Caption Generation Using Python

Submitted in partial fulfillment of the requirements of

Mini Project (CSL804)

by

(Name of the Student)

(Roll No.:04,05,26,60)

Guide:

Faizan Ahmed Ansari

Sarfaraz Ahmed Ansari

Mehtab Alam Khan

Md. Tajuddin Shaikh

Computer Engineering Department

Rizvi College of Engineering

University of Mumbai

2019-2020

CERTIFICATE

This is to certify that the project entitled **“Caption Generation Using Python”** is a bonafide work of ‘‘**Faizan Ahmed Ansari, Sarfaraz Ahmed Ansari, Mehtab Alam Khan, Md. Tajuddin Shaikh” (Roll No.:04,05,26,60)** submitted to the University of Mumbai in partial fulfillment of **“MINI PROJECT(CSL804)”** in “**Computer Engineering”**.

Dr. Anupam Chaudhary
Project Guide

Prof. Shiburaj Pappu	**Dr. Varsha Shah**
Head of Department	Principal

Project Report Approval for B.E.

This Project report entitled ***Caption Generation Using Python*** by ***Faizan Ahmed Ansari, Sarfaraz Ahmed Ansari, Mehtab Alam Khan, Md. Tajuddin Shaikh*** is approved for MINI PROJECT(CSL 804) of Computer Engineering.

Examiners

1.--

2.--

Guide

1.--

2.--

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Faizan Ahmed Ansari 04

Sarfaraz Ahmed Ansari 05

Mehtab Alam Khan 26

Md. Tajuddin Shaikh 60

Date:

ABSTRACT

In the past few years, the problem of generating descriptive sentences automatically for images has garnered a rising interest in natural language processing and computer vision research. Image captioning is a fundamental task which requires semantic understanding of images and the ability of generating description sentences with proper and correct structure. In this study, the authors propose a hybrid system employing the use of multilayer Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. The convolutional neural network compares the target image to a large dataset of training images, then generates an accurate description using the trained captions. We have successfully finished the implementation of caption generation using python and are able to train our network on Google CoLab. In this paper, Python is used to form this caption generating platform with the help of TensorFlow library which can easily generate the LSTM model for a given images. In this research work, machines are trained by deep learning approach. To improve the efficiency of the caption generation, the training has to be quite deep with more sample images.

Keywords : automatically, Convolutional Neural Network, TensorFlow, deep learning, generation.

Index

List of Figures

List of Tables

Chapter 1

Introduction

Caption generation is an interesting artificial intelligence problem where a descriptive sentence is generated for a given image. It involves the dual techniques from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Recently, deep learning methods have achieved state-of the-art results on examples of this problem. It has been demonstrated that deep learning models are able to achieve optimum results in the field of caption generation problems. Instead of requiring complex data preparation or a pipeline of specifically designed models, a single end-to-end model can be defined to predict a caption, given a photo. These results show that our proposed model performs better than standard models regarding image captioning in performance evaluation.

Convolutional Neural networks 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.

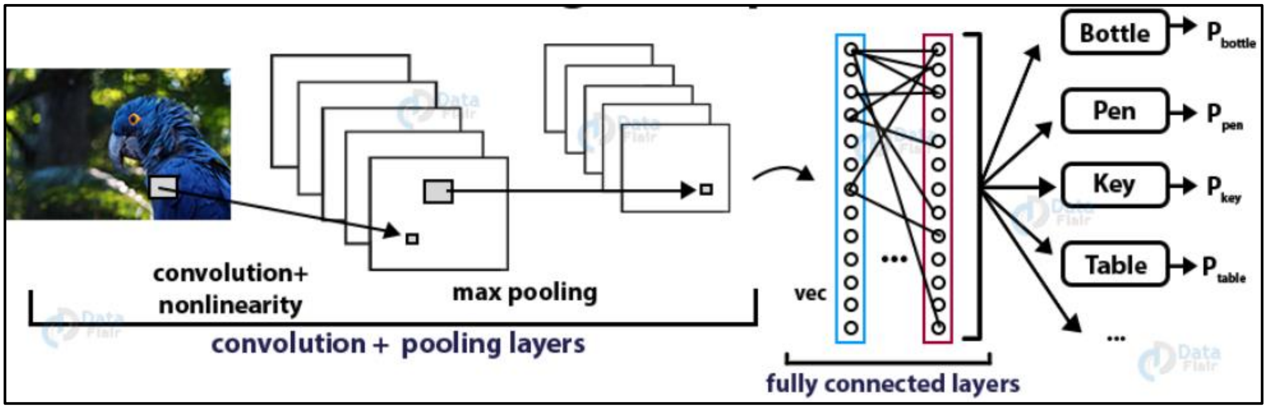


Figure 1. Working of Deep CNN

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.

