Image Caption Learning

Yulin Shen, Yiyan Chen, Shenghui Zhou
[1] *New York University-Center for Data Science*
ys2542,yc2462,sz2396(@nyu.edu)

Abstract

In this project, we aimed to implement models that convert an image into a natural language description. The proposed method construct input into encoder-decoder framework, which conduct image encoder with Deep Residual Convolutional Neural Network on ILSVRC-2012-CLS image classification dataset and text decoder with Elman RNN, long short-term memory (LSTM) network and Gated Recurrent Unit (GRU).

Introduction

Image caption, which contains a short description about explaining or elaborating on pictures, is common used on publish photographs or exhibitions. Moreover, the application of provide a accurate, syntactically reasonable text description for the a picture and find all the important features is also significant for many uses, for instance, we can generate descriptions for movies or videos, transit pictures to words for further use.

In this project, we want to implement encoder-decoder framework which gives accurate and well structured image caption automatically. In the encoder part, we plan to implement three CNN models (ResNet-34 & ResNet-101 & ResNet-152[6]) respectively, compare and performances of each models and eventually gives best image caption. In the decode part, we use Elman RNN, long short-term memory (LSTM) network and Gated Recurrent Unit (GRU), we compare the 9 combination results and evaluate each model with Bleu score, then select a most significant model for this project.

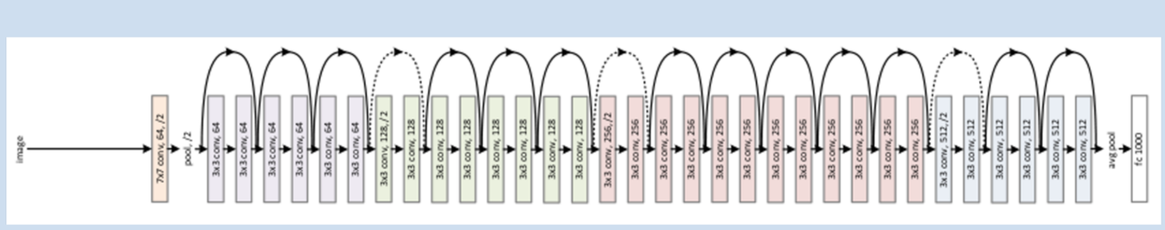
Literature Review

It has been studied for a long time in computer vision for the language description of the videos, which has a complex system composed structured formal language on sports shows or traffic scenes[8,9]. With the development of CNN models, there are lots of applications focusing on image recognition, which is able to find the specific subjects in the pictures with high accuracy. On top of that, Li et al.[19], who started to conduct the correlation between items in the pictures and the correlation of the structured sentences, provided a more accurate description of image caption.

In the most of the recent researches of generating the description of the images are mainly involve the combination of recurrent network (RNN) and convolutional network. In the structure, CNN aim to deliver the image feature extraction as a encoder, the RNN model plans to treat the language model as decoder. In [6], the image caption is delivered by a simple recurrent neural network with LSTM, we also find several models combinations to implement image captioning, however, mots of the research choose to use LSTM as the RNN model but lose the consideration of GRU, also, there is no structured comparison or evaluation about different CNN models with Elain RNN, LSTM or GRU.

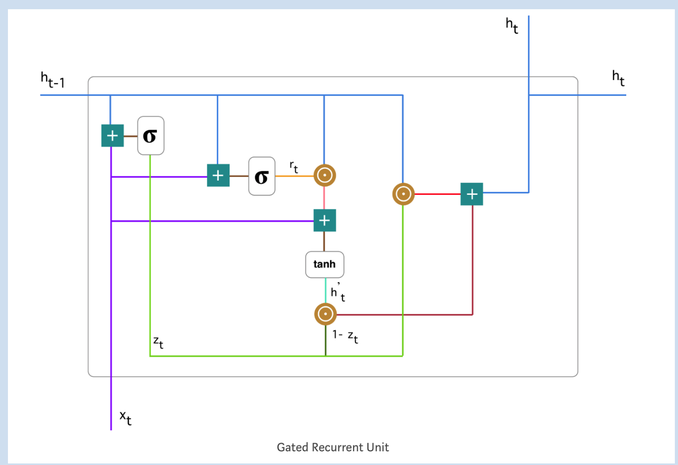
CNN Structure: ResNet152

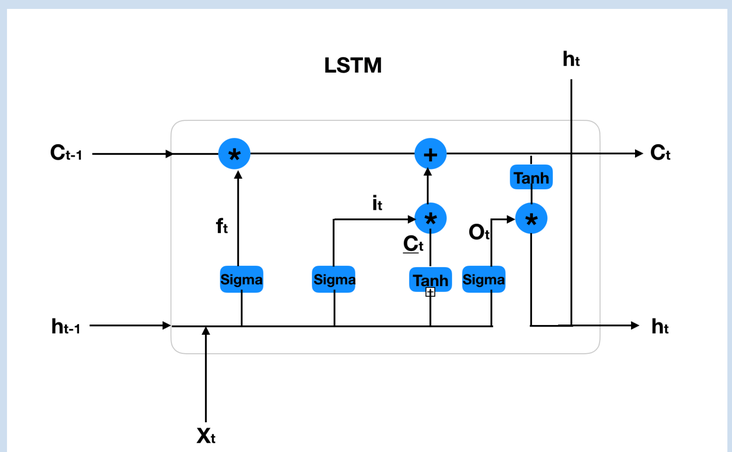
We take Resnet152 structure as an example:



RNN Structure: GRU & LSTM

We mainly talks about the structures of GRU and LSTM:





Evaluation

	ERNN & ResNet34	ERNN & ResNet101	ERNN& ResNet152
BLEU1 Score	0.645	0.639	0.643
BLEU2 Score	0.459	0.455	0.460
BLEU3 Score	0.316	0.315	0.318
BLEU4 Score	0.218	0.217	0.219

	GRU & ResNet34	GRU & ResNet101	GRU& ResNet152
BLEU1 Score	0.656	0.674	0.673
BLEU2 Score	0.475	0.492	0.492
BLEU3 Score	0.334	0.348	0.347
BLEU4 Score	0.234	0.246	0.244

	LSTM & ResNet34	LSTM& ResNet101	LSTM& ResNet152
BLEU1 Score	0.658	0.673	0.668
BLEU2 Score	0.479	0.495	0.489
BLEU3 Score	0.338	0.352	0.346
BLEU4 Score	0.239	0.250	0.244

Conclusion

- From the result, we can see that GRU 34 usually describe things in a quite simple and straightforward way compared to GRU 101 and GRU 134 which can catch more details.
- The RNN compared to other two neural networks of same layer size has higher possibility to have non-complete sentences and have captions that make no sense. GRU and LSTM with more layers tend to perform better.
- From the evaluation we can find that ResNet101 with LSTM gives the best result with BLEU Score 0.250, which is quit close to the result from Google (0.27). Moreover, all CNN models with Elaine RNN give worse results and ResNet101 provides better result with all RNN models.
.

Acknowledgements

This project was supported and advised by Prof. Lecun Yang, Dr.Mikael Henaff and Dr. Alfredo Canziani. We would like to express our thanks to their who provided insight and advice that greatly assisted this project.