DLCV HW3 REPORT

R11521701 程懷恩

Problem 1：Zero-shot Image Classification with CLIP

1. Methods analysis (3%)

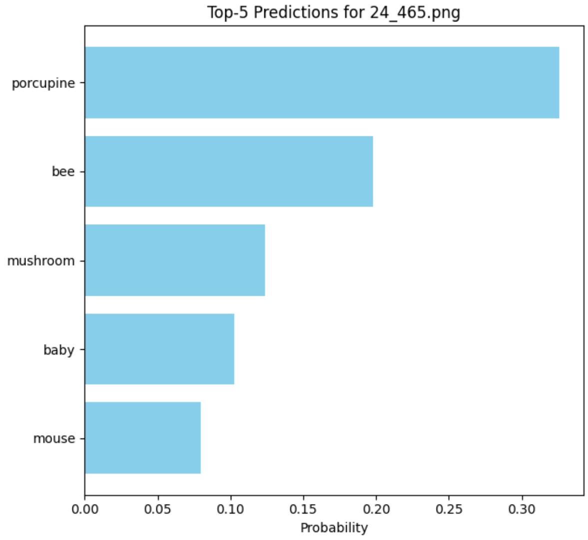
CLIP's ability to achieve competitive zero-shot performance is largely due to its extensive and diverse training data. CLIP uses a contrastive learning framework where it learns to align the embedding of an image with the embedding of its corresponding textual description. The model is trained to maximize the similarity between correct image-text pairs and minimize it for incorrect ones. This approach helps CLIP understand the semantic relationship between images and text, enabling effective transfer learning.

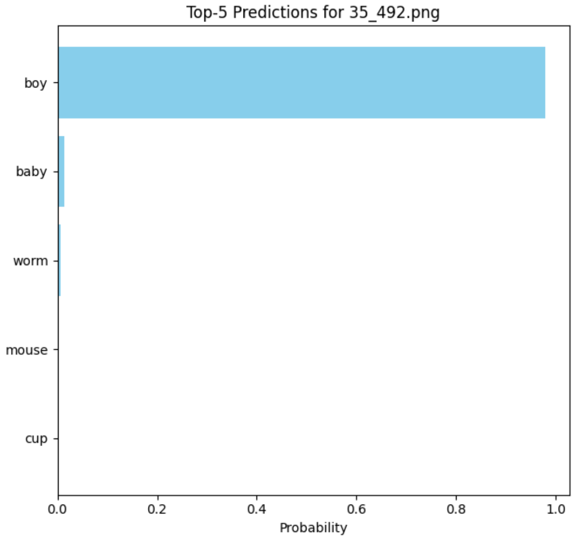
2. Prompt-text analysis (6%)

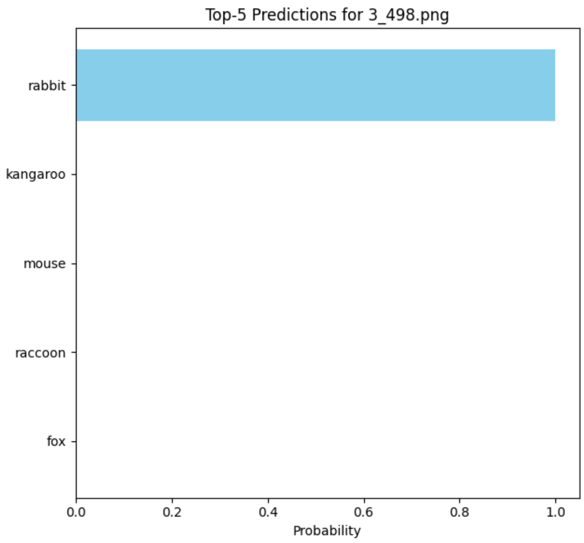
text prompt	**acc**
This is a photo of {object}	0.6556
This is not a photo of {object}	0.7232
No {object}, no score	0.5344

1. **"This is a photo of {object}" (Accuracy: 0.6556):**
   This prompt directly asks the model to confirm the presence of the specified object in the image. An accuracy of 65.56% indicates that the model is only moderately accurate at identifying the object.
2. **"This is not a photo of {object}" (Accuracy: 0.7232):**
   Interestingly, the model performs better when tasked with identifying what is not in the image, with an accuracy of 72.32%. This could suggest that the negative prompt provides a stronger signal for the model, perhaps due to it being less common or more distinctive than affirmative statements, making the contrast easier to detect.
3. **"No {object}, no score" (Accuracy: 0.5344):**
   This prompt, which seems to suggest a penalty for incorrectly identifying the object, unlike the second scenario with a negative prompt got the highest score, this one has the lowest accuracy at 53.44%. This might be due to the more complex structure of the prompt, which is less straightforward than a simple confirmation of more ambiguous interpretations by the model.

