CS272 - Computer Vision II, 2022-2023 Fall

Assignment #2

T.A. Huo Chaofan (2021233274)

Acknowledgements:

1) Deadline: **2022-12-04 23:59:59**.
2) You should write you report in English using given LaTeX template. Please submit your assignment report in PDF format via Gradescope.
3) Please write your codes using provided code templates (https://github.com/MoChen-bop/assignment2).
4) You should submit your codes and evaluation logs according to the requirements stated in README.md of code templates. Please upload your code to ShanghaiTech cloud disk (http://pan.shanghaitech.edu.cn/cloudservice/outerLink/decode?c3Vnb24xNjY3NzI4MDk2OTE0c3Vnb24=) and remember to rename your zip file to CS272_NAME_ID_hw2.zip.
5) **Plagiarism or cheating is strictly prohibited**. **DO NOT** share your assignment with your classmates. You can refer to existing codes in Github but mark clearly where you refer to it.

Task 1. (30 points) Image captioning aims at give a short desciption of an image using properly formed English sentences. Show and Tell proposes the generic neural encoder-decoder framework for image captioning. The framework typically contains a CNN as encoder and a RNN as decoder. Given an image $\mathbf{I}$, encoder extracts global lattent semantic feature from it and the caption is decoded from this semantic feature in a auto-regressive manner.

$$\mathbf{h}_{t+1} = \text{LSTM}(\mathbf{h}_t, \mathbf{s}_t),\ t \in \{0, \dots, L-1\} \tag{1}$$

where $\mathbf{h}_t$ is the hidden feature of LSTM in t-th step, $\mathbf{s}_t$ is the state feature of LSTM in t-th step. The inital state $\mathbf{s}_0$ is initialized using CNN features $\mathbf{s}_0 = \text{CNN}(\mathbf{I})$ and $\mathbf{h}_0$ is embedded using the start token <START>. The $t+1$-th word in caption is predicted by

$$\mathbf{p}_{t+1} = \text{MLP}(\mathbf{h}_{t+1}) \tag{2}$$

$\mathbf{p}_{t+1}$ is a vector with the length of vocabulary size in dataset indicating the probability for each word.

Given an image and the corresponding caption, the encoder-decoder model directly maximizes the following objective:

$$\theta^* = \arg\max_{\theta} \sum_{(\mathbf{I},\mathbf{y})} \log p(\mathbf{y}|\mathbf{I};\theta) \tag{3}$$

where θ are the parameters of the model, $\mathbf{I}$ is the image, and $\mathbf{y} = \{y_1, \dots, y_t\}$ is corresponding caption. In traing stage, we optimize the negative log likelihood of the correct words at each step

$$L(\mathbf{I}, \mathbf{y}) = -\sum_{t=1}^{L} \log \mathbf{p}_t(\mathbf{y}_t) \tag{4}$$

where $\mathbf{y}_t$ is the index of t-th word in ground-truth caption.

We have provided an implementation of this basic encoder-decoder framework described above. In this task, you are required to follow the instructions of provided codes, setup experiment environment, download the COCO 2014 dataset, preprocess the dataset and train the base model (20 points). You can try one of following techniques or other techniques to improve the performance of base model at least one BLEU-1 point (10 points).

- Use more deeper and powerful backbone.
- Adjust the hyper-parameter of the base model.
- Use glove word vector to embed tokens.
- Use beam search instead of greedy search.
- Finetune the parameters of backbone after training LSTM.

Please explain how you improve the performence of the base model and fill the table below.

Answer *Show how you improve the performance of the base model here.*

Table I

TASK 1: IMPROVE THE PERFORMANCE OF THE BASE MODEL ON COCO 2014 CAPTIONING TASK.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
base model	-	-	-	-	-	-	-
base + your techque 1	-	-	-	-	-	-	-
base + ...	-	-	-	-	-	-	-

Task 2. (30 points) Show, Attend and Tell introduces an attention based model for image caption. In this task, you are required to implement one variance of this attention-based model[1]. The core this attention-based model is how to adaptively select features from visual feature map. Please read the descriptions of the spatial attention model in this paper, and implement the model depicted in figure 2(b) of original paper (20 points). You can refer to existing implemented codes but you need to adapt them into the code framework provided by this assignment cleverly. We recommand you write it by yourself, since it may take more effort to adapt existing codes into our code framework. You don't need to follow the training details in the original paper. Please follow the training process of base model provided in task 1. The model converge quickly if we freeze the backbone and can achieve 0.64 BLEU-1 score after one epoch. In report, you should report final performance of attention-based model (maybe a slight improvement) and compare it with base model and visualize the generated captions and attention map using visualization tools which have been provided in code templates (10 points).