To make our image caption generator model are merging these architectures. 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.

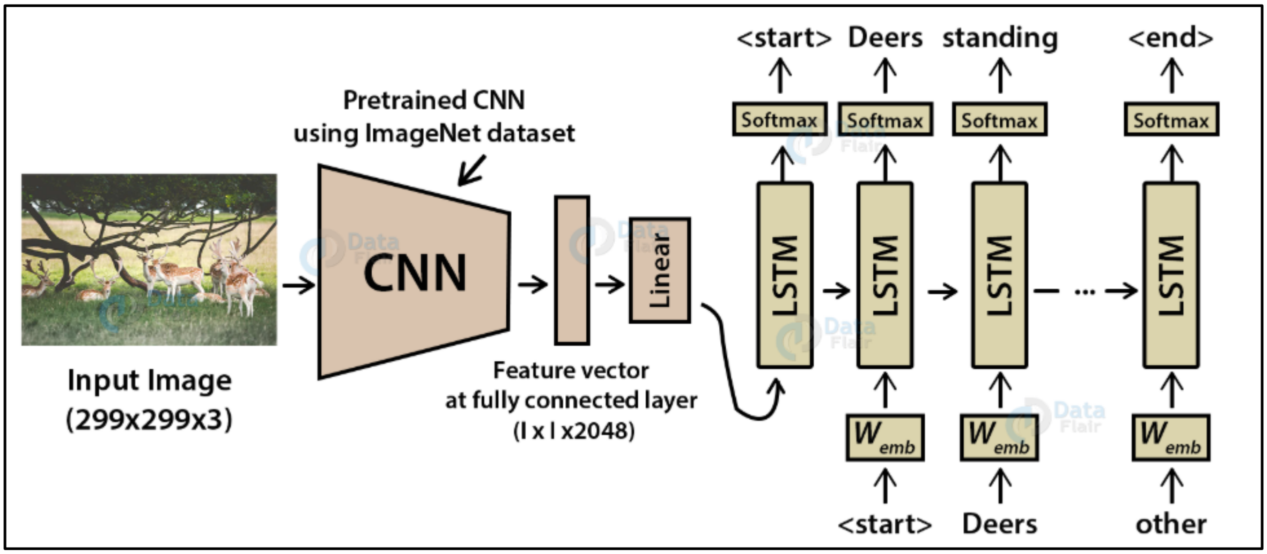


Figure 2. Model- Image Caption Generator

Chapter 2
Review of Literature

The image captioning problem and its proposed solutions have existed since the advent of the Internet and its widespread adoption as a medium to share images. Numerous algorithms and techniques have been put forward by researchers from different perspectives. Krizhevsky et al. [1] implemented a neural network using non-saturating neurons and a very efficient a unique method GPU implementation of the convolution function. By employing a regularization method called dropout, they succeeded in reducing overfitting. Their neural network consisted of maxpooling layers and a final 1000-way softmax.

Deng et al. [2] introduced a new database which they called ImageNet, an extensive collection of images built using the core of the WordNet structure. ImageNet organized the different classes of images in a densely populated semantic hierarchy. Karpathy and FeiFei [3] made use of datasets of images and their sentence descriptions to learn about the inner correspondences visual data and language. Their work described a Multimodal Recurrent Neural Network architecture that utilises the inferred co-linear arrangement of features in order to learn how to generate novel descriptions of images.

Yang et al. [4] proposed a system for the automatic generation of a natural language description of an image, which will help immensely in furthering image understanding. The proposed multimodel neural network method, consisting of object detection and localization modules, is very similar to the human visual system which is able to learns how to describe the content of images automatically. In order to address the problem of LSTM units being complex and inherently sequential across time, Aneja et al. [5] proposed a convolutional network model for machine translation and conditional image generation. Pan et. al [6] experimented extensively with multiple network architectures on large datasets consisting of varying content styles, and proposed a unique model showing noteworthy improvement on captioning accuracy over the previously proposed models.

Vinyals et al. [7] presented a generative model consisting of a deep recurrent architecture that leverages machine translation and computer vision, used to generate natural descriptions of an image by ensuring highest probability of the generated sentence to accurately describe the target image. Xu et al. [8] introduced an attention based model that learned to describe the image regions automatically. The model was trained using standard backpropagation techniques by maximizing a variable lower bound. The model was able to automatically learn identify object boundaries while at the same time generate an accurate descriptive sentence.

Chapter 3

Theory, Methodology and Algorithm

3.1. Theory

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.

The objective of our project is to learn the concepts of a CNN and LSTM model and build a working model of Image caption generator by implementing CNN with LSTM.

