Lab3 - Caption Generation with Visual Attention

Wen-Jie Tseng (0556146)
Deep Learning and Practice - 2018 Spring Semester

1. INTRODUCTION

In this Lab, We are asked to run a caption generator by using CNN and Recurrent Neural Networks (RNN) language generator. I train this model on COCO dataset 2014 with show, attend and tell model and generate a caption to describe each given image.

In Show, Attend and Tell model, they combine the encoder-decoder framework with attention mechanism, which became a state-of-art method in image captioning. For the encoder part: the model extracts features by CNN, generates low-level features so the decoder can focus on different part of image. In decoder part, they use LSTM, RNN to generate sequence of caption.

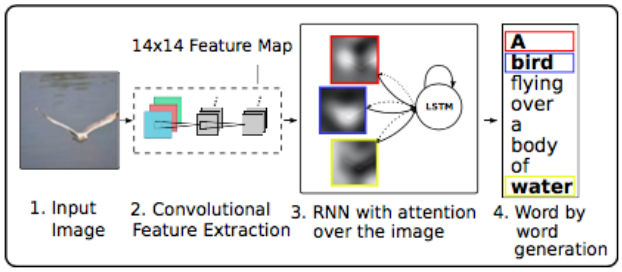

Figure 1. The work flow of Show, Attend and Tell model.

2. EXPERIMENT SETUP

- I only tried show attend tell model
- batch size: 10
- input encoding size: 512
- RNN size: 512
- att hid size: 512
- fc feat size: 2048
- att feat size: 2048
- rnn type: LSTM
- feature map: 14 x 14
- Train with COCO 2014 dataset

To produce the attention over time, one has to find weight of the model. I started tracing it from **ShowAttendTellCore**. I return the weight inside its forward function. Then I collect the returned weight in **sample**, **sample_beam**, **forward** and **get_logprobs_state** of **OldModel**. Also needed to record the weights in **CaptionModel**. This is where caption model generate the attention of each word. After that, we can obtain attention overtime and plot them in eval_utils.py.

3. RESULT

The training loss curve of show, attend and tell model was shown as Figure 2. The minimum loss reached 1.7147. Follow the instruction from sample code provided in Lab 3, I obtain the output images in Figure 3 and Figure 4. Most of the captions were corresponded to the given image, but some of them still had error 4. The results of attention over each word of caption are shown in Figure 5 and 6.

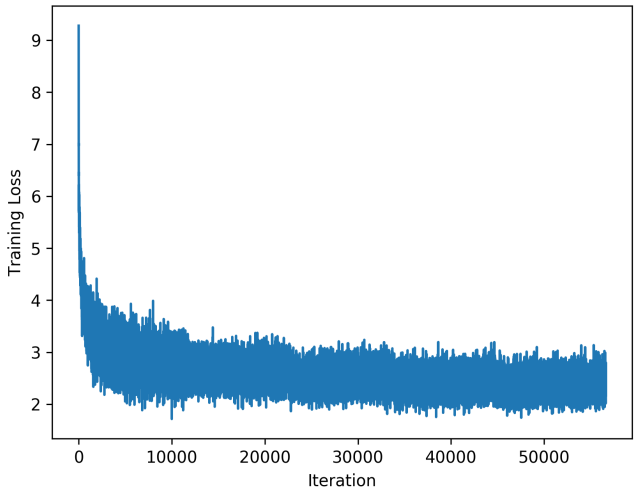

Figure 2. The training loss curve of show, attend and tell model.

