Image Caption Generator with CNN & LSTM

Report submitted in fulfillment of the requirements
for the Course Project of Data Mining

Third Year B.Tech.

by

Bharat Kumar,Ayush Singh,A Agrawal,ABP rastogi

Under the guidance of

Dr. Bhaskar Biswas

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI
Varanasi 221005, India
November 2020

Dedicated to

Our parents, teachers,.....

Declaration

I certify that

1. The work contained in this report is original and has been done by us and the general supervision of our supervisor.

2. The work has not been submitted for any project.

3. Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

4. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: IIT (BHU) Varanasi
Date:20 November 2020

Bharat Kumar,Ayush Singh,A Agrawal,ABP rastogi
B.Tech.
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Certificate

*This is to certify that the work contained in this report entitled "**Image Caption Generator with CNN & LSTM**" being submitted by **Bharat Kumar,Ayush Singh,A Agrawal,ABP rastogi** (**Roll No. 180750/16,15,11,13**), carried out for the project of Data mining course in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of our supervision.*

Place: IIT (BHU) Varanasi
Date:20 November 2020

Dr. Bhaskar Biswas
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Acknowledgments

We would like to express our sincere gratitude to the people who have helped us the most throughout our project.I am grateful to my project supervisor(Dr. Bhaskar Biswas) for providing me an opportunity to implement the caption generator using "CNN and LSTM" and his constant support for the project.

Place: IIT (BHU) Varanasi

Date:20 November 2020 **Bharat Kumar,Ayush Singh,A Agrawal,ABP rastogi**

Abstract

Basically,we are implementing an image caption generator.Image caption generator is a task that involves computer vision and natural language processing concepts to recognize the context of an image and describe them in a natural language like English. This Project deals with generating image caption using CNN and LSTM.Here we used 8k images in which 6k images are used for training our model and rest 1k images are used for testing.Remaining 1k images are used for devlopment of our model.

