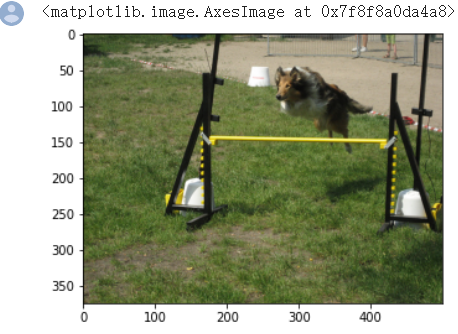
**COMP5623 Coursework on Image Caption Generation**

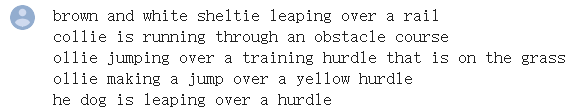
Fanhui Meng (sc19fm) 201373470

1. Lemmatize means to restore a language vocabulary of any form to a general form (can express complete semantics), which means people can use fewer words to express the meaning of a sentence. So the sentence made of lemmatize words can be short. It can be much more efficient and useful when applying this method to information retrieval and text and natural language process. As for the regular sentence, it may contain a lot of meanless words, and it may cause inefficiency during the natural language process.

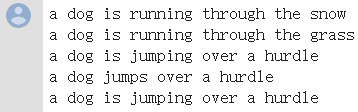
However, the process of lemmatizing words can be complicated and hard to implement. It needs to return the original form of the word, analyze the form of the word, etc. No one can guarantee those processes are accurate and perfect. Especially for those complicated processes, which are more likely to get wrong and cause imprecise during the natural language process.

1. Below is the first image and it’s original captions.

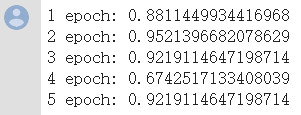




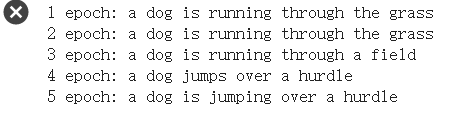
Here’s the generated captions after each epoch of training for RNN decoder:



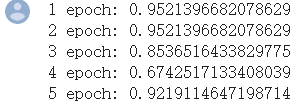
And their BLEU scores for RNN decoder:



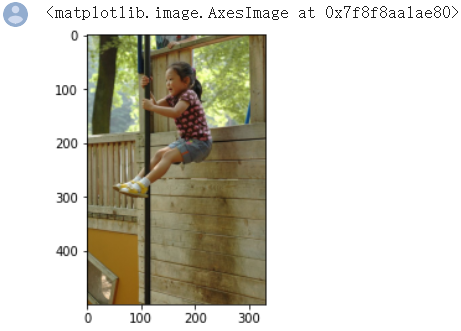
Here’s the generated captions after each epoch of training for LSTM decoder:

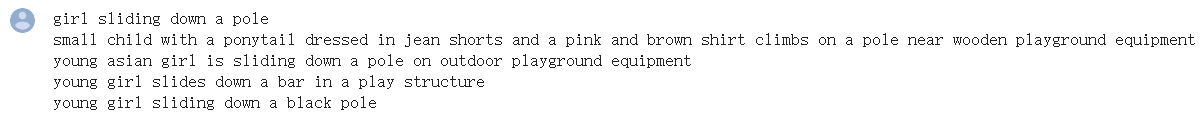


And their BLEU scores for LSTM decoder:

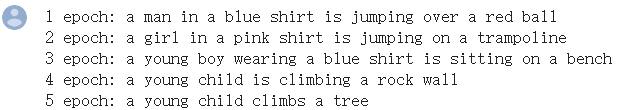


Below is the second image and it’s original captions.

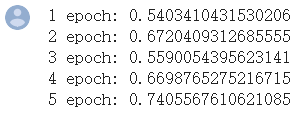




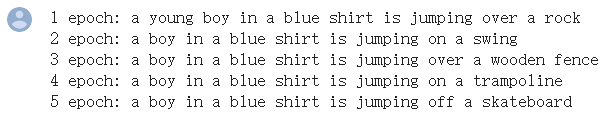
Here’s the generated captions after each epoch of training for RNN decoder:



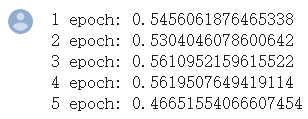
And their BLEU scores for RNN decoder:



Here’s the generated captions after each epoch of training for LSTM decoder:



And their BLEU scores for LSTM decoder:



1. As for loss, I copy these loss into 2 txt files ,so that I can compare them easily.

At the beginning, both RNN and LSTM are start with about 8.1 loss. After the first three epoch training, the loss are quite close, but the RNN network seems has lower loss.