3.2. Methodology

We have written data preprocessing scripts to process raw input data (both images and captions) into proper format; A pre-trained Convolutional Neural Network architecture as an encoder to extract and encode image features into a higher dimensional vector space; An LSTM-based Recurrent Neural Network as a decoder to convert encoded features to natural language descriptions; Attention mechanism which allows the decoder to see features from a specifically highlighted region of the input image to improve the overall performance.

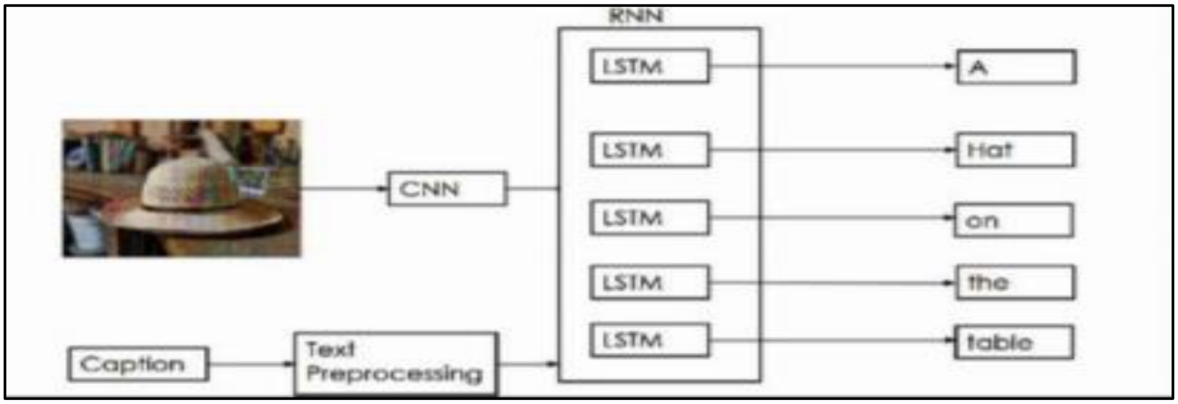


Figure 3. Architecture

The model consists of 3 phases:

A. Image Feature Extraction

The features of the images from the Flickr 8K dataset is extracted using the VGG 16 model due to the performance of the model in object identification. The VGG is a convolutional neural network which consists of consists of 16 layer which has a pattern of 2 convolution layers followed by 1 dropout layers until the fully connected layer at the end. The dropout layers are present to reduce overfitting the training dataset, as this model configuration learns very fast. These are processed by a Dense layer to produce a 4096 vector element representation of the photo and passed on to the LSTM layer.

B. Sequence processor

The function of a sequence processor is for handling the text input by acting as a word embedding layer. The embedded layer consists of rules to extract the required features of the text and consists of a mask to ignore padded values. The network is then connected to a LSTM for the final phase of the image captioning.

C. Decoder

The final phase of the model combines the input from the Image extractor phase and the sequence processor phase using an additional operation then fed to a 256 neuron layer and then to a final output Dense layer that produces a softmax prediction of the next word in the caption over the entire vocabulary which was formed from the text data that was processed in the sequence processor phase. The structure of the network to understand the flow of images and text is shown in the Figure 3.

3.3. Algorithm

Step 1: The code ensures that Google CoLab is running the correct version of TensorFlow. Running the following code will map your GDrive to /content/drive.

Step 2: Importing the needed libraries.

Step 2: Needed data for download the source and upload the drive in glove.6B, Flicker8k_Dataset and Flicker8k_Text.

Step 3: Running Locally in connected to CoLab and Gdrive.

Step 4: Dataset From Flicker8k in clean/Build.

Step 5: Read all image names and use the predefined train/test sets.

Step 6: Build the sequences. We include a start and stop token at the beginning/end.

S**tep 7:** The summary for the Chosen Neural Network to Transferred in displayed.

Step 8: Separate the captions that will be used for training. There are two sides to this training, the images and the captions.

Step 9: Using a Data Generator- The generator will create new data, as it is needed. The generator provided here creates the training data for the caption neural network, as it is needed.

Step 10: Loading the Glove Embeddings.

Step 11: Building the Neural Network and Train the Neural Network.

Step 12: Evaluate Performance on Test Data from Flicker8k.

