Image Captioning with Several RNN Models
MATH 6380O Mini-Project 2

ZHANG Hongming[1], ZHANG Jianhui[1], ZHU Weizhi[2], FAN Min[2]
[1]Department of Computer Science and Engineering, [2]Department of Mathematics, HKUST

ABSTRACT

In this project, we used several Recurrent Neural Network (RNN) models to do image captioning with MSCOCO, which is a popular large-scale dataset for image captioning, object detection and segmentation.

Three RNN models were adopted including Long Short Term Memory (LSTM), LSTM with attention mechanism and one of the-state-of-art model, LSTM with ARNet, which was accepted by CVPR 2018.

Compared with LSTM, the performance of LSTM with attention mechanism and LSTM with ARNet is much better. Moreover, we randomly picked some examples in the test dataset of MSCOCO to demonstrate the advantages of two modified LSTM models.

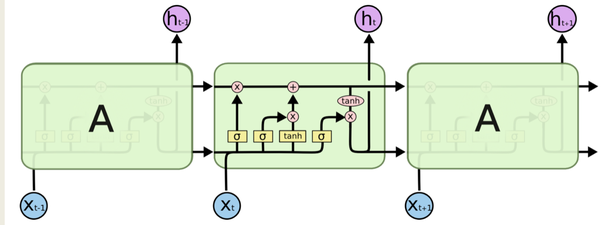
INTRODUCTION

Image captioning is drawing more and more attention from both computer vision and natural language processing communities. The task is to predict a syntactically and semantically correct target sequence consisting of consecutive words based on the provided source information.

To achieve this goal, we implemented three Recurrent Neural Network (RNN) models, which are Long Short Term Memory (LSTM)[1], LSTM with attention mechanism[2] and one of the-state-of-art model, LSTM with ARNet[3] which was accepted by CVPR2018. In the project, we used the most popular MSCOCO dataset instead of Flicker8K.

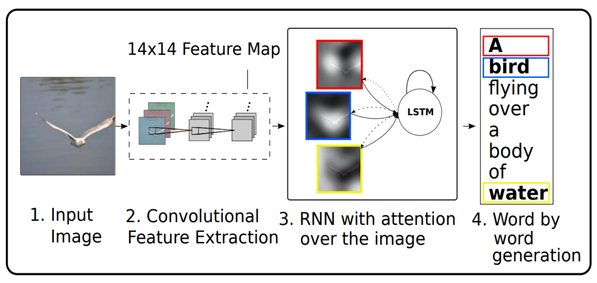
LSTM

The idea of long short term memory was introduced by Hochreiter in 1997. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. Intuitively, they can be thought as regulators of the flow of values that goes through the connections of the LSTM.



LSTM WITH ATTENTION

Attention is simply a vector, often the outputs of dense layer using softmax function. And attention mechanism and LSTM were combined together in [2]. The model with attention could automatically learn to describe the content of images, whereby improve the performance substantially.



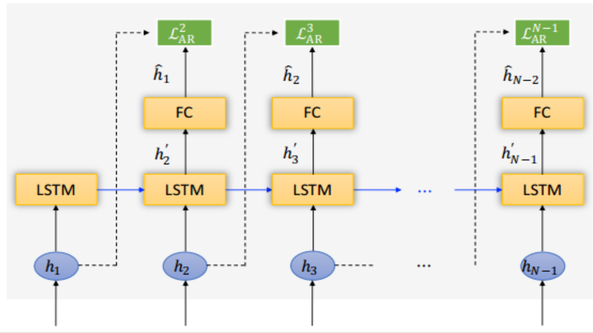
LSTM WITH ARNET

In this project, we also implemented the current state-of-the-art model: LSTM with ARNet. Similar to Attention model, ARNet is an additional structure which can be easily applied to any existing frame works. From the experiment result, such structure can improve the performance of all of the models.

The motivation of ARNet is very straightforward, it tries to strengthen the connection between neighboring hidden states by reconstructing the past with present.

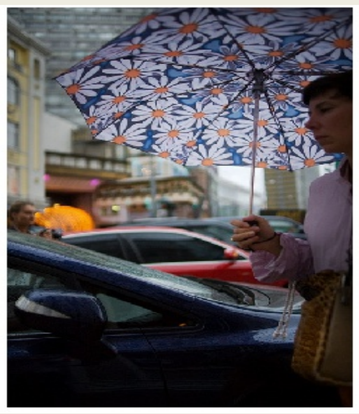
As shown in the following graph, it takes one additional LSTM decoder as the basic structure. Among that, ARNet adds a new framework to allow the hidden state h(i) to predict its previous hidden state h(i-1).

This whole process simulates how people go back to check their old knowledge when they are thinking and experiment results prove that such frame work can help improve performance of nearly all the base networks.



EXAMPLES

LSTM: men of on a blue
Attention: A group of people sitting on a wooden bench.
ARNet: ['a couple of people sitting on a park bench',
'a person sitting on a bench in a park',
'a couple of people sitting on a bench in a park']



LSTM: The man is in outside with outside
Attention: A man standing in front of a motorcycle
ARNet: ['a woman holding an umbrella in the rain', 'a woman holding an umbrella in front of a car',
'a woman holding an umbrella in front of a building']

LSTM: of A children with a red red shirt in a field .
Attention: a small boat is parked in the grass
ARNet: ['an old truck is parked next to a boat',
'an old truck is parked in the grass',
'an old truck is parked next to a small boat']

ANALYSIS

For the attention model, the sentence structure seems to be more reasonable comparing with pure LSTM, since RNN knows where to look at in the next time. We seldom saw a sentence totally meaningless.

Compared with the pure attention model, ARNet adds a new layer to strengthen the relation between neighboring words. As we can see from the second example, a boat is not supposed to be parked. ARNet can capture such information, but the base model can't. Thus ARNet can provide better caption for this example.

However, from the experiment results, we can see that some pictures are still very hard for the state-of-the-art model. The reason behind is that those pictures require some commonsense knowledge to understand. This also show the potential improvement room of our future work.

Result

From table below, we can see the results of LSTM with attention mechanism and LSTM with ARNet are much better than LSTM itself.

Model	BELU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
LSTM	0.657	N/A	N/A	N/A	N/A
Attention	0.707	0.492	0.344	0.243	0.239
ARNet	0.740	0.576	0.440	0.335	0.261

REFERENCES

1. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
2. Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International Conference on Machine Learning. 2015: 2048-2057.
3. Chen X, Ma L, Jiang W, et al. Regularizing RNNs for Caption Generation by Reconstructing The Past with The Present[J]. arXiv preprint arXiv:1803.11439, 2018.

CONTRIBUTIONS

ZHANG Jianhui: implement LSTM.
ZHU Weizhi: implement LSTM with attention.
ZHANG Hongming: implement LSTM with ARNet.
FAN Min: write the poster.