Image Captioning

by
Mihir Joshi (mxj180007)
Pawan Patil (pxp180029)
Vishal Shah(vjs180000)
The University of Texas at Dallas

ABSTRACT

Image captioning is one of the demanding fields in Machine learning. We are using Recurrent Neural Network in our project which are playing important role field of image captioning. It is one of the primary components in the field of image captioning. As per title of project, we predict the captions for a given input images. For that we have to not only use image processing but also natural language processing. So in this project our program prints/ assigns caption to the images automatically. We do image analysis using featured vectors and words are used to assign a caption automatically. We have used Flickr8k dataset. The reasons behind using this dataset are: Compact size of dataset. (Compact size makes it easier to train/ upload dataset on local devices, low end devices. You have free access to dataset (you can download the same from lot of different online sites.) Also, you will get a caption/ output for most of input images because this dataset has provided 5 captions to all images present in it.

Keywords: CNN, Image captioning, RNN,

I. INTRODUCTION

In past days we used to get large number of images from many sources which users were supposed to interpret on their own. Although those images were caption-less, users could understand those images (captioning process was not necessary.)

Image captioning deals with image understanding/ processing and producing captions for those images.

1) Image understanding/ processing:
   Here algorithm is supposed to detect images and recognize all the objects in a given input images.
2) Captioning:
   Here algorithm is supposed to predict/ produce text/ captions which are not only grammatically correct but also semantically right. That is machine on which algorithm is running must understand the meaning of sentence.

For example, in past days machine were not able to understand the meaning of sentence but now semantic web concept has been developed which allows the machine to understand the text. That is machine works on intuition like human beings.

The purpose of RNN is to predict text of the given/ provided input image. The purpose of using CNN is to process all images in the Flickr8k dataset.

Basically, there are many ways to detect and caption a given image. These ways are as follows:

a) We can use computer vision algorithms to detect various object features from given input image.
b) We can use neural models for like RNN and CNN for image detection and caption prediction.
   This report includes approach (b).

As stated earlier both of the above networks plays equally important roles in the given task of image detection and predicting caption automatically.
Though tasks of RNN and CNN are mentioned above, The role of RNN depends largely on the way CNN is used. The following diagram represents inject and merge architectures for caption generation.

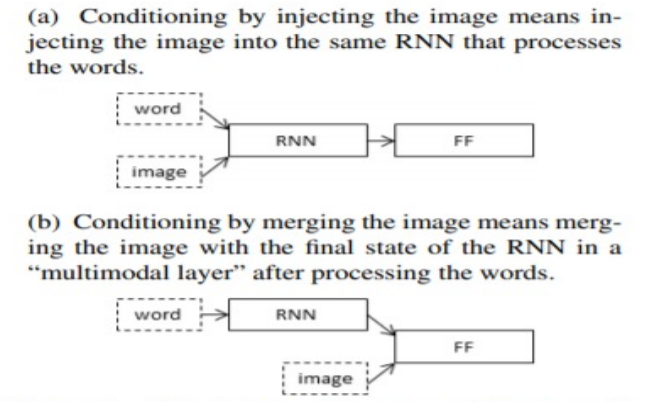


Figure 1: The inject and merge architectures for caption generation. The RNN's previous state going into the RNN is not shown. Legend: RNN - Recurrent Neural Network; FF - Feed Forward layer.

CS 6375.501 FINAL PROJECT – IMAGE CAPTIONING

Picture Courtesy: Link

II. BACKGROUND WORK

With the advance in technology human beings made it possible to recognize an object in a given input image. Although this was one of the impressive tasks, only predicting/ identifying name of the objects recognized from image was not sufficient. Because only guessing an objects' names from given images may be ambiguous. For example: consider 2 pictures one where a group of children are studying and another where group of children are playing. So, if we give these images as a input to the old object recognition systems they will only produce text as "Children" that is objects. But they could not specify actions those boys were doing. And machine cannot think on intuition so predicting actions for machines is quite difficult.

