Deep Learning for Computer Vision HW3

Pin-Han, Huang: R10946003

November 21, 2022

1 Problem1

1. **Methods analysis**

   - The advantage of CLIP model is combining images and texts to train the model, which a wide variety of text annotations can be gathered. Images and texts are projected to a specific embedding space to compute similarity. Traditional models require labeled images in specific tasks while CLIP does not. The generality of CLIP is therefore better than previous methods.

2. **Prompt-text analysis**

   - Original template 'A photo of a ...' reaches the highest accuracy of 71.12, while template 'No ..., no score' has lowest accuracy of 54.96.

Prompt templates	Accuracy
A photo of a {}:	71.12
This is a photo of {}:	60.84
This is a {} image:	67.88
No {}, no score:	54.96

Figure 1: Model performance of different prompts.

3. **Quantitative analysis**

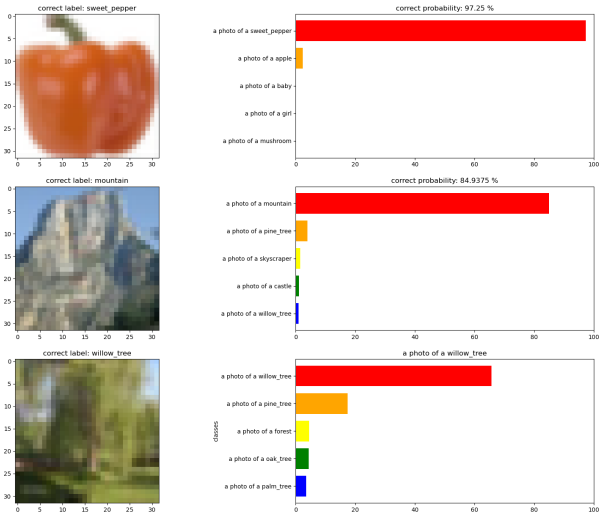

Figure 2: Sample three images from the validation dataset and then visualize the probability of the top-5 similarity scores.

2 Problem2

1. **Report best performance of CIDEr and CLIPScore**

   - The best performance of the model with pretrained backbone and encoder has CIDEr: 0.828 and CLIPScore: 0.694

2. **Report 3 other settings of experiment.**

   - Under full pretrained setting, the model achieves 0.937 CIDEr and 0.712 CLIPScore. Under pretrained and freezing ResNet backbone and transformer encoder, the model achieves 0.858 CIDEr and 0.681 CLIPScore. Under a simplified model architecture with scratch, the model achieves 0.803 CIDEr and 0.674 CLIPScore.

3 Problem3

1. **Visualize attention maps of five images**

   - The image is shown in figure 3 to 7.

2. **Visualize the top and the last score image caption pairs.**

   - The image is shown in figure 8 and figure 9. The last-1 image has clipscore 0.329, while the top-1 image has clipscore 1.043.

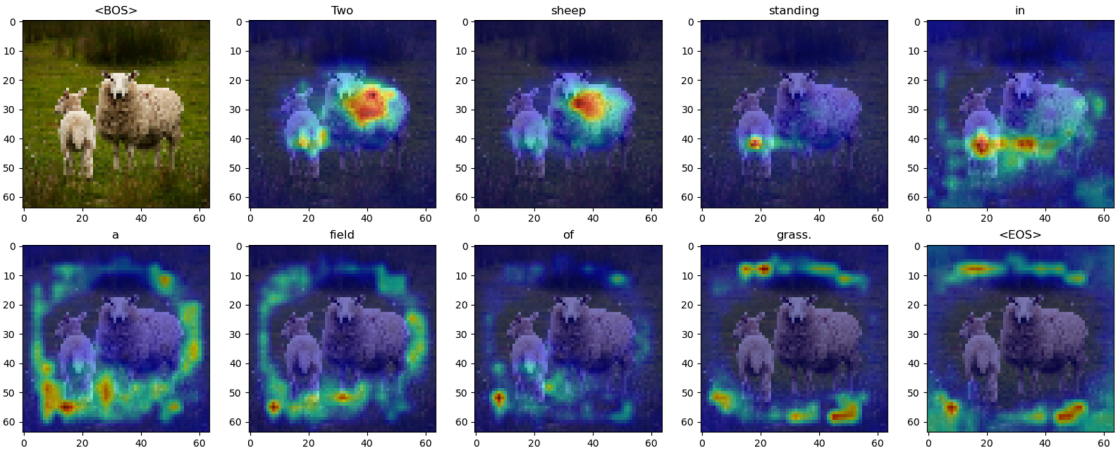

Figure 3: The sheep image.

