International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Image Caption Generator using Big Data and Machine Learning

Dr. Vinayak D. Shinde[1], Mahiman P. Dave[2], Anuj M. Singh[3], Amit C. Dubey[4]

[1]Head of Department of Computer Engineering, Shree L.R. Tiwari College of Engineering, Maharashtra, India
[2,3,4]B.E. Student, Computer Engineering, Shree L.R. Tiwari College of Engineering, Maharashtra, India

---

Abstract – *Image captioning aims to automatically generate a sentence description for an image. Our project model will take an image 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 the recent years. The task is rather complex, as the concepts of both computer vision and natural language processing domains are combined together. We have developed a model using the concepts of a Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) model and build a working model of Image caption generator by implementing CNN with LSTM. The CNN works as encoder to extract features from images and LSTM works as decoder to generates words describing image. After the caption generation phase, we use BLEU Scores to evaluate the efficiency of our model. Thus, our system helps the user to get descriptive caption for the given input image.*

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

1. INTRODUCTION

Problem Statement: To develop a system for users, which can automatically generate the description of an image with the use of CNN along with LSTM.

Automatically describing the content of images using natural language is a fundamental and challenging task. With the advancement in computing power along with the availability of huge datasets, building models that can generate captions [13] for an image has become possible. On the other hand, humans are able to easily describe the environments they are in. Given a picture, it's natural for a person to explain an immense amount of details about this image with a fast glance. Although great development has been made in computer vision, tasks such as recognizing an object, action classification, image classification, attribute classification and scene recognition are possible but it is a relatively new task to let a computer describe an image that is forwarded to it in the form of a human-like sentence.

For this goal of image captioning, based on semantics of images should be captured here and expressed in the desired form of natural languages. It has a great impact in the real world, for instance by helping visually impaired people better understand the content of images on the web.

So, to make our image caption generator model, we will 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 [16] 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 in emergent situations as they, essentially replicate the information present in the images and detailed descriptions [14] regarding events and entities present in the images are not 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 [9] description of the image and to make a classification system in order to differentiate images as per their description. It can also make the task of SEO easier which is complicated as they have to maintain and explore enormous amounts of data.

2. LITERATURE REVIEW

In Literature review, various references of the existing projects are taken into consideration which are similar to this current project.

1. [1] In this paper one of the most popular deep neural networks is the Convolutional Neural Network (CNN) is explained. There are multiple layersin CNN; such as convolutional layer, & non-linearity layer, & pooling layer and fully- connected layer as well. The CNN has an excellent performance in machine learning problems and one of the most common algorithms.

2. [2] In this paper Sepp Hochreiter explain about the deep neural network algorithm long short team Memory (LSTM). LSTM is local in both space as well as in time; the computational complexity is per time of step and also the weight pattern representation. In comparison to other algorithm LSTM leads to many more successful runs, and learn much faster. It's even solve complex, artificial long time lag tasks that have never been solved by previous recurrent network

3. [3] The fundamental problem in artificial intelligence that connects computer vision and Natural language processing is automatically

describing the content of an image. In this paper, A.L systematically analyze a deep neural networks based image caption generation method. Here an image is given as the input, and the method as output in the form of sentence in English describing the content of the image. They analyze three components of the method: convolutional neural network (CNN), recurrent neural network (RNN) and sentence generation. This model analyze image and generate more trival and relevant words for images.

4. [4] Current image captioning approaches generate descriptions which lack specific information, such as named entities that are involved in the images. Here Di Lu, Spencer Whitehead had proposed a very new task which generates descriptive image captions, given images as input. A simple solution to this problem that we are proposing is that we will train a CNN-LSTM model so that it can generate a caption based on the image.

5. [5] Automatically describing the content of an image using properly arranged English sentences is a tough challenging task, but it could is something very necessary for helping visually impaired people. Modern smartphones are able to take the photographs, which can help in taking surrounding images for visually impaired peoples. Here images as input can generate captions that can be loud enough so that visually impaired can hear, so that they can get a better sense of things present in there surrounding. Here Christoper Elamri uses a CNN model to extract features of an image. These features are then fed into a RNN or a LSTM model to generate a description of the image in grammatically correct English sentences describing the surroundings.