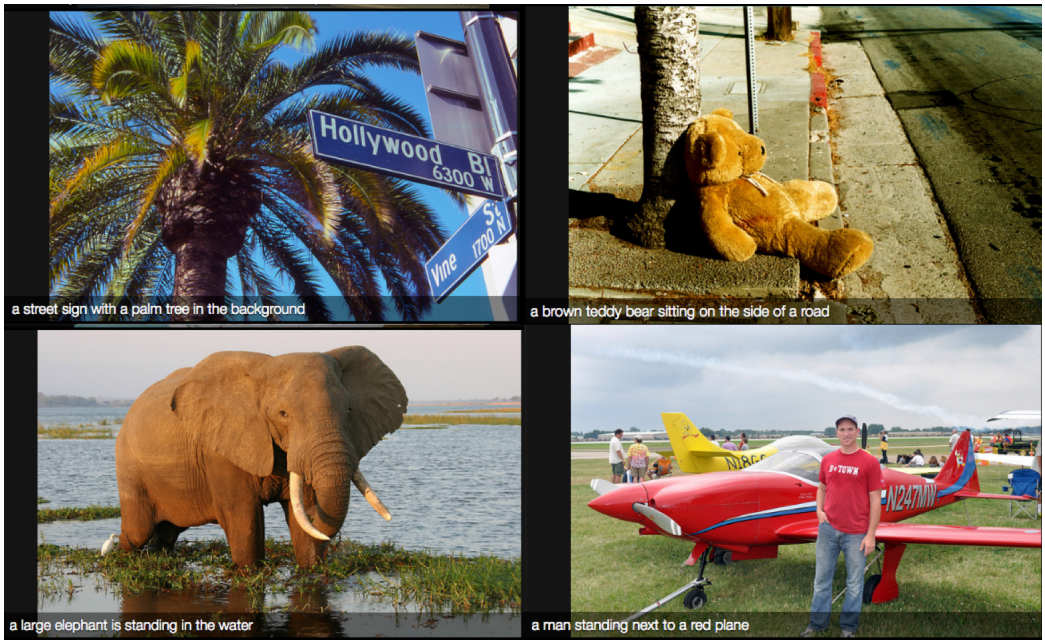

Figure 3. The output result. Most of the time, the caption model can produce correct result.

Figure 4. However, some of the result did not perform well. There is one elephant in the left image and no shower in the right image.