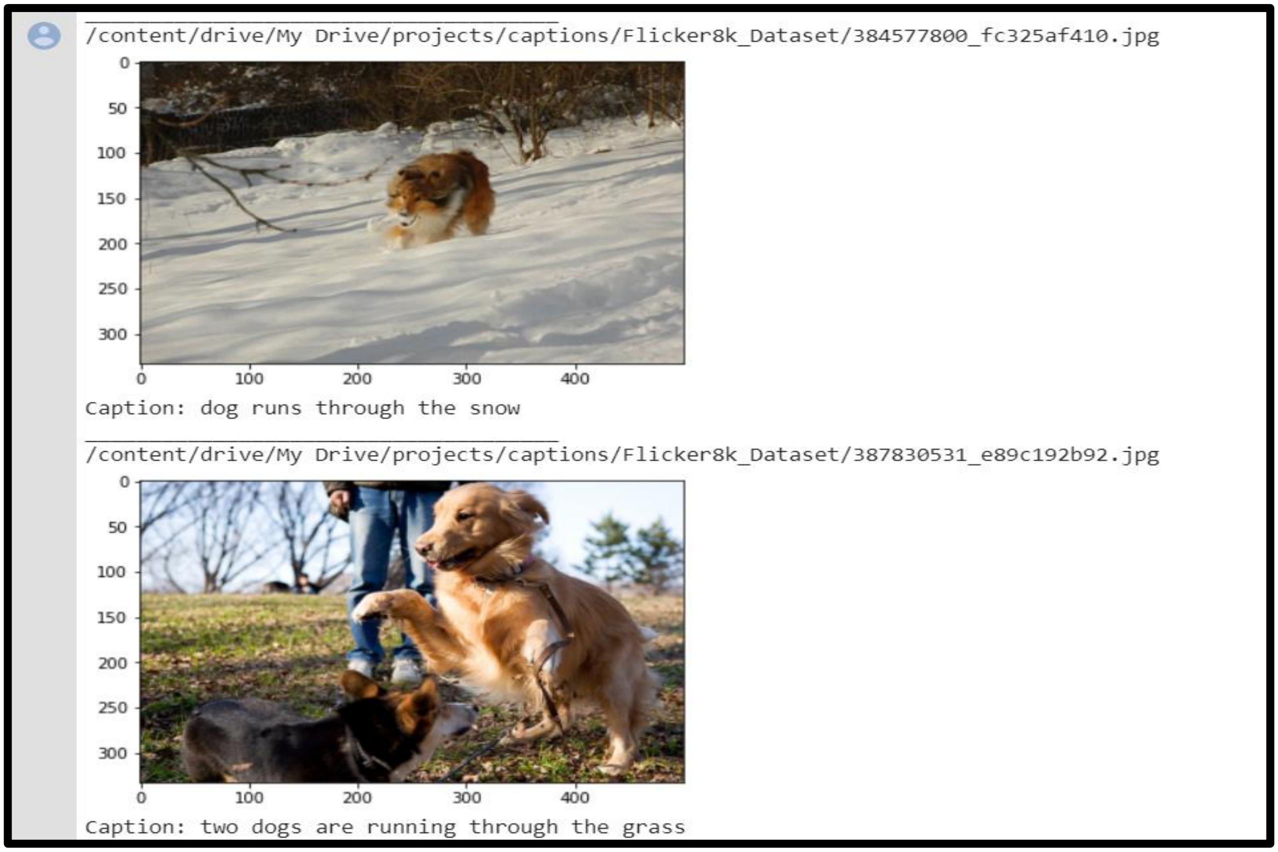
or

Evaluate Performance on My Own Photos.

