3{\cdot}T_2$ $I_3{\cdot}T_3$ ... $I_3{\cdot}T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N{\cdot}T_1$ $I_N{\cdot}T_2$ $I_N{\cdot}T_3$ ... $I_N{\cdot}T_N$

Figure 1: CLIP training process

2. Prompt-text analysis

Please compare and discuss the performances of your model with the following three prompt templates: "This is a photo of object","This is not a photo of object","No object, no score."

	This is a photo of object	This is not a photo of object	No object, no score.
CLIP score	0.69988	0.6675	0.6235

3. Quantitative analysis

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

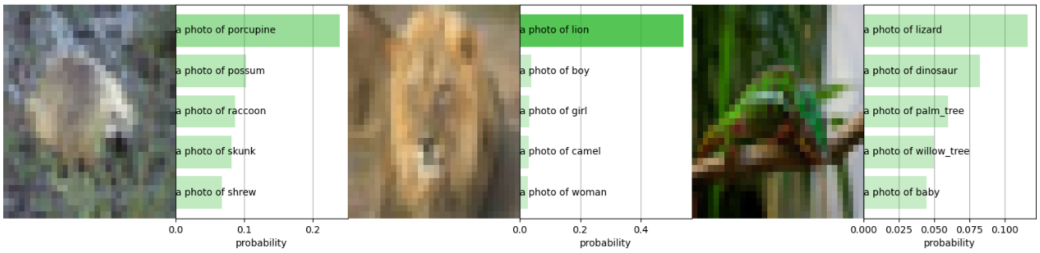

Figure 2: Top-5 similarity scores of three images

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

1. Evaluation metrics report

Best setting: Adapter

- CIDEr = 0.936
- CLIPScore = 0.728

3 different attempts of PEFT and their corresponding CIDEr & CLIPScore

	Adapter	Prefix tuning	Lora
CIDEr	0.936	0.843	0.897
CLIP score	0.728	0.6998	0.717

2. Visualization of Attention in Image Captioning

Please visualize the predicted caption and the corresponding series of attention maps

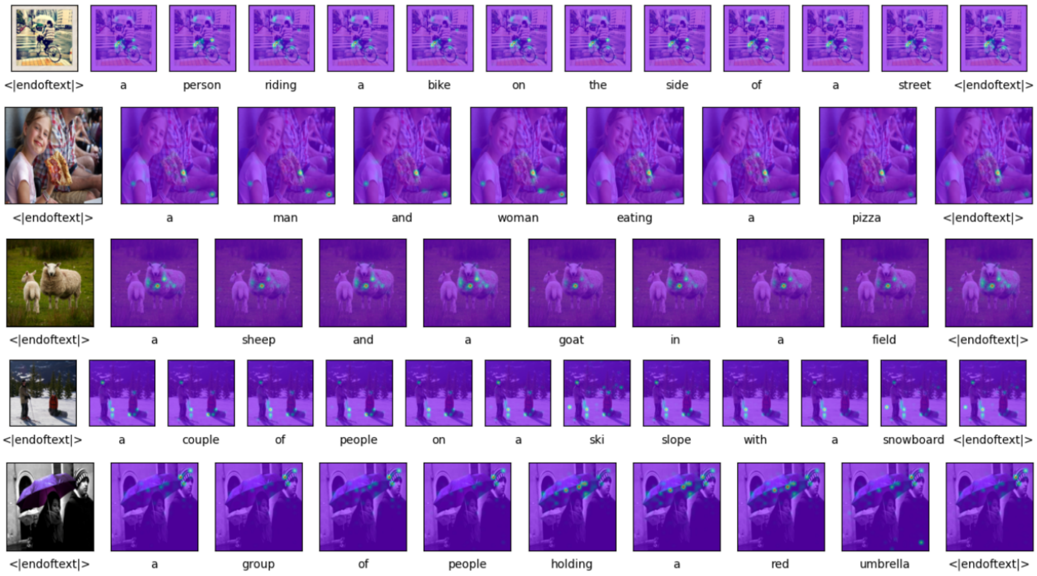
<|endoftext|> a person riding a bike on the side of a street <|endoftext|>
<|endoftext|> a man and woman eating a pizza <|endoftext|>
<|endoftext|> a sheep and a goat in a field <|endoftext|>
<|endoftext|> a couple of people on a ski slope with a snowboard <|endoftext|>
<|endoftext|> a group of people holding a red umbrella <|endoftext|>

Figure 3: Visualization of Attention in Image Captioning

According to CLIPScore, you need to: visualize top-1 and last-1 image-caption pairs report its corresponding CLIPScore in the validation dataset of problem 2.

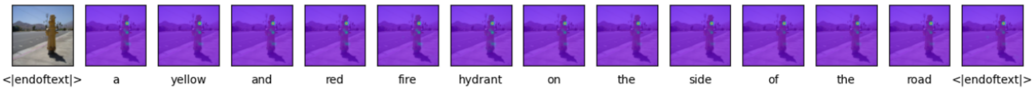

Figure 4: Top-1 img-caption pair CLIPScore=0.99975

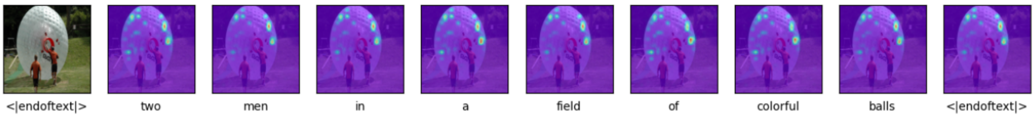

Figure 5: Last-1 img-caption pair CLIPScore=0.38360

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?

In Figure 3, row 1-5 are almost reasonable, particularly in row 3, notice that the model can classify the sheep and goat, but in row 5 the model predict wrong color of the umbrella which is purple not red. In Figure 4 the img only have yellow hydrant, but the caption also output red fire hydrant, I think it's due to the data basis in train data. The model commonly see the red hydrant, so the probability of the red hydrant may be very high. In Figure 5, the caption can predict it's a ball, which is correct for the image. In Figure 3 to 5, the attended region doesn't reflect well on the corresponding word in the caption, I think the reason is that we only can finetune the cross-attention layer & adapter so actually we only can finetune a relatively small parameters. Hence, the visualization of attention in Image Cpationing is not pretty well.