IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Experiment based on Deep Learning: Image Caption Generator

Manisha M. Patil
Asst. Professor,
Indira College of Commerce and Science, Pune.

Abstract *– The purpose of the Image caption Generator is to create a sentence description for given image. Our project model will take a picture as input and generate an English sentence as output, describing the contents of the image. It has attracted much research attention in cognitive computing in recent years. The task is quite complex because the concepts of both computer vision and tongue processing domains are combined. In this research study, we proposed a model based on CNN (Convolution neural network) and LSTM (Long- and Short-Term Memory) algorithms to describe image in the sentence form. The CNN works as an encoder to extract features from images and LSTM works as a decoder to generates words describing the image. After the caption generation phase, we use BLEU Scores to gauge the efficiency of our model. Thus, our system helps the user to urge descriptive caption for the given input image.*

Key Words – Convolutional Neural Network, Long Short-Term Memory, Natural Language Processing, Computer Vision.

1. INTRODUCTION

Problem Statement- To develop a system for users, which may automatically generate the outline of a picture with the utilization of CNN alongside LSTM.

Automatically describing the content of images using tongue may be a fundamental and challenging task. Now a days we have powerful tools and huge datasets are available, which helps to build models to generate captions for an image. On the opposite hand, humans are ready to easily describe the environments they are in. Given an image, it is natural for an individual to elucidate an immense amount of details about this image with a glance. Although great development has been made in computer vision, tasks like recognizing an object, action classification, image classification, attribute classification, and scene recognition are possible but it is a new type of task, where computer recognize and describe the image in the form of sentence like human.

For this goal of image captioning, based on the semantics of images should be captured here and expressed in the desired form of natural languages. In reality one can use this solution for visually impaired people, to make them understand images through the description of images, those are available on the internet.

So, to form our image caption generator model, we'll be merging CNN-RNN architectures. Feature extraction from images is done using CNN. We have used the pre-trained model Exception. The information received from CNN is then used by LSTM for generating a description of the image.

However, sentences that are generated using these approaches are usually generic descriptions of the visual content, and background information is ignored. Such generic descriptions do not satisfy emergent situations as they, essentially replicate the information present in the images, and detailed descriptions regarding events and entities present in the images aren't provided, which is imperative to understanding emergent situations.

The objective of our project is to develop a web-based interface for users to get the description of theimage and to make a classification system to differentiate images as per their description. It also can make the task of SEO easier which is complicated as they need to take care of and explore enormous amounts of knowledge.

2. SYSTEM ARCHITECTURE

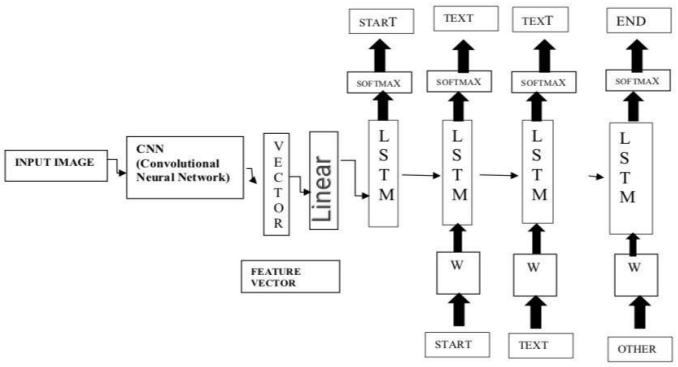


Figure 1: Proposed Model of Image Caption Generator

The proposed model of Image Caption Generator is as shown within the above figure 1. Here, in this model, an input image is given and then A convolutional neural network is used to create a dense vector, also called an embedding, this vector can be used as input into other algorithms, and it generates a suitable caption for the given image as output.

For an image caption generator, this embedding becomes a representation of the image and is used asthe initial state of the LSTM for generating meaningful captions, for the image.

The System Architecture of our system is shown below in Figure 2.

This is what our proposed system architecture will look like.