Contents

Chapter 1

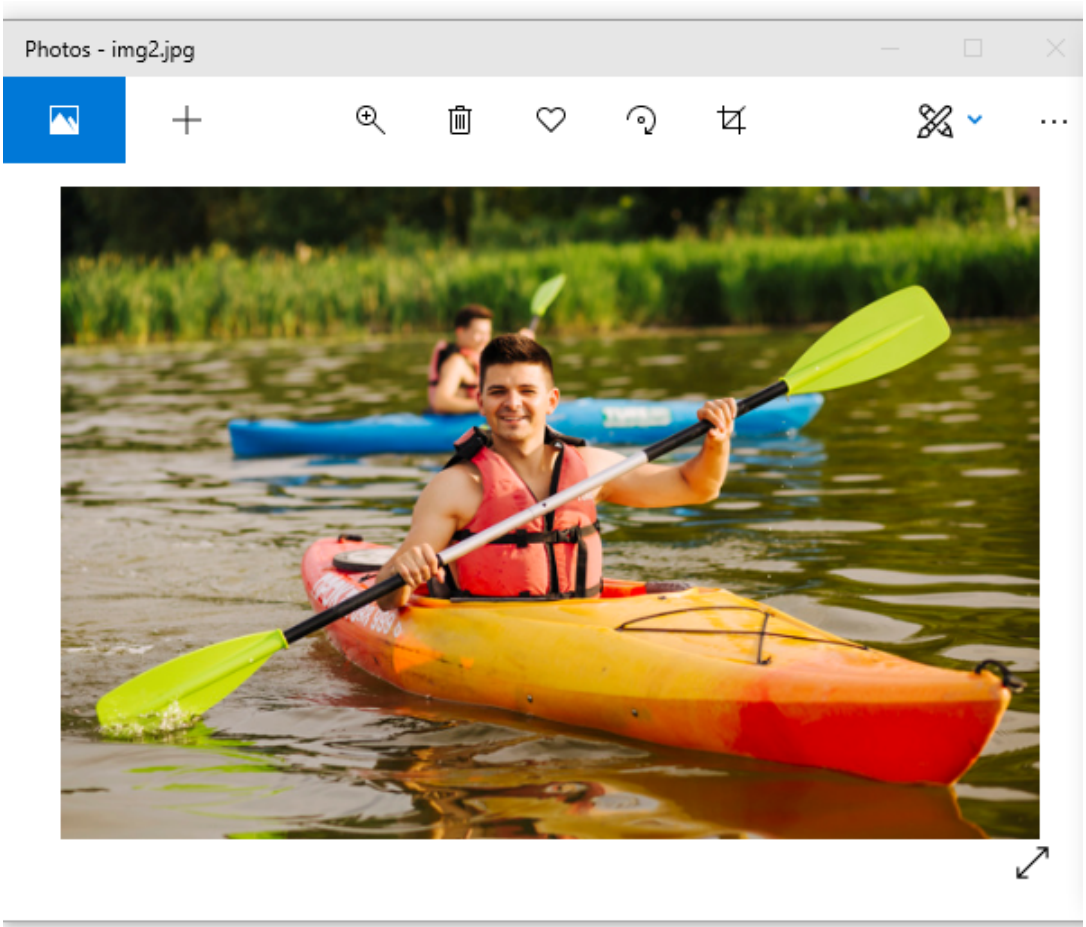
Introduction

1.1 Overview

Image caption generator is a task that involves computer vision and natural language processing concepts to recognize the context of an image and describe them in a natural language like English.In this we are given an image,we store the feature vector of image and using that feature vector and our model which basically consisting of CNN and LSTM.We trained our model to generate caption for other images.CNN is basically used for image classifications and identifying if an image is a bird, a plane or Superman, etc.

1.2 Motivation of the Research Work

The ability to generate image caption is very useful in many areas such as education and entertainment.Main problem in this area is of poor accuracy and very big dataset.But in this paper we have resolved this issue to some extent and working with CNN with LSTM for image caption generator results in better output.Previously also,many researchers have solved this problem using different kinds of models consisting of CNN and others RNN(recurrent neural network).




```
start man in yellow kayak is reflecting up river end
```

Chapter 2

Methodology and Implementation

2.1 Methodology

Our purposed method deals with implementing a caption generator using CNN and LSTM.In this method we are extracting feature vector using CNN and then using LSTM we are storing the results and using the feature vector and previous state vector we are predicting next word for our image. For this purpose we used pretrained model of Xception for extracting feature vector and then using LSTM we are creating our model.

Convolutional Neural networks(CNN) are specialized deep neural networks which can process the data that has input shape like a 2D matrix. Images are easily represented as a 2D matrix and CNN is very useful in working with images.

CNN is basically used for image classifications and identifying if an image is a bird, a plane or Superman, etc. for this purpose we pretrained our model and our discriminator and generator from the given dataset.

LSTM stands for Long short term memory, they are a type of RNN (recurrent neural network) which is well suited for sequence prediction problems. Based on the previous text, we can predict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short

term memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information. It is also called a CNN-RNN model. CNN is used for extracting features from the image. We will use the pre-trained model Xception.

LSTM will use the information from CNN to help generate a description of the image.