Hence as long as machine cannot think like a human being, automatic image captioning will be one of difficulties/ challenges in today's world. But now a days there are lot of advanced techniques which are trying to solve this issue. Although it is hard to make a perfect algorithm which will make the machine think like human beings (because machines will always lack the way humans communicate), as mentioned earlier a new field of semantic web is playing an important role in the development.

In last 10 years, a tremendous amount of research is done on image caption generation using deep learning technology. These learning algorithms can help us to solve problems and challenges in the field of image captioning.

Hence now days image captioning using multimodal neural net has been one of the challenging topics in the field of machine learning.

In neural models RNN detects a text and then the same network detects the next words, or it passes the word to the next level, and that layer will make prediction. This prediction is carried out in SoftMax function based on probability distribution.

RNN is used in 2 ways to produce caption:

1) Conditioning by inject
   Injecting image in same RNN
2) Conditioning by merging
   Combine the image with final output of RNN

III. CNN and RNN

RNN and CNN are most important part of our project.

CNN

CNN is a neural network where connection between node does not form a cycle that is it is a type of "Feed Forward" neural net. CNN is a multi-layer neural network. It consists of input layer, output layer and multiple hidden layer. More the numbers of hidden layer more accurate the output will be that is image detection will be more accurate. It requires a little processing, as it learns the feature by itself from input data.

RNN

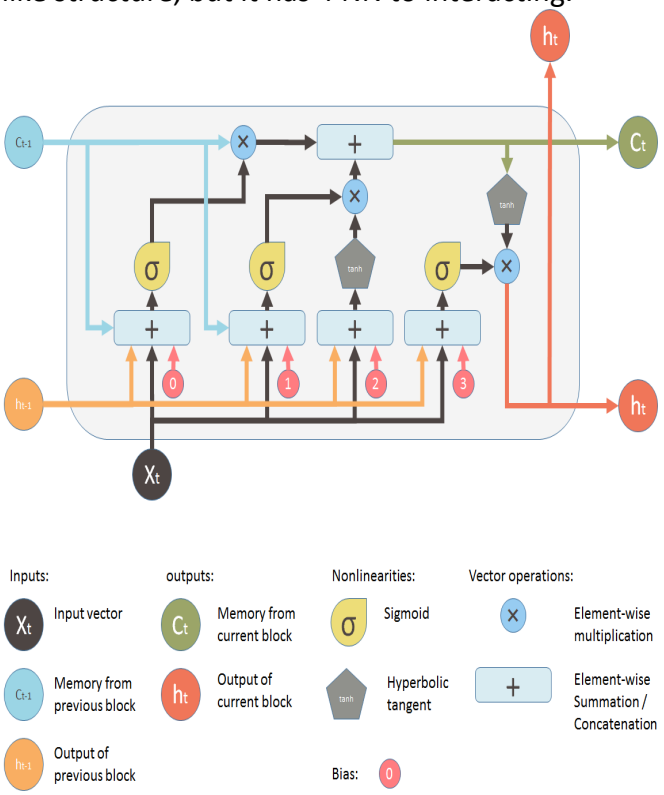
Nodes in RNN neural net form directed links. So, it is like a directed graph structure. They are responsible for handling series operation like text recognition, speech detection etc. In RNN, nodes form a loop, so information flows between loops. RNN predicts the output based on present input. It also considers previous inputs for producing outputs

IV. Dataset Details

The dataset we are using is the flickr8k.The dataset contains the about 8000 images and it also contains 5 captions per images and also there are also model weights and the encoding in the file present with the dataset