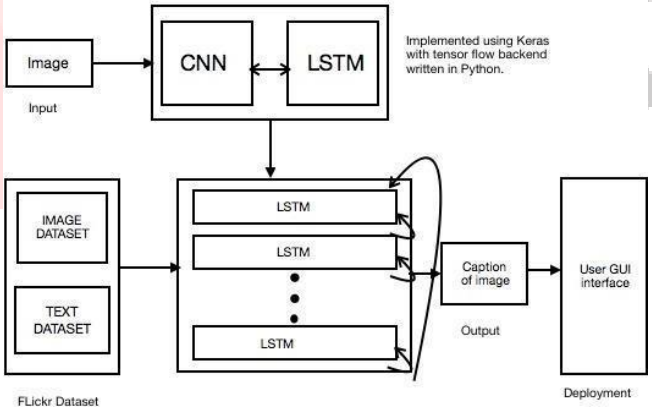


Figure 2: System Architecture of Image Caption Generator

System Requirements:

- OS: Windows 7 and above, Recommended: Windows 10.
- CPU: Intel processor with 64-bit support.
- Disk Storage: 8GB of free disk space.

For Execution: Spyder using Anaconda Framework in Python.

2.1 : Algorithms

2.1.1 : Convolutional Neural Network

Convolutional Neural networks are specialized deep neural networks that process the data that has input shape like a 2D matrix. CNN works well with images and is easily represented as a 2D matrix. Image classification and identification are often easily done using CNN. It can determine whether an image is a bird, a plane or Superman, etc.

An important feature of an image can be extracted by scanning the image from left to right and top to bottom and finally, the features are combined to classify images. It can affect the pictures that are translated, rotated, scaled, and changes in perspective.

2.1.2 : Long Short-Term Memory

LSTM is a type of RNN (Recurrent Neural Network) that is well suited for sequence prediction problems. One can predict the next words based on the previous text used. This shows how effectively limitations of RNN are overcomes. LSTM can carry out relevant information throughout the processing, it discards non-relevant information.

2.2 : Data Exploration

For the image caption generator, we have used the Flickr8K_dataset. There are also other big datasets like Flickr_30K and MSCOCO dataset but it can take weeks for systems having only CPU support just to train the network, so we used a small Flickr8K_dataset. Using a huge dataset helps in developing a far better model.

3. PROPOSED IMAGE CAPTION GENERATOR

Here we have shown the DFD's (Data Flow Diagrams) of our system. DFD's provide us the basic overview of the whole Image Caption Generator System or process being analyzed or modeled.