3. SYSTEM ARCHITECTURE

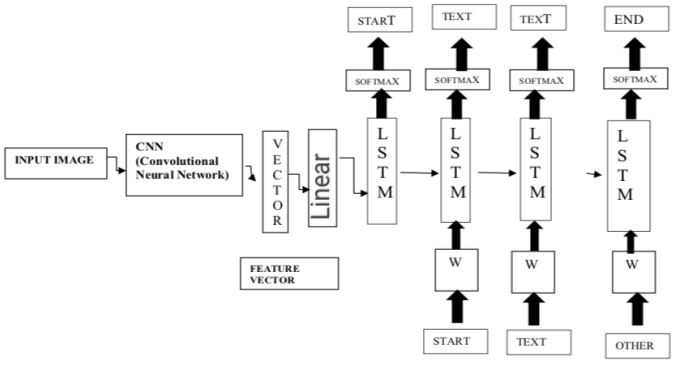


Figure 1: Proposed Model of Image Caption Generator

The proposed model of Image Caption Generator is as shown in the above figure 1. Here in this model, input image is given & then A convolutional neural network is used to create a dense feature vector as shown in figure. This dense vector, also called an embedding, this vector can be used as input into other algorithms, and its generates [11] suitable caption for given image as output.

For an image caption generaor, this embedding becomes a representation of the image and used as the initial state of the LSTM for generating meaningfull captions, for the image.

Sytem Architecture of our system is shown below in Figure2.

This is our proposed system architecture will look like .

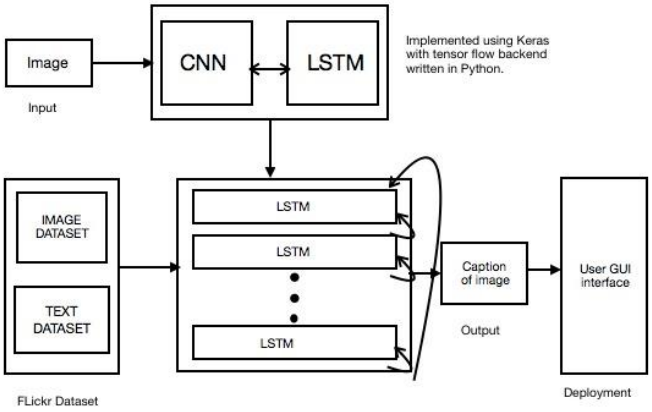


Figure 2: System Architecture of Image Caption Generator

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.
For Deployment: Python GUI - Tkinter

3.1. Algorithms

3.1.1. Convolutional Neural Network

Convolutional Neural networks [10] are specialized deep neural networks which processes the data that has input shape like a 2D matrix. CNN works well with images and are easily represented as a 2D matrix. Image classification and identification can be easily done using CNN. It can determine whether an image is a bird, a plane or Superman, etc. [7]

Important features of an image can be extracted by scanning the image from left to right and top to bottom

and finally the features are combined together to classify images. It can deal with the images that have been translated, rotated, scaled and changes in perspective.

3.1.2. Long Short Term Memory

LSTM are type of RNN (recurrent neural network) [2] which is well suited for sequence prediction problems. We can predict what the next words will be based on the previous text. It has shown itself effective from the traditional RNN by overcoming the limitations of RNN. LSTM can carry out relevant information throughout the processing, it discards non-relevant information [16].

3.2. Data Exploration

For the image caption generator, we have used the Flickr_8K dataset. There are also other [6] 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 better model.

4. PROPOSED IMAGE CAPTION GENERATOR

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

Figure 3: DFD Diagram Level 0

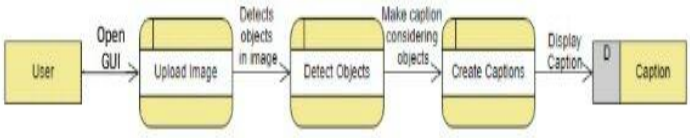


Figure 4: DFD Diagram Level 1

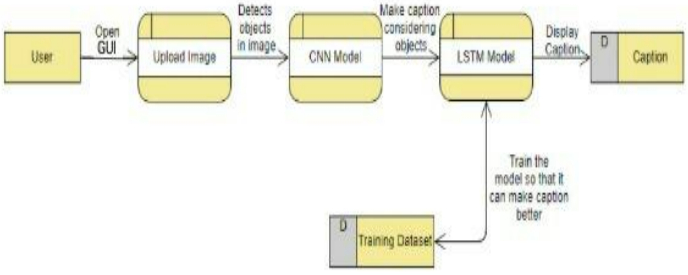


Figure 5: DFD Diagram Level 2

The Figure 6. shows State Chart Diagram of the system. 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 [8] Training Dataset, which comprises of Image Data set and Text Data Set, after the training a suitable caption will be generated and displayed top the user.

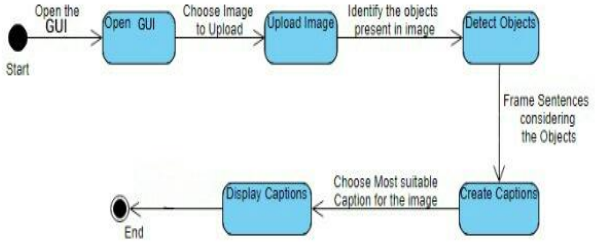


Figure 6: State Chart Diagram of Steps taken by the System.