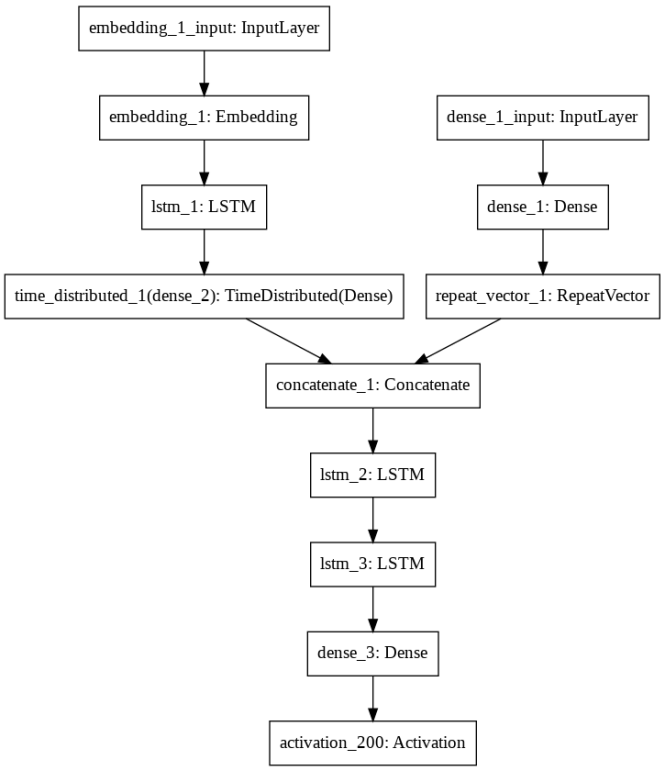
V. LSTM

RNN suffers from two issues: exploding gradient and vanishing gradient. So, LSTM were used to solve the problem. Sometimes we need to remember the input for later use, to look at the data for a specific time frame. Default nature of LSTM is to remember data for elongated period of time. LSTM has a chain like structure, but it has 4 NN to interacting.



Picture-link: https://miro.medium.com/max/2312/1*laH0_xXEkFE0lKJu54gkFQ.png

VI. Model Structure

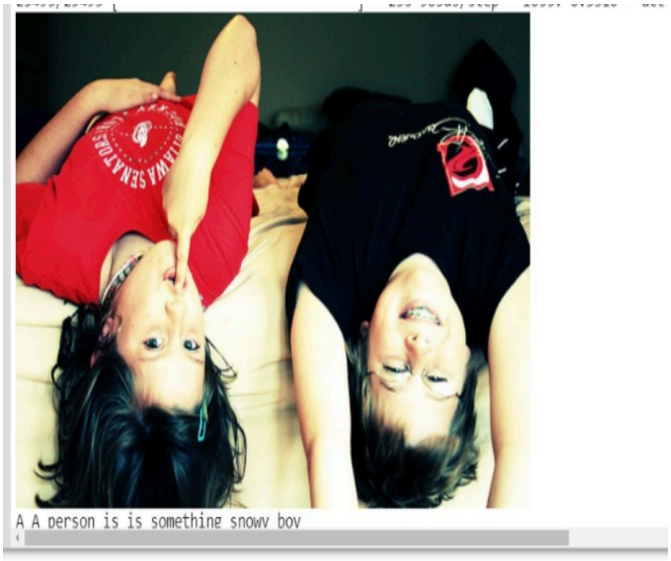


VII. RESULTS AND ANALYSIS

After number of experiments each time with different number of epochs and batch size. We also cannot get perfect caption because of small dataset and model weights used. We were using google colab to run the project for different epoch sizes such as 50,100,150,200,250,300,500,1000.Till 250 epoch the result was not good enough but after doing for more than 300 epoch we were getting a good caption for the images and hence the results aren't accurate and the model gives random word in captions

The below image is after 150 epoch with a batch size of 512

The below image is after 175 epochs with a batch size of 256



The below image is after 300 epochs and a batch size of 512

Analysis of Results:

The analysis is done by creating a caption for the an unknown image from the dataset. The captions were different after different epoch for the same image

VIII. Conclusion and Future Work

The results still are not very accurate. The LSTM models with help of RNN overcomes shortcoming which solves the vanishing and exploding gradients. We can change our model and create its depth by using many layers and we could get more data and captions set for better results

XI. References

[1] Wikipedia: Long short-term memory.

[2] Medium: Understanding LSTM and its diagrams.

[3] Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network [https://arxiv.org/pdf/1808.03314.pdf]

[4] https://arxiv.org/pdf/1708.02043.pdf

[5] Image Retrieval Using Image Captioning [https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1686&context=etd_projects]

[6] Kaggle

[7]Image captioning[https://cs.stanford.edu/people/karpathy/sfmltalk.pdf]

CS 6375.501 FINAL PROJECT – IMAGE CAPTIONING