Chapter 4

Results and Discussions

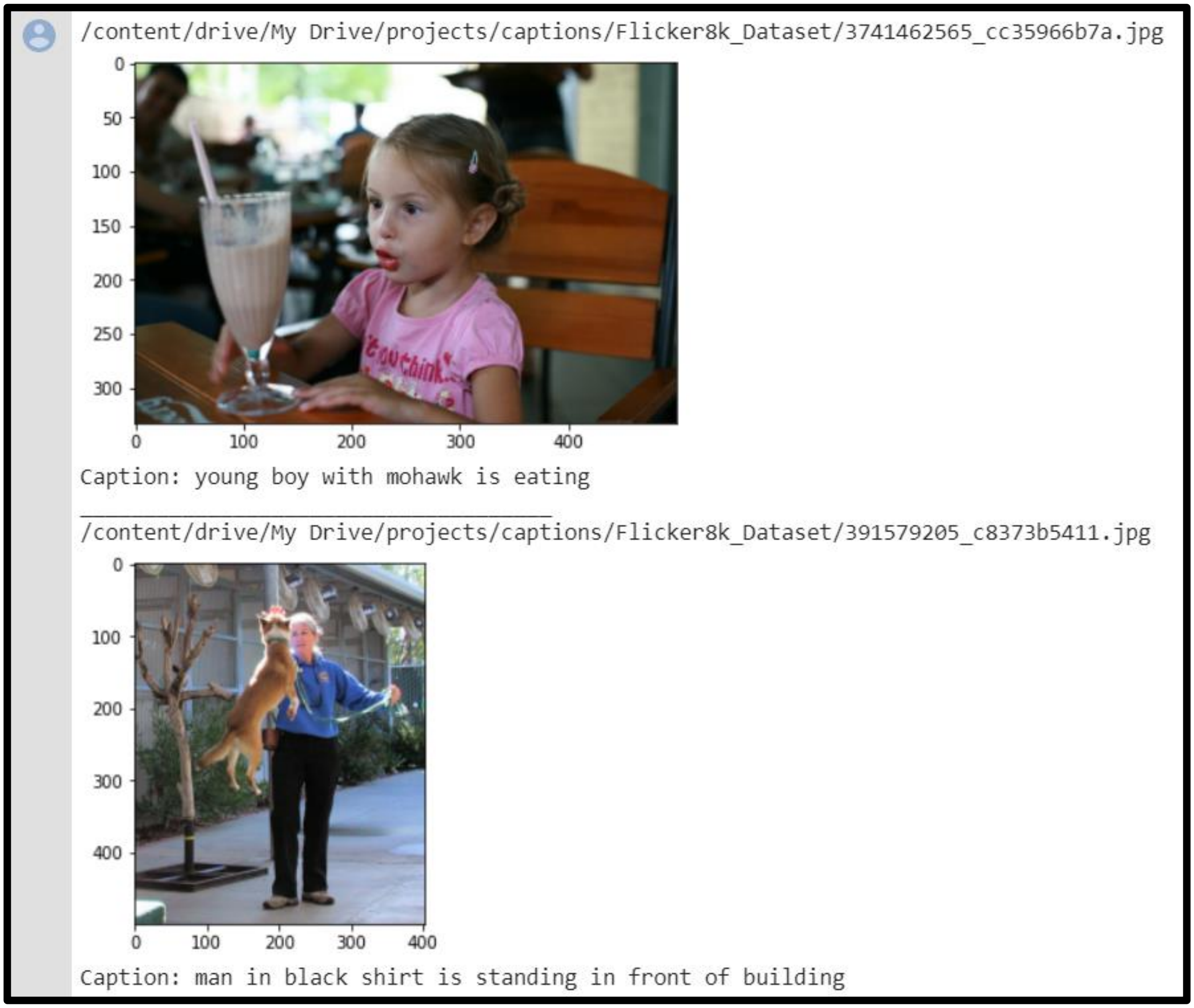
The image captioning model was implemented and we were able to generate moderately comparable captions with compared to human generated captions. The model converts the image into word vector. This word vector is provided as input to LSTM cells which will then form sentence from this word vector. The example of Generated sentence are dog runs through the snow, while actual human generated sentences are dog runs through the grass, dog runs through the ocean, dog runs through the ball. The Caption Generation using python in google CoLab are evaluate performance on Test Data from Flicker8k in choosed 10 image to result in Generated sentence.



```
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/384577800_fc325af410.jpg
Caption: dog runs through the snow
_____________________________________________
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/387830531_e89c192b92.jpg
Caption: two dogs are running through the grass
```