Answer

Table II

TASK 2: COMPARSION OF ATTENTION-BASED MODEL AND BASE MODEL.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
base model	-	-	-	-	-	-	-
attention-based model	-	-	-	-	-	-	-

[1] Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. In *CVPR* 2017.

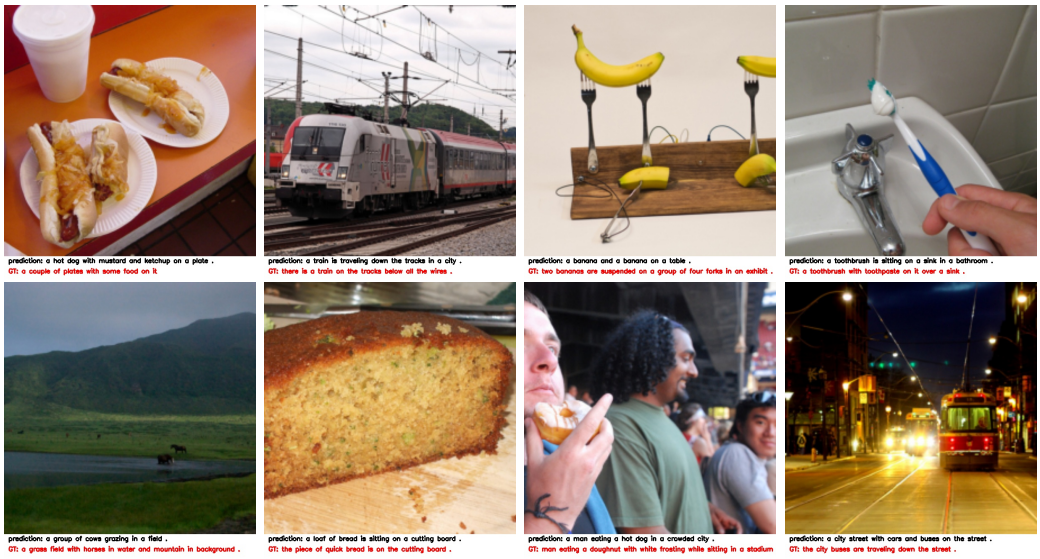

Figure 1. Generated captions.

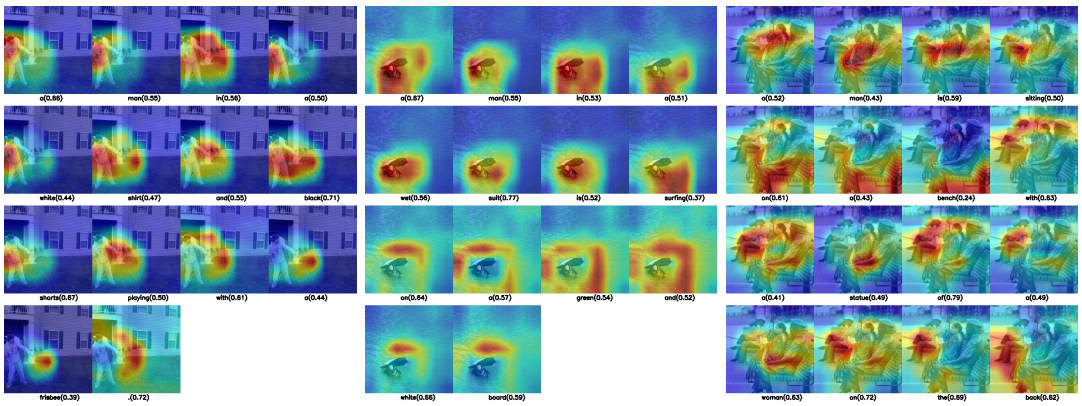

Figure 2. Attention Visualization.

Task 3. (40 points) In this task you are required to try other attention mechanisms such as multi-head attention in Transformer, channel attention in Squeeze-and-Excitation Networks, channel-spatial joint attention in Triplet Attention or other attention variance. You need to adapte one of them into base model (30 points) to check if there is performance gain (10 points). Here, we provide several directions:

- Attention is All You Need. In *NIPS* 2017.
- Squeeze-and-Excitation Networks. In *CVPR* 2018.
- Rotate to Attend: Convolutional Triplet Attention Module. In *WACV* 2021.

In report, you should write clearly the motivation of your selected attention mechanism and how you adapt it into our image captioning framework. (If it sounds making sense, you will get full 30 pts.) And also the performance the attention-augmented model should be compared with the base model. (If there is performance improvement, you will get the rest 10 pts.)