The proposed system of Image Caption Generator has the capabilities to Generate Captions for the Images, provided during the Training purpose & also for the New images as well. Our Model takes an Image as Input and by analyzing the image it detects objects present in an image and create a caption which describes the image well enough for any machine to understand what an image is trying to say. [15]

5. IMPLEMENTATION OF THE SYSTEM

Here we will discuss the implementation of the system.

5.1. Object Detection

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

5.2. Sentence Generation

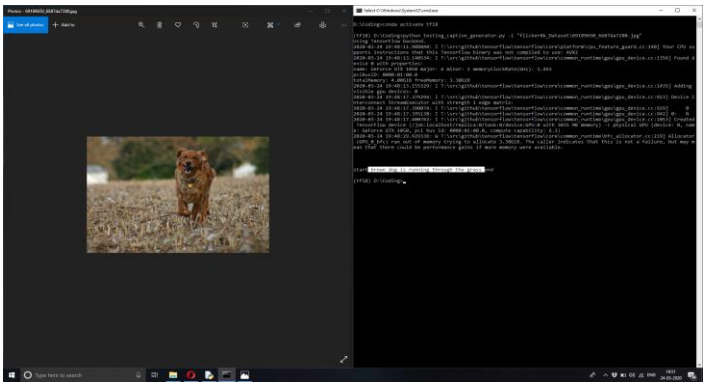
[12] By using LSTM, sentences are generated. Each predicted word is employed to get subsequent words. Using these words, appropriate sentence is formed with the help of

Optimal beam search. Here, Softmax function will be used for prediction of word.

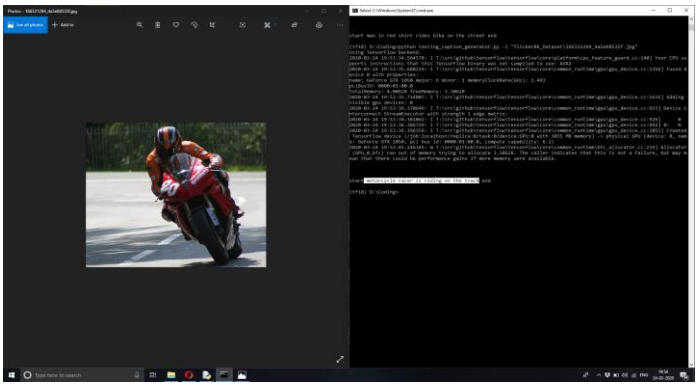
5.3. Deployment

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

6. RESULTS



Generated Caption: Brown Dog is running through the grass.



Generated Caption: Motor Cycle racer is Riding on the Track.

7. CONCLUSION

In this advanced Python project, an image caption generator 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.

8. REFERENCES

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

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

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

[4] D. S. Whitehead, L. Huang, H. and S.-F. Chang, "Entity-aware Image Caption Generation," in Empirical Methods in Natural Language Processing, Brussels, 2018.

[5] C. Elamri and T. Planque, "Automated Neural Image Caption Generator for Visually Impaired People," California, 2016.

[6] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han and Q. Liu, "Neural Image Caption Generation with Weighted Training and Reference," Cognitive Computation, 08 August 2018.

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

[8] S. Bai and S. An, "A Survey on Automatic Image Caption Generation," Neurocomputing, 13 April 2018.

[9] R. Staniute and D. Sesok, "A Systematic Literature Review on Image Captioning," Applied Sciences, vol. 9, no. 10, 16 March 2019.

[10] J. Hessel, N. Savva and M. J. Wilber, "Image Representations and New Domains in Neural Image Captioning," ACL Anthology, vol. Proceedings of the Fourth Workshop on Vision and Language, p. 29–39, 18 September 2015.

[11] M. Z. Hossain, F. SOHEL, M. F. SHIRATUDDIN and H. LAGA, "A Comprehensive Survey of Deep Learning for Image Captioning," ACM Journals, vol. 51, no. 6, 14 October 2018.

[12] A. Farhadi, . M. Hejrati, . M. A. Sadeghi and . P. Young, "Every Picture Tells a Story: Generating Sentences from Images," in ACM Digital Library, 2010.

[13] S. Yan, F. wu, J. Smith and W. Lu, "Image Captioning via a Hierarchical Attention Mechanism and Policy Gradient Optimization," LATEX CLASS FILES, vol. 14, 11 January 2019.

[14] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," CVPR 2015 Paper, December 2014.

[15] K. Simonyan and A. Zisserman, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION," in ICLR, 2015.

[16] J. Donahue, L. A. Hendricks and M. Rohrbach, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," CVPR 2015, vol. 14, 31 May 2016.