Fig. 4. Output

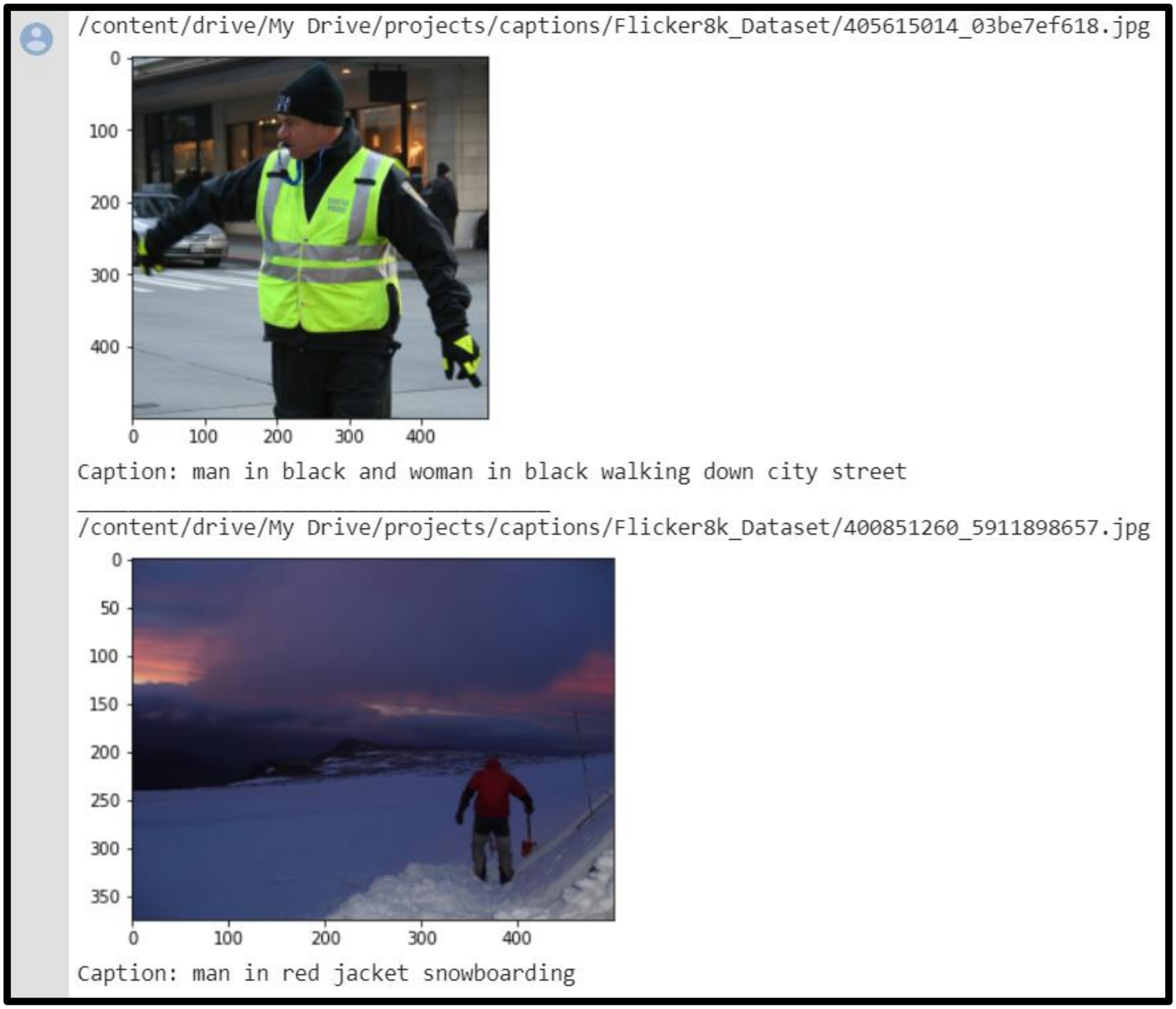
1. Generated sentence are **“dog runs through the snow”**.
2. Generated sentence are **“dog runs through the grass”**.



```
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/3741462565_cc35966b7a.jpg
Caption: young boy with mohawk is eating
_____________________________________
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/391579205_c8373b5411.jpg
Caption: man in black shirt is standing in front of building
```