3. Quantitative analysis (6%)







Problem 2：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%)**

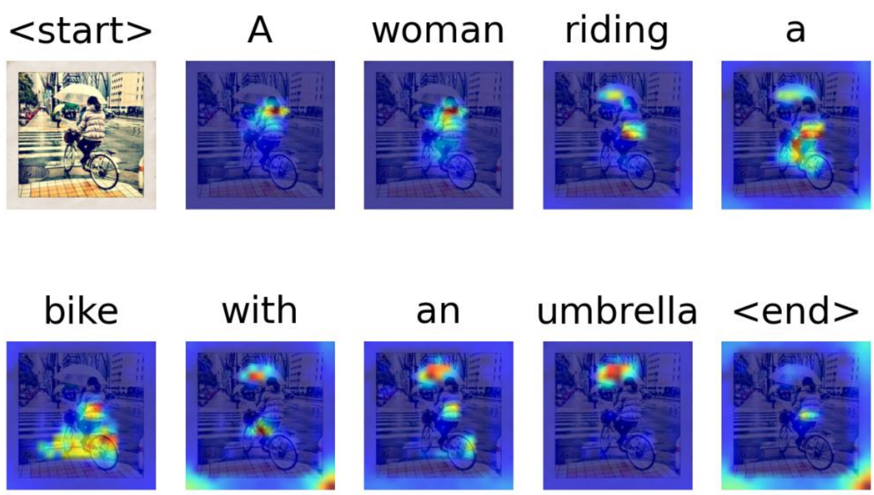
CIDEr	CLIPScore	Param
0.825545172	0.723558813	29,175,296

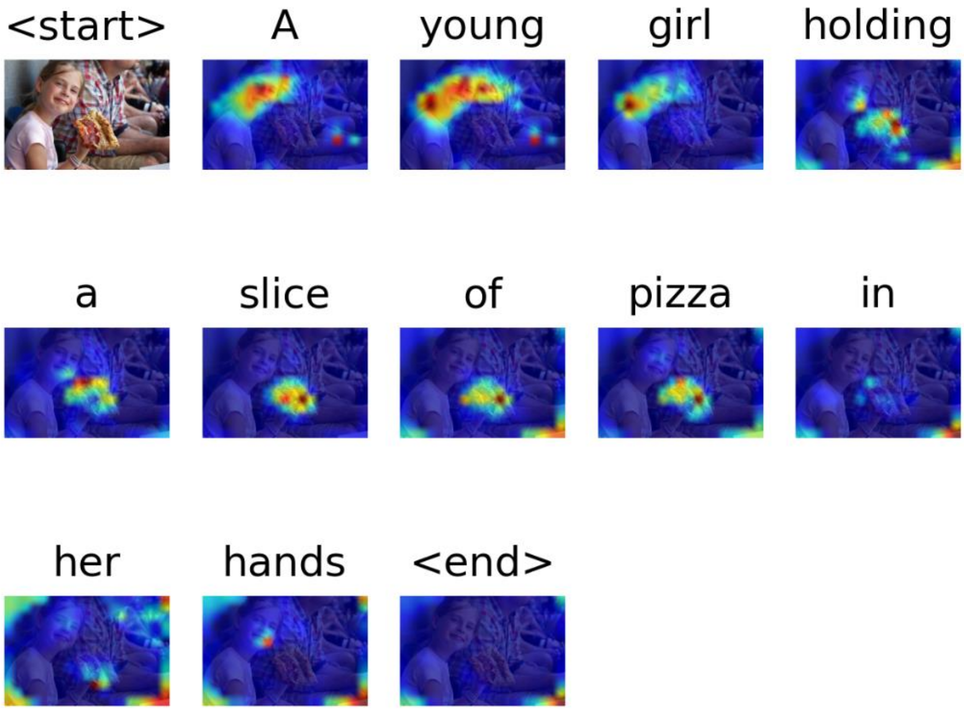
My current optimal setup involves incorporating the adapter into the final two layers of the block and configuring the bottleneck layer to have 256 units. This configuration is implemented alongside an encoder utilizing the OpenAI CLIP ('ViT-L/14') model.