3. **Analyze the predicted captions and the attention maps for each word.**

   - The result in the top-1 image is more reasonable, where words like 'a', 'boy', 'couch', 'playing', and 'video' are quite accurate. However, some of the captions in the last-1 image are poor, including 'hat', 'holding', 'gun', 'woman', 'standing', and 'boat'.

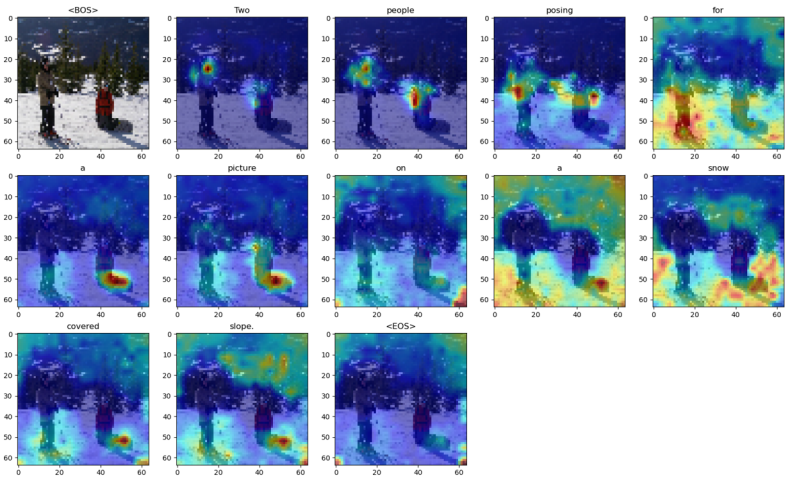

Figure 4: The ski image.

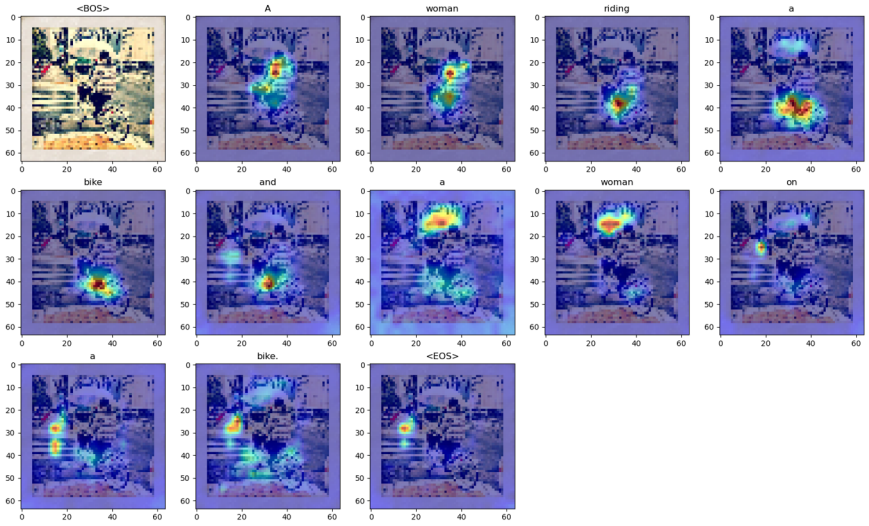

Figure 5: The bike image.

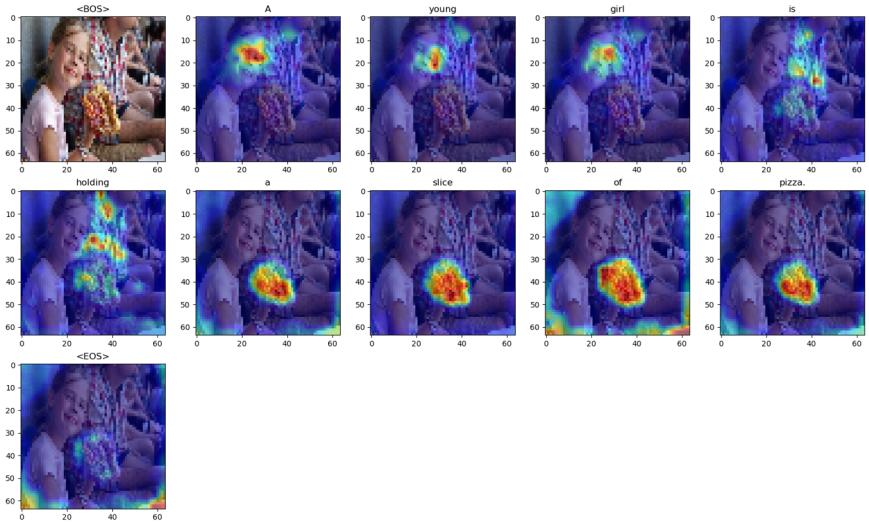

Figure 6: The girl image.