Fig. 5. Output

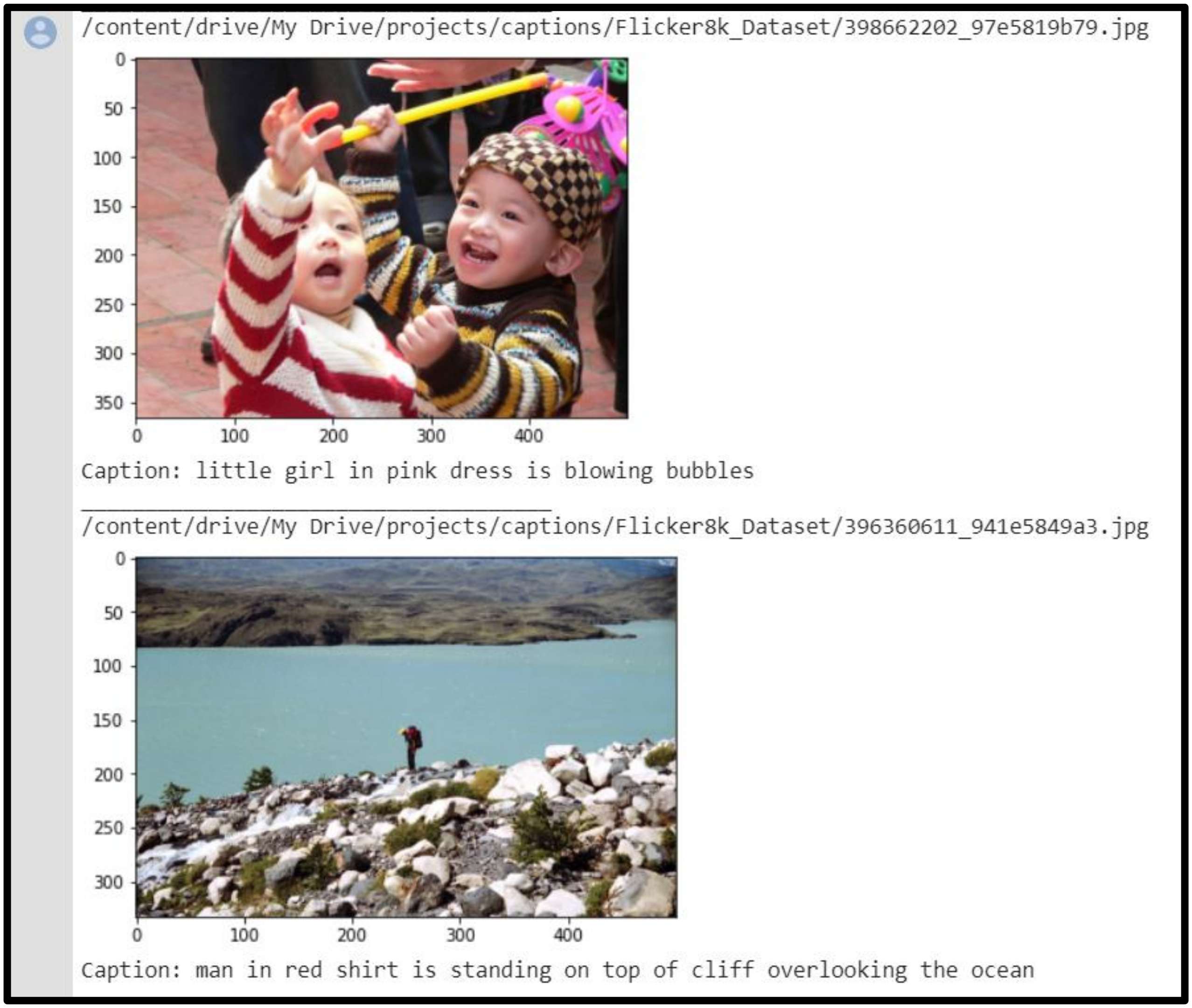
3. Generated sentence are **“young boy with mohawk is eating”**.
4. Generated sentence are **“man in black shirt is standing in front of building”**.



```
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/405615014_03be7ef618.jpg
Caption: man in black and woman in black walking down city street
________________________________________
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/400851260_5911898657.jpg
Caption: man in red jacket snowboarding
```

Fig. 6. Output

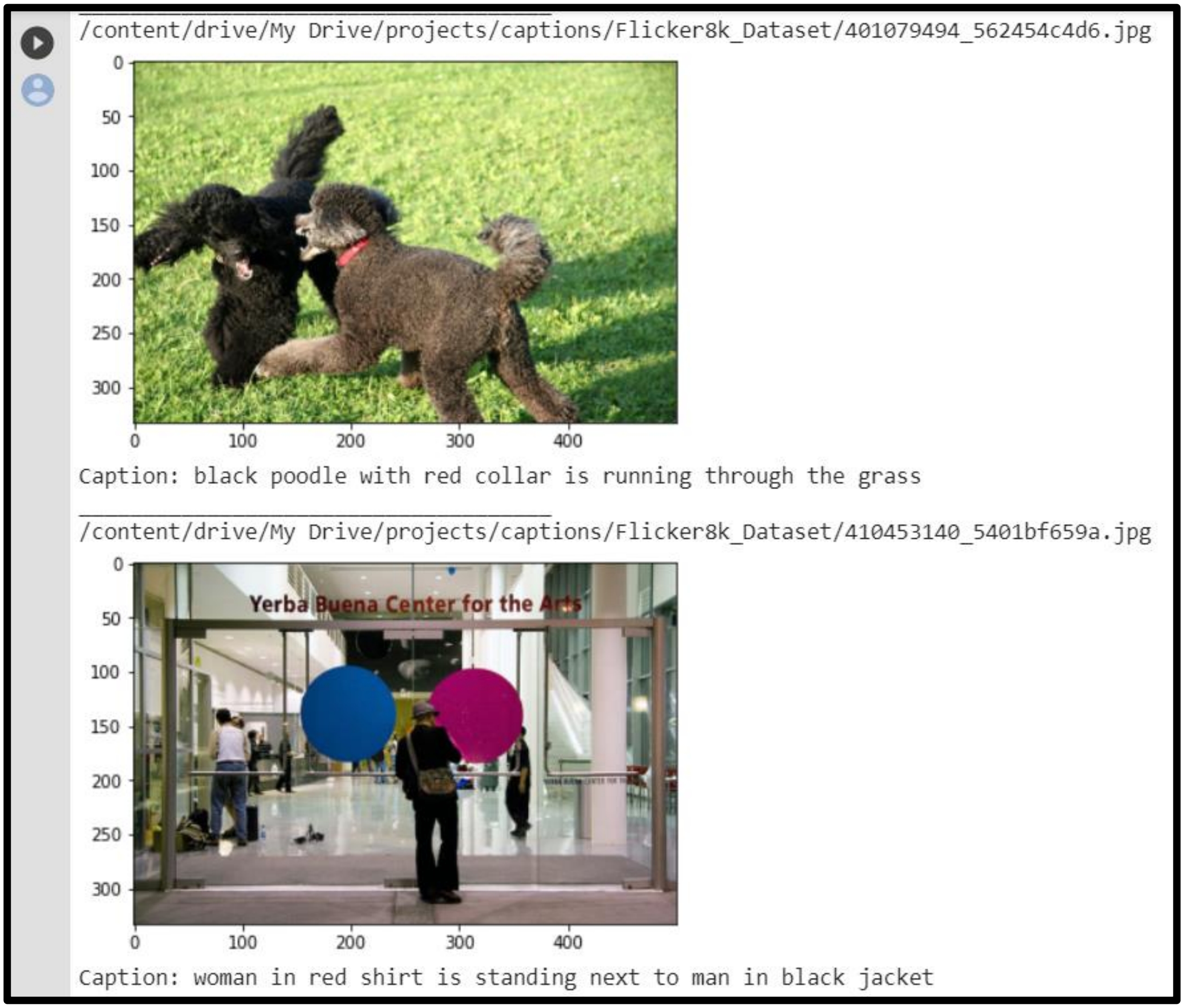
5. Generated sentence are **“man in black and woman in black walking down city street”**.
6. Generated sentence are **“man in red jacket snowboarding”**.



```
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/398662202_97e5819b79.jpg
Caption: little girl in pink dress is blowing bubbles
_____________________________________
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/396360611_941e5849a3.jpg
Caption: man in red shirt is standing on top of cliff overlooking the ocean
```