2.2 Implementation

1.Initially we used the following datasets

1.1 Flicker8k Dataset – Dataset folder which contains 8k images.

1.2 Flickr 8k text – Dataset folder which contains text files and captions of images.

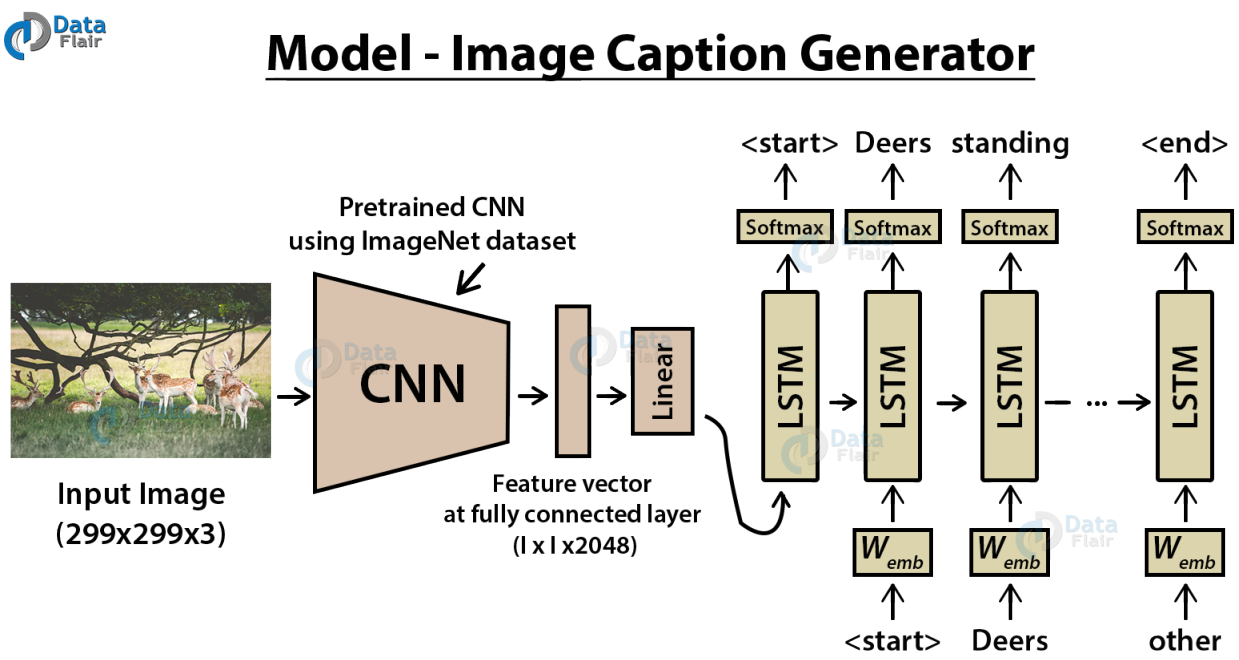
2.We performed preprocessing on the given datasets with given captions.

3.We created vocabulary of all unique words.Then we tokenized the vocabulary.

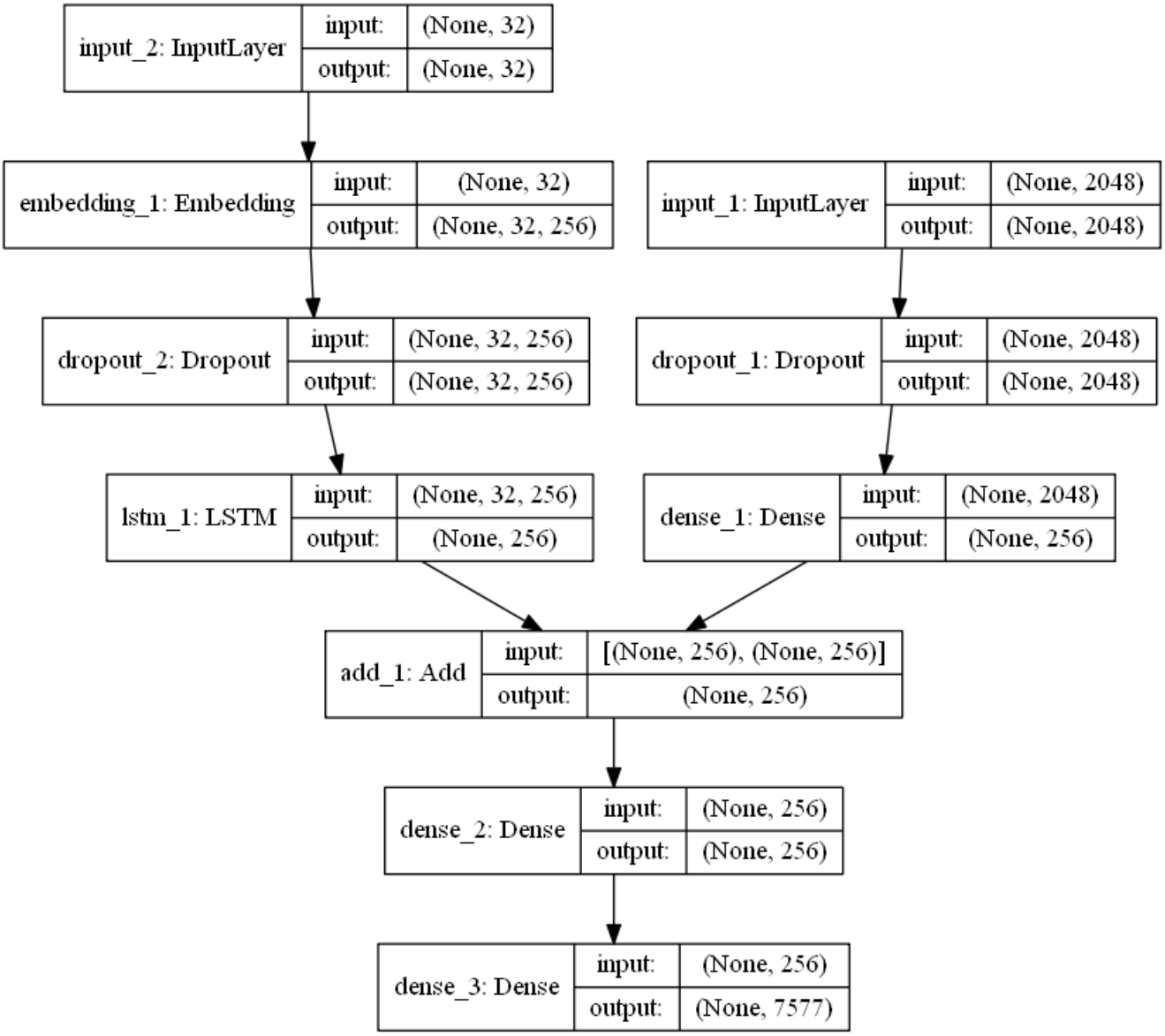
4.Then we created a data generator.This is basically converting all image feature vectors into normalised form taking their dimensions into factor.