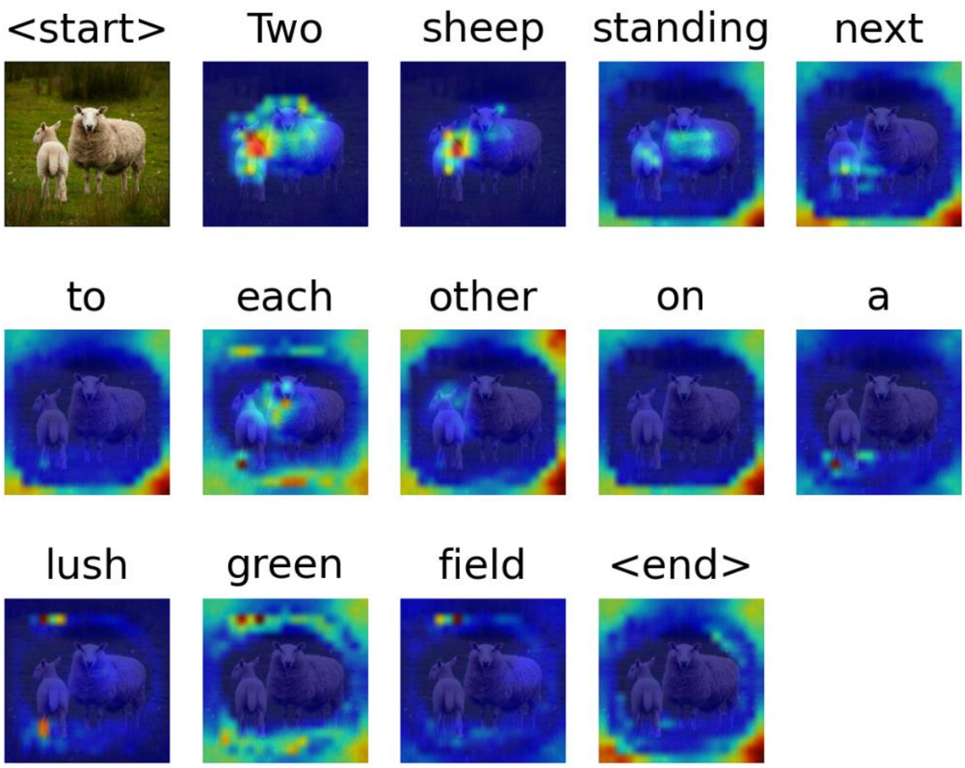
2. **Report 3 different attempts of PEFT and their corresponding CIDEr & CLIPScore. (7.5%, each setting for 2.5%)**

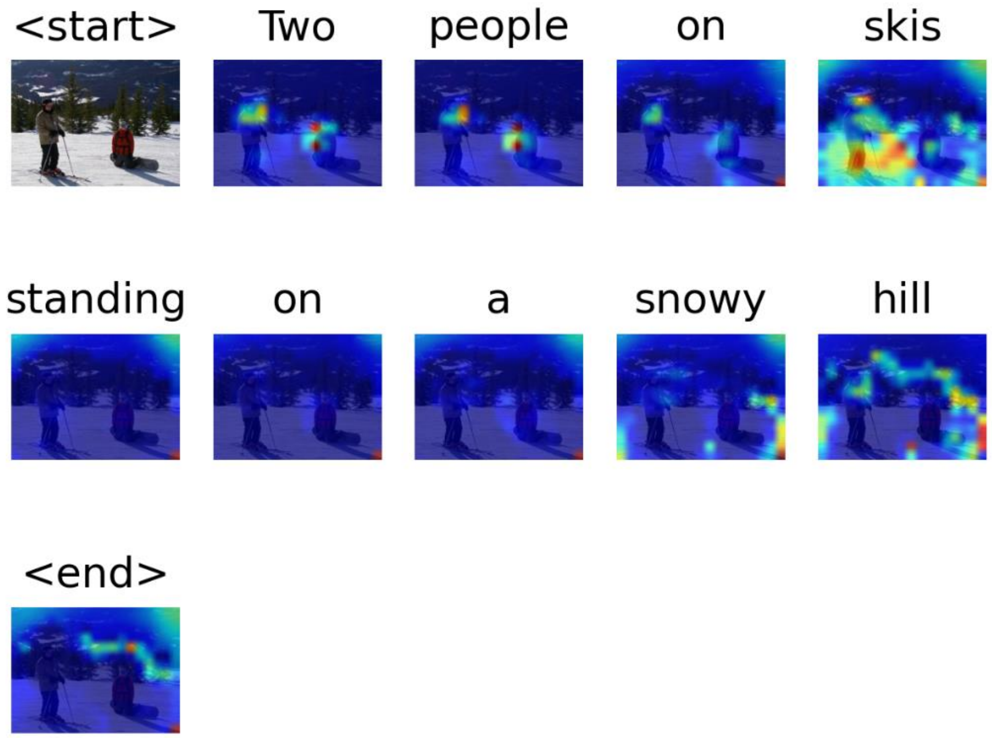
	CIDEr	CLIPScore	param
adapter	0.825545172	0.723558813	29,175,296
prefix	0.452588029	0.679535032	28,386,816
lora	1.98E-05	0.477362815	3,538,944