Figure 3: Data Flow Diagram Level 0

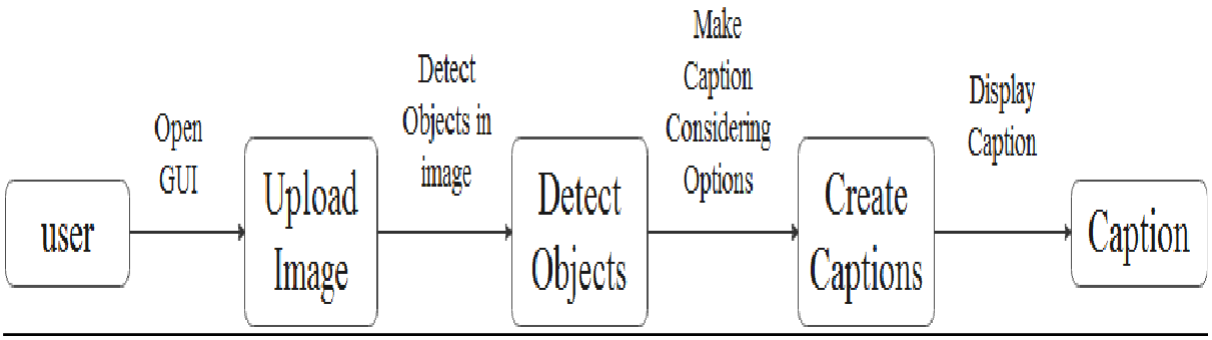


Figure 4: Data Flow Diagram Level 1

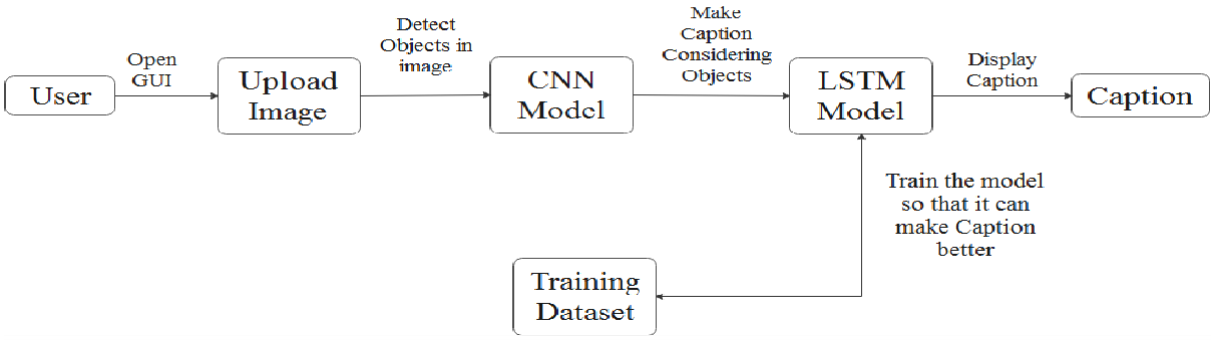


Figure 5: Data Flow Diagram Level 2

Figure 6. shows the State Chart diagram of the system. The first user will browse the site. Then he will upload the image, CNN will identify the objects present in the image then LSTM will start preparing captions considering the objects present in the image using Training Dataset, which comprises of Image Data Set and Text Data Set, after the training an appropriate caption is going to be generated and displaying top the user.

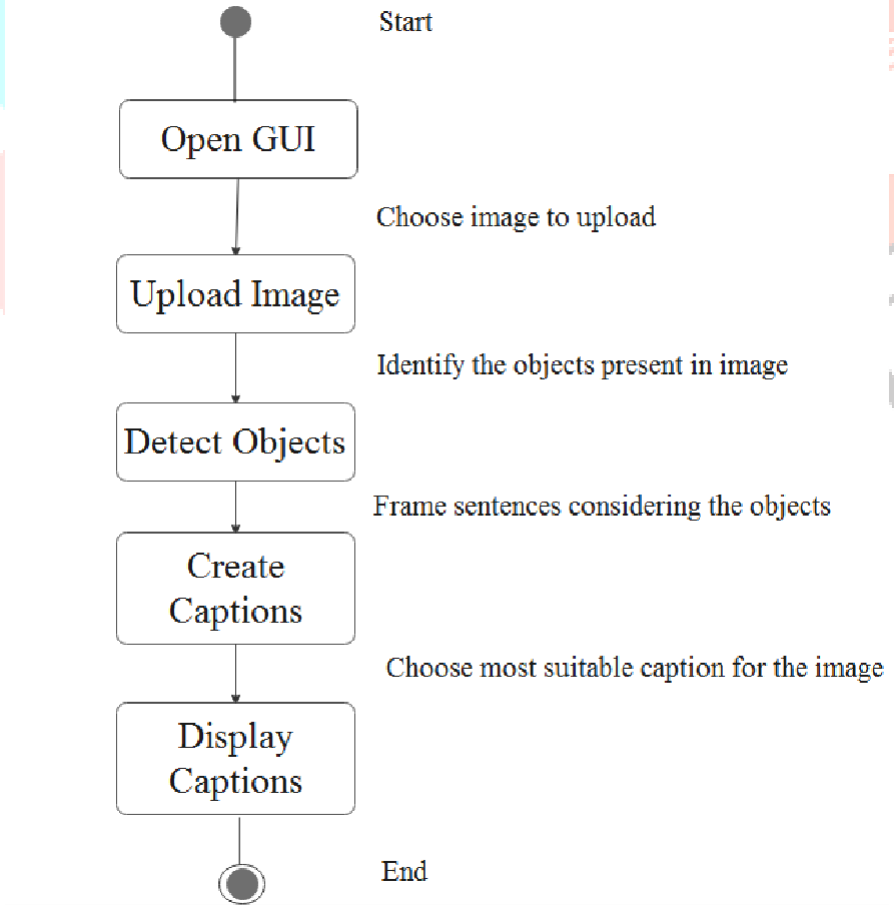


Figure 6: State Chart Diagram of Steps taken by the System.

The proposed system of Image Caption Generator has the capabilities to Generate Captions for the Images, provided during the Training purpose and also for the New images as well. Our model takes a picture as Input and by analyzing the image it detects objects present in a picture and makes a caption that describes the image tolerably for any machine to know what a picture is trying to mention.

