DLCV hw3

ID: r12922075

Name: 吳韋論

Zero-shot image classification with CLIP

1. Methods analysis

Question: Why CLIP could achieve competitive zero-shot performance on a great variety of image classification datasets?

Answer:

- Contrastive Learning with images-text pairs

  CLIP uses contrastive learning. It focuses on making similar pairs of images and texts closer together in a shared space. By maximizing the cosine similarity for closer pairs and minimizing it for farther pairs, CLIP learns to understand complex relationships between images and texts.

- Large scale and diverse images-texts data

  CLIP is trained on diverse data to understand images and texts together. It helps to learn flexible features that work well for different tasks. By not focusing on specific tasks during training, CLIP becomes good at new tasks it hasn't seen before, showing strong performance even without specific training for those tasks.

2. Prompt-text analysis

In addition to the three prompts specified in the report, I achieved the best results in the final inference by using prompt text templates from CIFAR100.

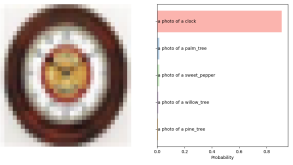
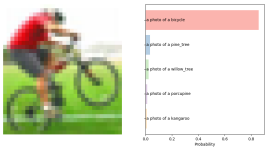
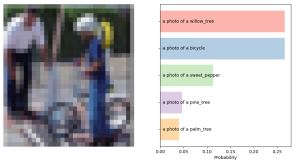
Ref: templates from CIFAR100

	Validation Accuracy
“This is a photo of {object}”	67.48%
“This is not a photo of {object}”	69.64%

“No {object}, no score.”	45.24%
CIFAR100 prompt templates	82.48%

3. Quantitative analysis

In the three example images, the first two models correctly identify a clock and a bicycle. However, the third image confuses the model, making it uncertain whether it's a tree or a bicycle. As a result, the probability distribution is more spread out.

PEFT on Vision and Language Model for Image Captioning

1. Best setting and its corresponding CIDEr & CLIPScore

Model

- Encoder: 'vit_large_patch14_clip_224.openai’
- Decoder
  - n_layer: 12
  - n_head: 12
  - n_embd: 768
  - adapter_size: 384
  - dropout: 0.1
  - cross_n_embd: 384

PEFT (Adapter)

- Downsample: nn.Linear(n_embd, adapter_size)
- Activation: nn.ReLU()
- Upsample: nn.Linear(adapter_size, n_embd)

Training

- Epochs: 5
- Optimizer: AdamW
- Scheduler: CosineAnnealingWarmupRestarts (warmup=0.1*total_steps, max_lr=5e-4, min_lr=1e-6)
- Loss: Cross Entropy Loss
- Params: 30776832
  - Adapter layer
  - Layer Normalization
  - Cross Attention

Result (Greedy search)

	CIDEr	CLIPScore
Adapter	0.964	0.733

2. Three attempts of PEFT and their corresponding CIDEr & CLIPScore

Model and Training are the same as the best settings.

PEFT

- Adapter
  - Downsample: nn.Linear(n_embd, adapter_size)
  - Activation: nn.ReLU()
  - Upsample: nn.Linear(adapter_size, n_embd)
- Lora
  - cross_n_embd: 768
  - c_attn = lora.Linear(cfg.n_embd, 3 * cfg.n_embd, r=8)
  - c_proj = lora.Linear(cfg.n_embd, cfg.n_embd, r=8)
- Prefix Tuning
  - cross_n_embd: 768
  - nn.Sequential(
    torch.nn.Linear(self.token_dim, self.encoder_hidden_size),

```
            torch.nn.Tanh(),
            torch.nn.Linear(self.encoder_hidden_size, self.num_layers * 2 *
        self.token_dim))
```

Result (Greedy search)