5.Then we trained our model on 6k images and we set number of epochs to 5.

6. Finally for testing our model we used Bleu score which represents how close our sentence is to the ground truth.


Model - Image Caption Generator
Pretrained CNN
using ImageNet dataset
CNN
Linear
Input Image
(299x299x3)
Feature vector
at fully connected layer
(I x I x2048)
<start>
Deers
standing
<end>
Softmax
LSTM
W emb
<start>
Deers
other

whole Model


input_2: InputLayer
input: (None, 32)
output: (None, 32)
embedding_1: Embedding
input: (None, 32)
output: (None, 32, 256)
input_1: InputLayer
input: (None, 2048)
output: (None, 2048)
dropout_2: Dropout
input: (None, 32, 256)
output: (None, 32, 256)
dropout_1: Dropout
input: (None, 2048)
output: (None, 2048)
lstm_1: LSTM
input: (None, 32, 256)
output: (None, 256)
dense_1: Dense
input: (None, 2048)
output: (None, 256)
add_1: Add
input: [(None, 256), (None, 256)]
output: (None, 256)
dense_2: Dense
input: (None, 256)
output: (None, 256)
dense_3: Dense
input: (None, 256)
output: (None, 7577)

Chapter 3

Conclusions and Discussion

This report has made many conclusions ....

This report gives rise to a number of important....

- For image caption generator we used CNN with LSTM model,We compared bleu score with other methods and Results are as follows.

S.No.	Method Name	Bleu4	Our Bleu4
1	Mao et al. 2015	0.170	0.52
2	Jia et al. 2015	0.216	0.52
3	Xu et al. 2015	0.213	0.52
4	Wu et al. 2018	0.270	0.52

- we have achieved better performance than all above mentioned methods.
- There can be more work done in this field using CNN and any other Recurrent neural network.

Bibliography

[1] MD Zakir Hossain , Ferdous Ahmed Sohel,Mohd Fairuz *"A Comprehensive Survey of Deep Learning for Image Captioning"*, *inProc. IEEE ICDM,* 2015, pp. 479–488.

[2] Haoran Wang ,1 Yue Zhang,1 and Xiaosheng Yu, *An Overview of Image Caption Generation Methods* inProc. IEEE ICNP, 2016, pp. 1–10.

[3] J. Aneja, A. Deshpande, and S. Alexander,, *Convolutional image captioning,*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018.

[4] C. C. Park, B. Kim, and G. Kim, *Towards personalized image captioning via multimodal memory networks,* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, p. 1, 2018.

[5] X. Wang, S. Takaki, and J. Yamagishi,*An RNN-based quantized F0 model with multi-tier feedback links for text-to-speech synthesis,*,in Proceedings of the Interspeech 2017, pp. 1059–1063, Stockholm, Sweden, August 2017.