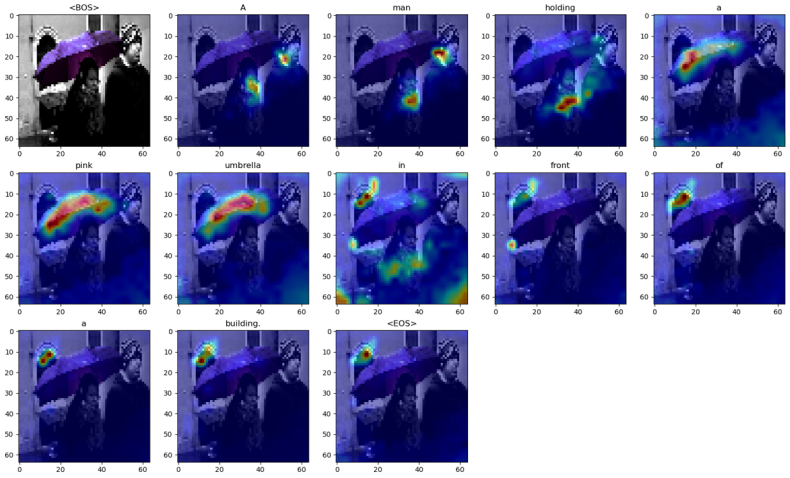

Figure 7: The umbrella image.

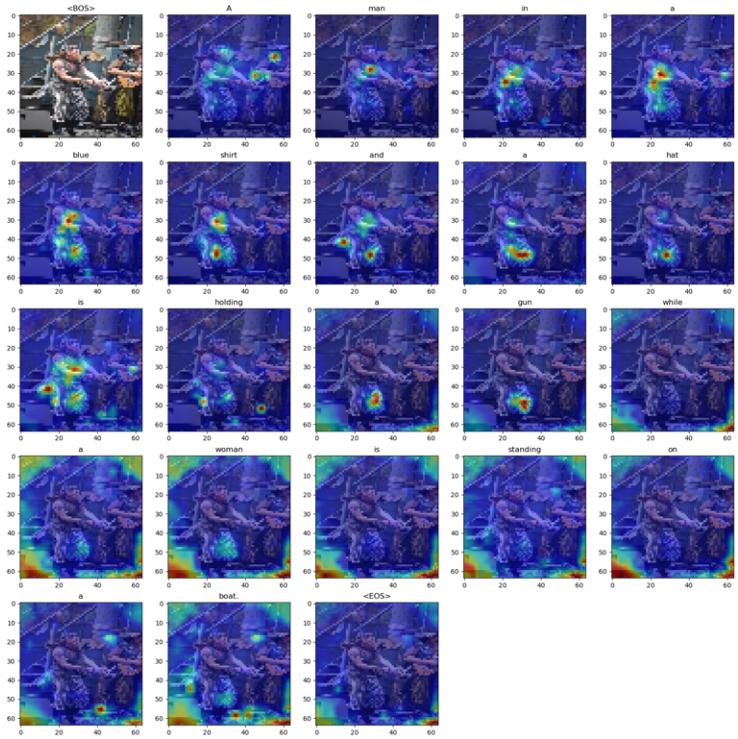

Figure 8: The man image.

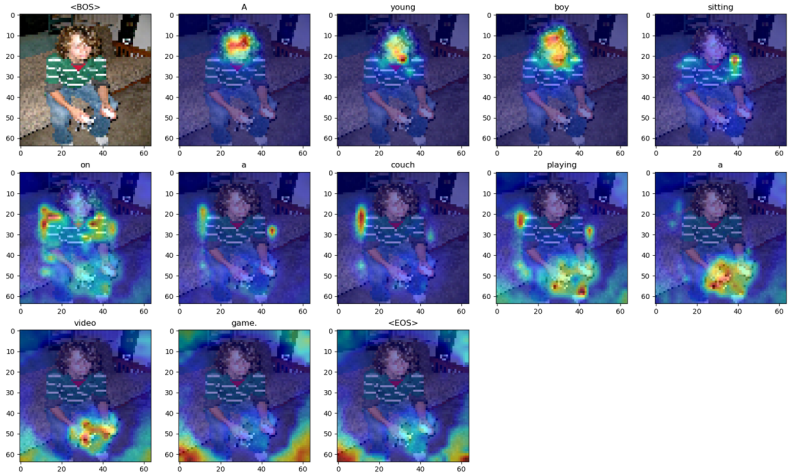

Figure 9: The video game image.