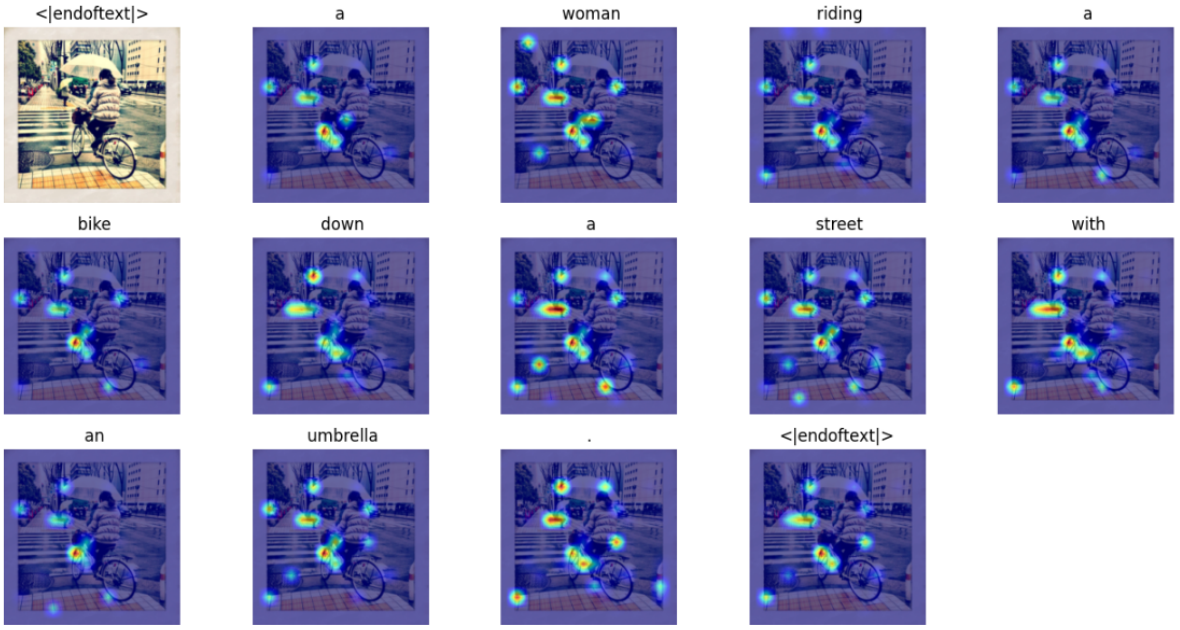
Among the three PEFT methods, adapters perform the best. Implementation-wise, lora is the simplest, while prefixTuning is more complex and less effective.

	CIDEr	CLIPScore
Adapter	0.964	0.733
Lora	0.901	0.726
Prefix Tuning	0.827	0.714

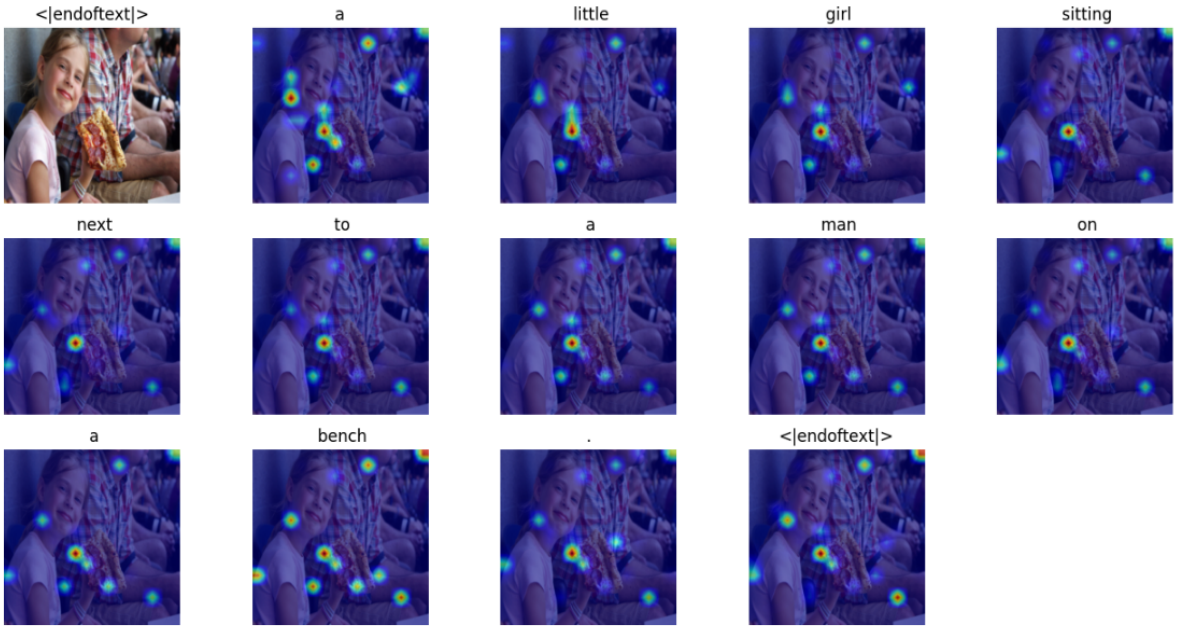
3. Visualization of attention in image captioning