Fig. 7. Output

7. Generated sentence are **“little girl in pink dress is blowing bubbles”**.
8. Generated sentence are **“man in red shirt is standing on top of cliff overlooking the ocean”**.



```
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/401079494_562454c4d6.jpg
```

```
Caption: black poodle with red collar is running through the grass
```

```
/content/drive/My Drive/projects/captions/Flicker8k_Dataset/410453140_5401bf659a.jpg
```

```
Caption: woman in red shirt is standing next to man in black jacket
```

Fig. 8. Output

9. Generated sentence are **“black poodle with red collar is running through the grass”**.
10. Generated sentence are **“woman in red shirt is standing next to man in black jacket”**.

Chapter 5

Conclusions

The project will help in Caption Generation is a task that involves computer vision and natural language processing. This project requires good knowledge of Deep learning, Python, working on Google CoLab notebook. The image captioning in new technology that combines LSTM that text generation with the computer visions power of a convolutional neural network. For the project to work properly the developer should go through the basic knowledge of python language and data manipulation. It is also important to understand that this model doesn't have a human-like understanding of what the images contain. If it sees an image of a giraffe and correctly produces a caption stating that, it doesn't mean that the model has a deep understanding of what a giraffe is; the model doesn't know that it's a tall animal that lives in Africa and Zoos. We have successfully finished the implementation of caption generation using python and are able to train our network on Google CoLab.

Chapter 6
References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, [Online] Available: https://papers.nips.cc/paper/4824-imagenetclassificationwith-deep-convolutionalneural-networks.pdf

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database

[3] Andrej Karpathy, Li Fei-Fei, Deep VisualSemantic Alignments for Generating Image Descriptions, [Online] Available: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

[4] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, [Online] Available: https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.p df

[5] Jyoti Aneja, Aditya Deshpande, Alexander Schwing, Convolutional Image Captioning, [Online] Available: https://arxiv.org/pdf/1711.09151.pdf

[6] Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, Automatic Image Captioning, Conference: Conference: Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on, Volume: 3

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and Tell: A Neural Image Caption Generator, [Online] Available: https://arxiv.org/pdf/1411.4555.pdf

[8] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, [Online] Available: https://arxiv.org/pdf/1502.03044.pdf

[9] M. Hodosh, P. Young and J. Hockenmaier (2013) ”Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics”, Journal of Artificial Intelligence Research, Volume 47, pages 853-899

Acknowledgements

I am profoundly grateful to Dr. Anupam Chaudhary for his expert guidance and continuous encouragement throughout to see that this project rights its target.

I would like to express deepest appreciation towards Dr. Varsha Shah, Principal RCOE, Mumbai and Prof. Shiburaj Pappu HOD Computer Department whose invaluable guidance supported me in this project.

At last I must express my sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped us directly or indirectly during this course of work.

FAIZAN AHMED ANSARI 04
SARFARAZ AHMED ANSARI 05
MEHTAB ALAM KHAN 26
MD. TAJUDDIN SHAIKH 60