4. IMPLEMENTATION OF THE SYSTEM

Here we will discuss the implementation of the system.

4.1 : Object Detection

Objects are detected from the image with the assistance of the CNN Encoder.

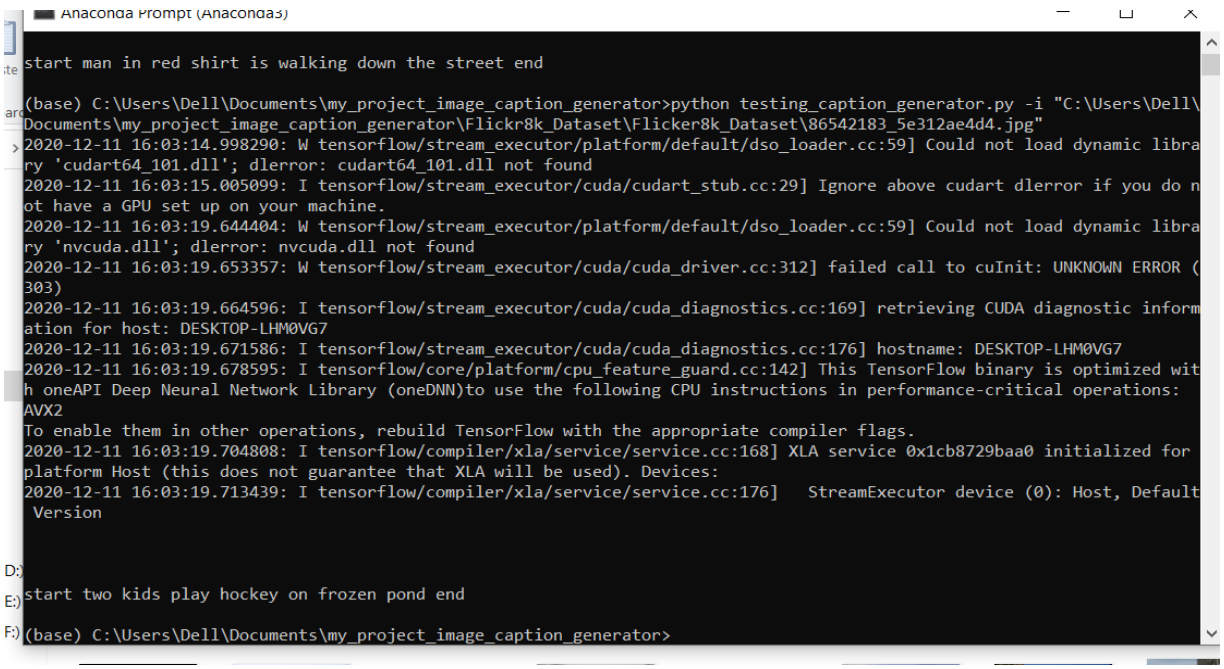
4.2 : Sentence Generation

By Using LSTM, sentences are generated. Each predicted word is employed to get subsequent words. Using these words, the appropriate sentence is formed with the help of an Optimal beam search. Here, the Softmax function is going to be used for the prediction of the word.

4.3 : Deployment

The final project is going to be deployed using Tkinter which is a Python-based GUI. It is the standard Python Interface for developing GUI's.

5. RESULTS



Generated Caption: Two kids play hockey on a frozen pond.

6. CONCLUSION

In this advanced Python project, an image caption generated has been developed using a CNN-RNN model. Some key aspects about our project to note are that our model depends on the data, so, it cannot predict the words that are out of its vocabulary. A dataset consisting of 8000 images is used here. But for production-level models i.e., higher accuracy models, we need to train the model on larger than 100,000 images datasets so that better accuracy models can be developed.

7. REFERENCES

[1] S. ALBAWI and T. A. MOHAMMED, "Understanding of a Convolutional Neural Network," in ICET, Antalya, 2017.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "A Neural Image Caption Generator," CVPR 2015 Open Access Repository, vol. Xiv, 17 November 2014.

[3] S. Hochreiter, "LONG SHORT-TERM MEMORY," Neural Computation, December 1997.

[4] J. Chen, W. Dong and M. Li, "Image Caption Generator Based On Deep Neural Networks," March 2018.

[5] "DataFlair site"