Five test images with predicted caption and attention maps (last decoder layer)

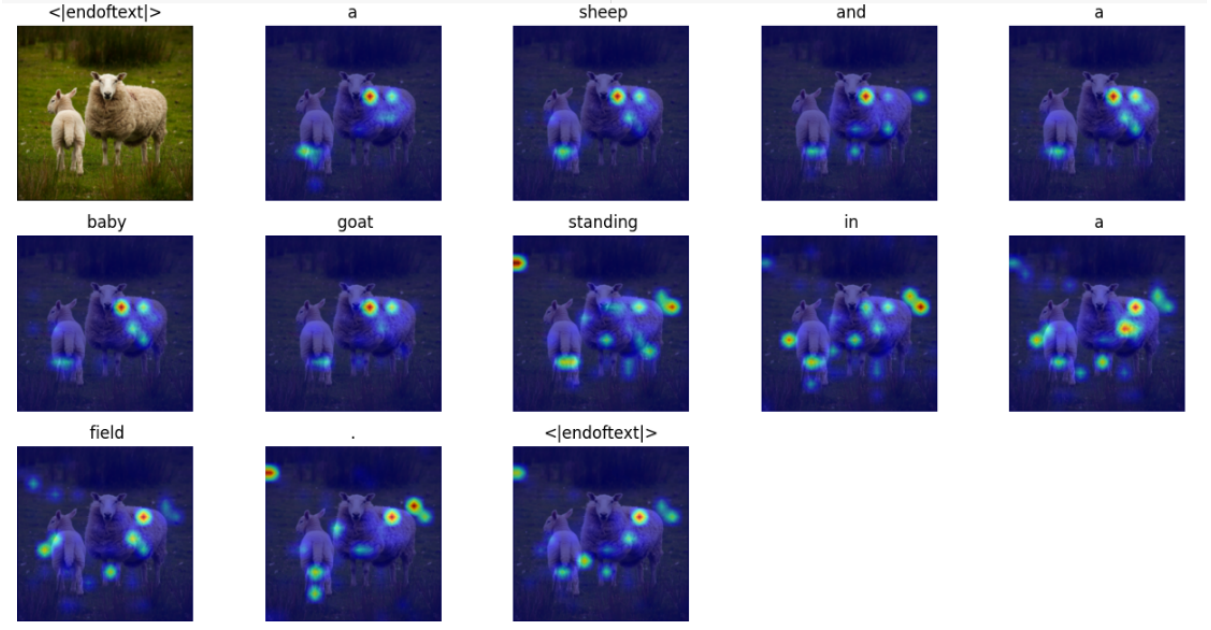
Bike.jpg



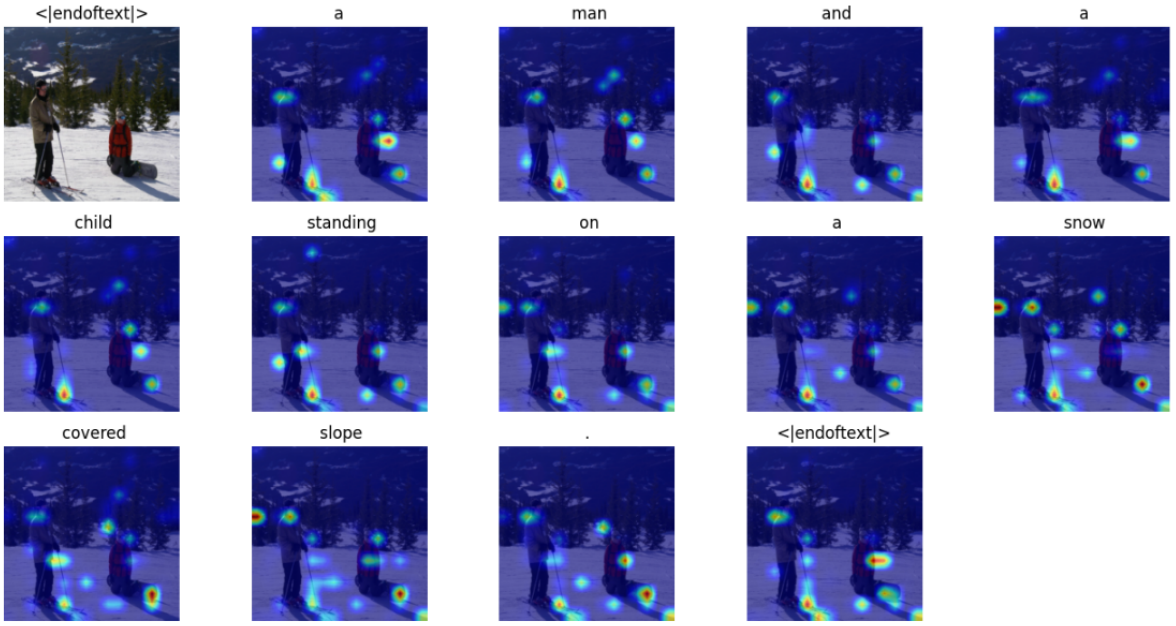
Girl.jpg



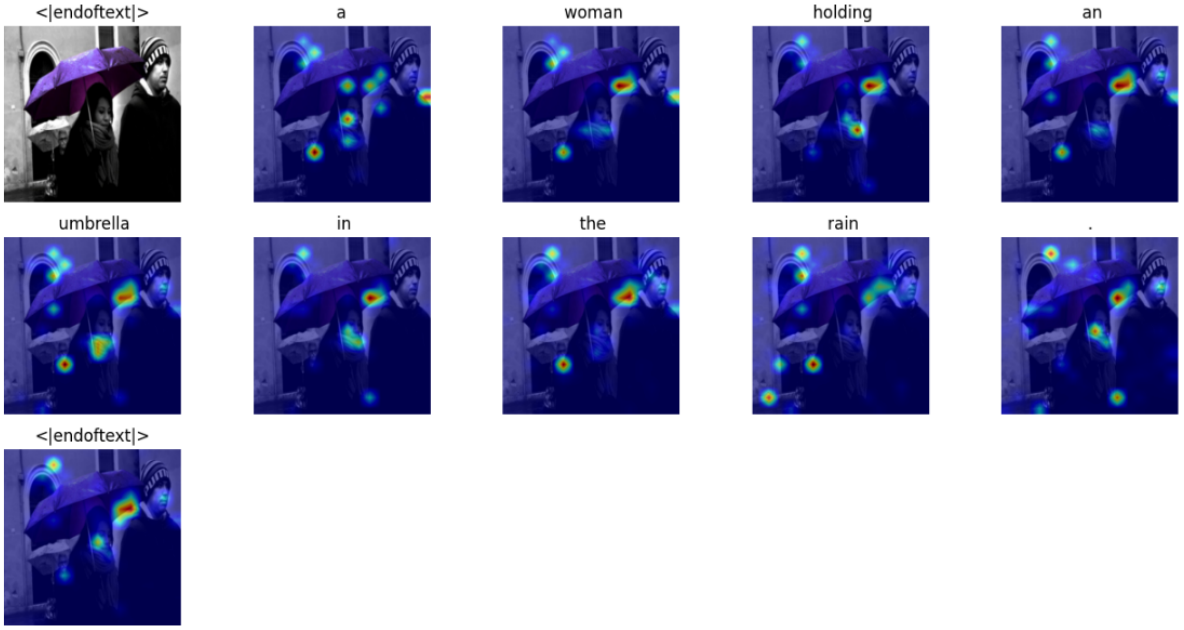
sheep.jpg



ski.jpg

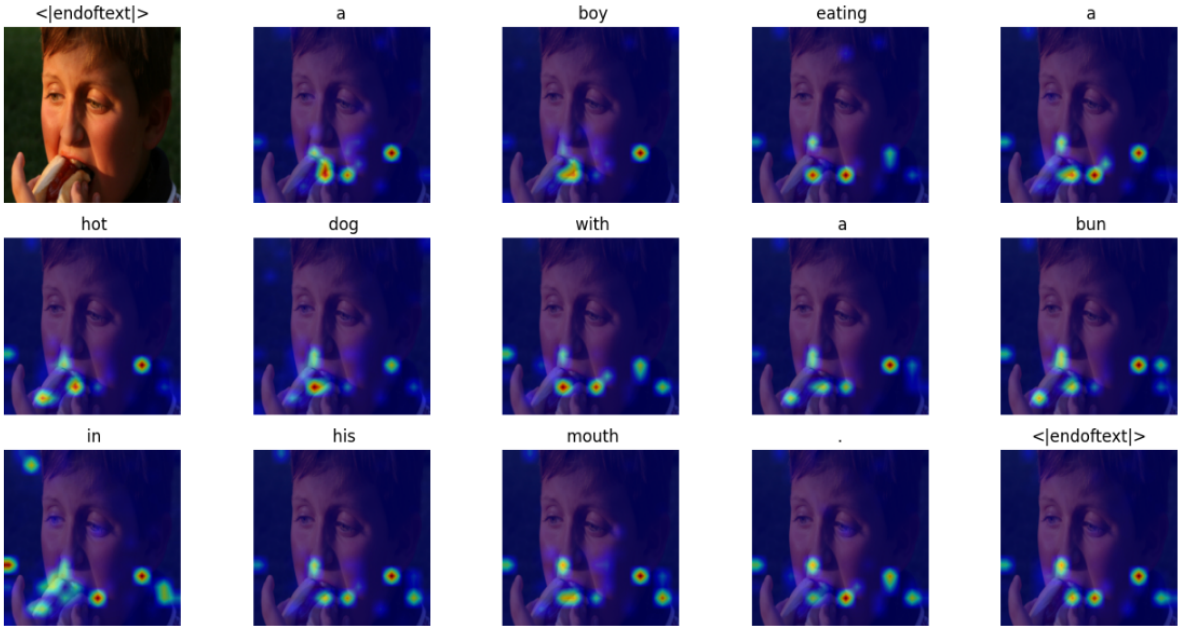


umbrella.jpg

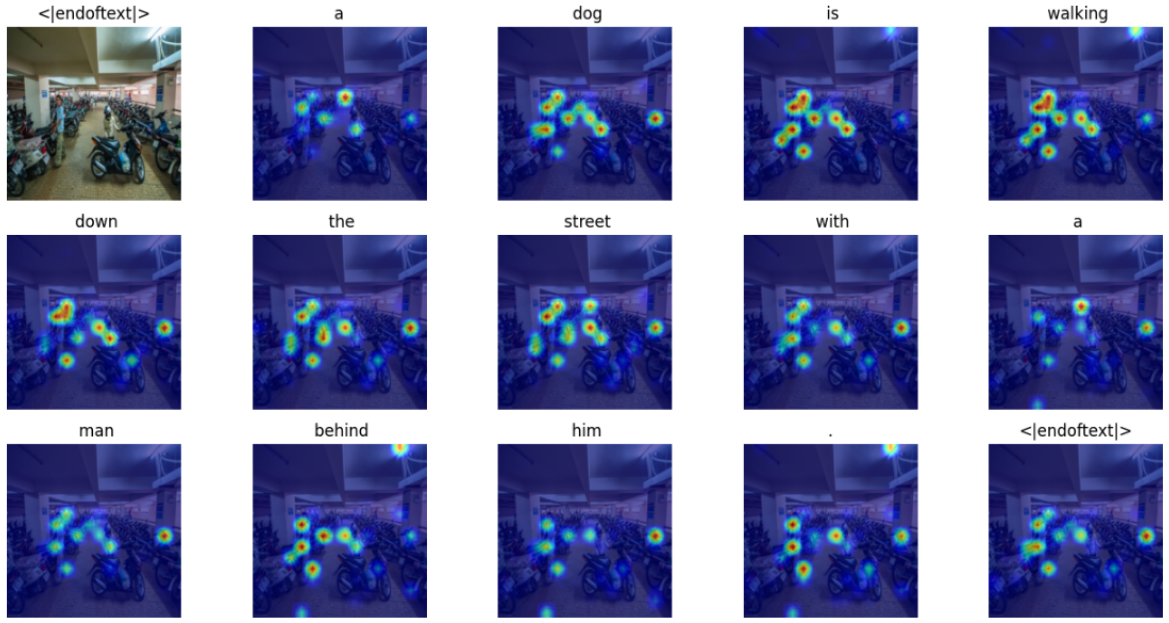


Visualize top-1 and last-1 image-caption pairs with CLIPScore

Best image id: 000000340181, ClipScore: 1.0076904296875



Worst image id: 000000028523, ClipScore: 0.3265380859375



Is the caption reasonable?

The majority of captions are reasonable, effectively describing the images.

For instance, "Bike.jpg" describes a picture of a girl, a bicycle, and an umbrella. "Umbrella.jpg" describes an umbrella and rainy weather. "000000340181.jpg" describes a boy, a hot dog, and a mouth in the image.

Does the attended region reflect the corresponding word in the caption?

The majority of attended regions are reasonable, successfully aligning key elements in the image with corresponding words.

For instance, in "Bike.jpg," the image region of the woman corresponds to the word "woman," and the region of the bike aligns with the word "bike." In "000000340181.jpg," the image region of the hot dog corresponds to the word "hot dog."