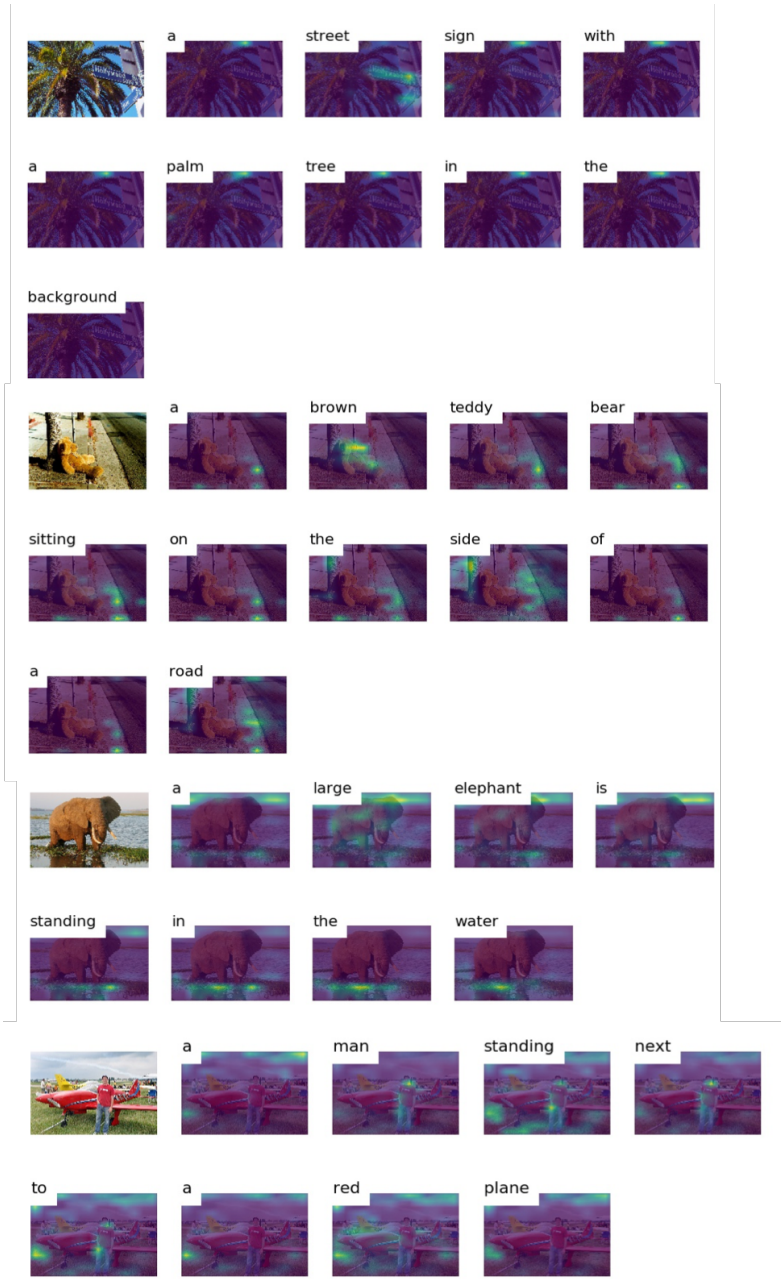

Figure 5. The attention over time of selected images.

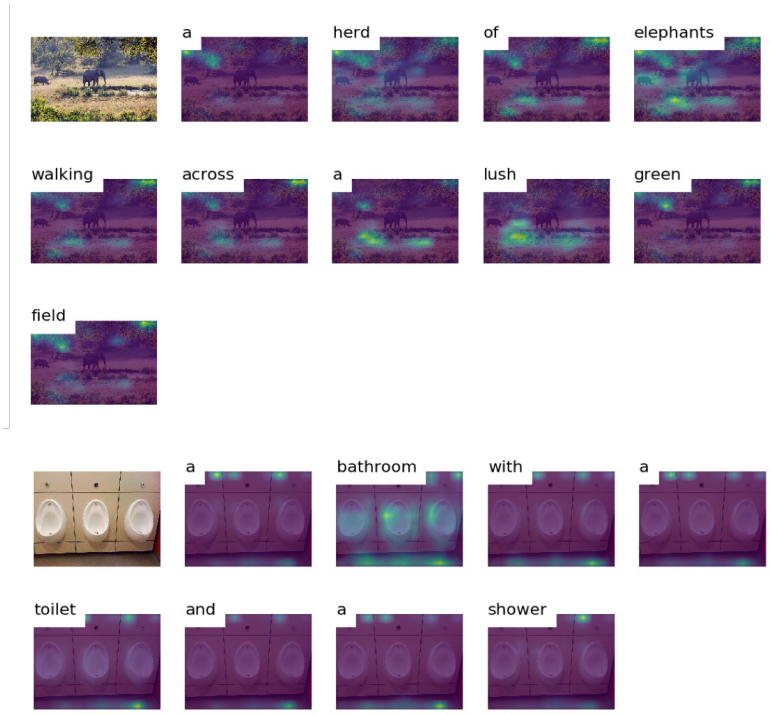

Figure 6. The attention over time of selected images. Can observe some attentions focused on incorrect places.

4. DISCUSSION

The paper proposed two attention mechanisms, hard attention and soft attention. In the Stochastic "Hard" Attention, weight represents the probability of an image region was chose as an input of decoder at time t. This variable follows a Multinoulli distribution. In Deterministic "Soft" Attention, weight represents the portion of information that an image region was chose as an input of decoder at time t. Since stochastic attention requires sampling the attention location each time, instead they take the expectation of the context vector directly. The deterministic attention model is an approximation to the marginal likelihood over the attention locations.

REFERENCES

1. An introduction to show, attend and tell
2. An introduction to show, attend and tell
3. Show, Attend and Tell website.
4. Sample code provided in Lab3.