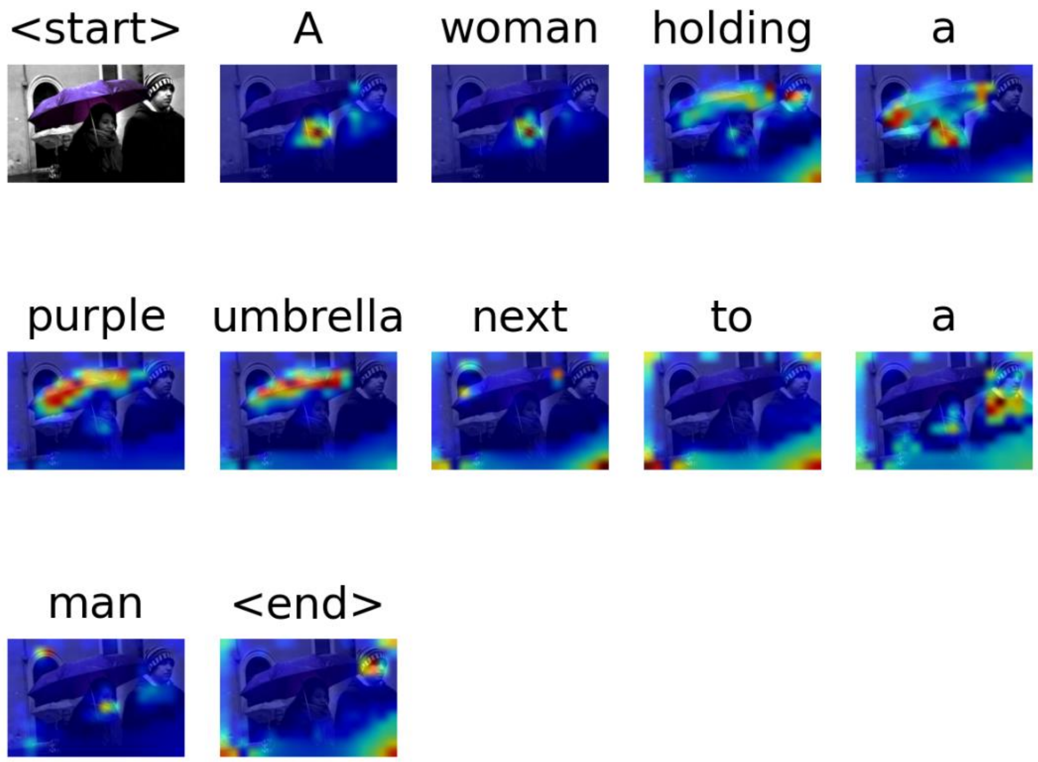
3. **Visualization of Attention in Image Captioning (20%)**
   1. **Attention maps**




<start> A young girl holding
a slice of pizza in
her hands <end>


<start> Two sheep standing next
to each other on a
lush green field <end>


<start>
Two
people
on
skis
standing
on
a
snowy
hill
<end>


<start>
A
woman
holding
a
purple
umbrella
next
to
a
man
<end>

2. **Visualize top-1 and last-1 image-caption pairs, report its corresponding CLIPScore**
   Highest CLIPScore Pair: 000000539189 - Score: 1.0101318359375

Prediction: a young man sitting on a couch holding a wii controller.

Lowest CLIPScore Pair: 000000541367 - Score: 0.439453125

Prediction: a photo of a vintage car and a computer.

3. **Analyze the predicted captions and the attention maps for each word according to the previous question**
   It appears that attention is mapped to the input tokens, and the attended region can effectively reflect the corresponding word in the caption. However, for the word "a," the heatmap is only displayed on larger regions in the image.

Reference:

1. Vision Transformer https://zhuanlan.zhihu.com/p/435636952
2. loralib: https://pypi.org/project/loralib/
3. open_clip: https://github.com/mlfoundations/open_clip
4. visual_cross-attention https://github.com/benkyoujouzu/stable-diffusion-webui-visualize-cross-attention-extension
5. PEFT csdn: https://blog.csdn.net/weixin_39663060/article/details/130724730?spm=1001.2101.3001.6650.1&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-1-130724730-blog-120255851.235%5Ev38%5Epc_relevant_default_base&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-1-130724730-blog-120255851.235%5Ev38%5Epc_relevant_default_base&utm_relevant_index=2
6. Prefix Tuning csdn: https://blog.csdn.net/qq_36426650/article/details/120255851
7. Autoregressive: https://blog.csdn.net/artistkeepmonkey/article/details/121793677

Collaborator:

Thanks to R11521613, R11942180, Lin, and Huang for the discussion, which helped complete the homework.