(Left is LSTM, and the right hand side is RNN)







And after epoch 4 and 5 training, the LSTM has lower loss.





In conclusion, the loss of RNN training would drop faster in the first 3 epoch, after this, the LSTM loss would be lower. And the LSTM loss would be more stable, it goes around 2.4 and 2.2 in the final epoch. The RNN loss would be more fluctuant,it goes around 2.6 and 2.2 in the final epoch.

I pick 2 picture in the test set, and test their generated captions and their score.

The first picture has 5 short sentences, and both RNN and LSTM have good score in this case.

The best score of RNN is approximate 0.95214, and the average is 0.870273

The best score of LSTM is approximate 0.95214, and the average is 0.870818

For the picture with clear features and short original sentences. Both RNN and LSTM can have good score in this case. LSTM seems has better score accoding to the average score, but their score are really close.

In terms of the quality of generated captions. Both RNN and LSTM captions describe the feature of the image correctly. However, the first sentence generated by RNN recognise the white object wrongly as snow. And the LSTM bare make mistakes. In conclusion, the qualities of RNN and LSTM are both good and closed, but the LSTM has a little bit better performance in this case.

The second picture that I pick from the test set as well as the original five sentences are more complicated than the first one.

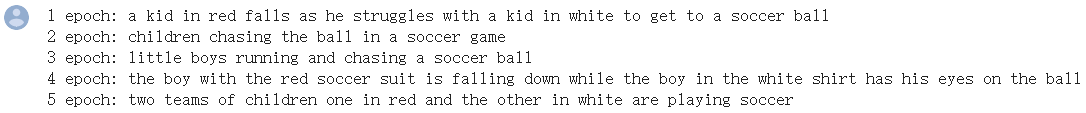
The best score of RNN is approximate 0.74056, and the average is 0.63636

The best score of LSTM is approximate 0.56195, and the average is 0.53311

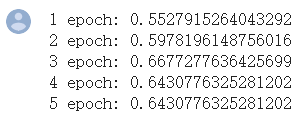
It’s quite clear that both RNN and LSTM have much lower score than the first simple image. And RNN have better score than LSTM, which is strange to me. Theoretically, RNN has trouble with long term memory due to the possibility of it’s gradient is particularly small or large. And LSTM can prevent gradient disappearance and has long term memory. So LSTM should preform better than RNN in long caption case.

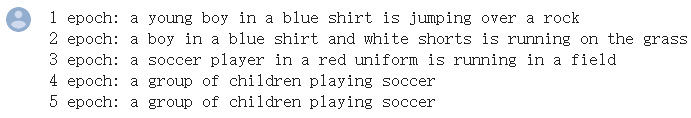
And I do some extra experiments on this, trying to find the reason.

For example, I use another image from the test set, and here’s the result.

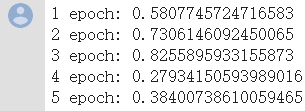
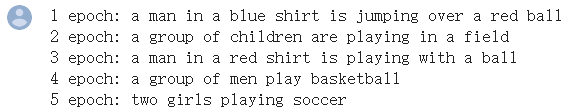


LSTM score and generated captions as below.





RNN score and generated captions as below.

In the aspect of BLEU score. The LSTM score tend to be more stable, which means the best and worst scores are close. And RNN scores are unstable. It’s best score can be very high, and the worst are pretty low.

In the aspect of generated sentences quality. LSTM can generate long sentence to decribe one image. And the RNN generated sentences are quite short, which means it contain less information and sometimes it will get low score due to lack of information.

1. One picture can have different descriptions. Some of them can be simple and short, others can be long and detailed. They are both right, but people want the generated sentences are as fully detailed as possible. Otherwise, some models can always generate some indistinct sentence and get good score.

In the ExpertAnnotations.txt, it come up with a ranking of each sentences. When evaluate model, we can put weight on each sentences. For example, if the generated captions match the most with the sentence with score 4, then this model performed extremly well. If the generated captions only match the sentence with score 1, then we think this model is not good enough.

In the CrowdFlowerAnnotations.txt file, has also created a ranking system. The sentence with the most yeses from the judgments must be the most valid sentence. Then put more weight on the most valid sentence when evaluate the model.