International Journal of Computational and Electronic Aspects in Engineering
Volume 3, Issue 2, June 2022, pp 26 - 31
https://doi.org/10.26706/ijceae.3.2.arset4383

Image Caption Generator Using CNN and LSTM

Davis S Cherian[1]
davischerian29@gmail.com

[1]Student, Department of Computer Science & Engineering, Christian College of Engineering & Technology, Bhilai, Chhattisgarh, India

***Abstract*—This project entitled "*Image Caption Generator Using CNN and LSTM*" is a work that demonstrates the automated generation of captions for a wide variety of images. This technology is used by major tech-giants like Google, Microsoft, IBM, etc. to generate captions for the huge dataset of images produced over various platforms and social media websites. This project embodies the use of various Artificial Neural Networks namely CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks) and LSTM (Long Short-Term Memory) Units. The functionality of the model developed using these neural nets. has been rendered to an interactive Web Application for the users to understand the methodology.**

***Keywords*—** Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory, Computer Vision, Natural Language Processing.

I. Introduction

"*Image Caption Generator Using CNN and LSTM*" is a project that illustrates how captions for a wide variety of unique images can be generated through an interactive Web Application. Generating captions for a huge number of images is a hectic task for humans but what if the machine itself starts telling the captions for images; that is a task with relevant importance in the Deep Learning domain. Tech-giants like Microsoft, Google, IBM, etc. use this image cation generation technique to create captions for a huge number of images.

The concepts of Computer Vision and NLP (Natural Language Processing) work at back-end to recognize the images and describe the captions in a simple English sentence. The model for accomplishing this task has been made using Convolutional Neural Networks (CNN), RNN (Recurrent Neural Networks) and LSTM (Long Short-Term Memory) units. As CNN extracts the features of the image, RNN forms a meaningful caption in English language. They together work with LSTM units, to perform automated image caption generation to make the machine capable of predicting captions for a set of images on which it has been trained.

The images above mentioned have been taken from the "*flickr8k_sau dataset*" which contains more than 8092 unique images mapped to five captions, in such a way that there are 8092*5 captions. The deep learning model has been trained in "*Kaggle Notebook*" with Python version 3.7.6. ResNet50 model has been used to extract the features from the image. The above notified Web Application has been developed using Flask framework in PyCharm, which is a Python IDE with version 3.7.9 being used for coding.

II. Related Work

A. Deep Learning

This Project belongs to the domain of Data Science and more specifically from Deep Learning based concepts. Deep Learning is a subset of Machine Learning and ML comes from the family of Artificial Intelligence [1].

Technical Article
Available online on – 06 August 2022

Cite this article – Davis S Cherian, "Image Caption Generator Using CNN and LSTM", *International Journal of Computational and Electronic Aspects in Engineering*, RAME Publishers, vol. 3, issue 2, pp. 26-31, 2022.
https://doi.org/10.26706/ijceae.3.3.arset4383

B. Convolutional Neural Networks (CNN)

CNN is a deep learning algorithm that takes an image as an input and extracts the features from it and classifies it as any entity [2]. CNN works on an image by assigning importance to specific features of the image and creates feature maps for the same. Feature maps are created for every feature in that image like a loopy circle pattern in the digit '9' is a feature [3], also it will catch out the vertical line and diagonal line patterns in it and multiply them with its original pattern detector and get activated only when that pattern is found.

If the image is of a 'Koala' then the features activated would be of its eyes, nose & ears, forming its head and there could be the features of hands & legs, forming its body.

Then there comes the ReLU function where all the non-negative numbers are kept as it is and the negative nos. are converted to zero (0) [4].

It is the Max pooling which comes next with the maximum digit in the grid chosen as the representative of the grid.

Fig.1 explains how a loopy circle pattern of the digit '9' after going through its filter and the ReLU function gets its max pooled layer with a 2*2 grid and a stride of 1 being used here. CNN with its Convolution + ReLU layers and Max pooling layers for different feature maps like that of a koala's head and body are explained in *Fig.2*, with flatten layer for reducing the dimensions of the array and a dense layer of Neural networks to predict whether it is a Koala.

C. Recurrent Neural Networks (RNN)

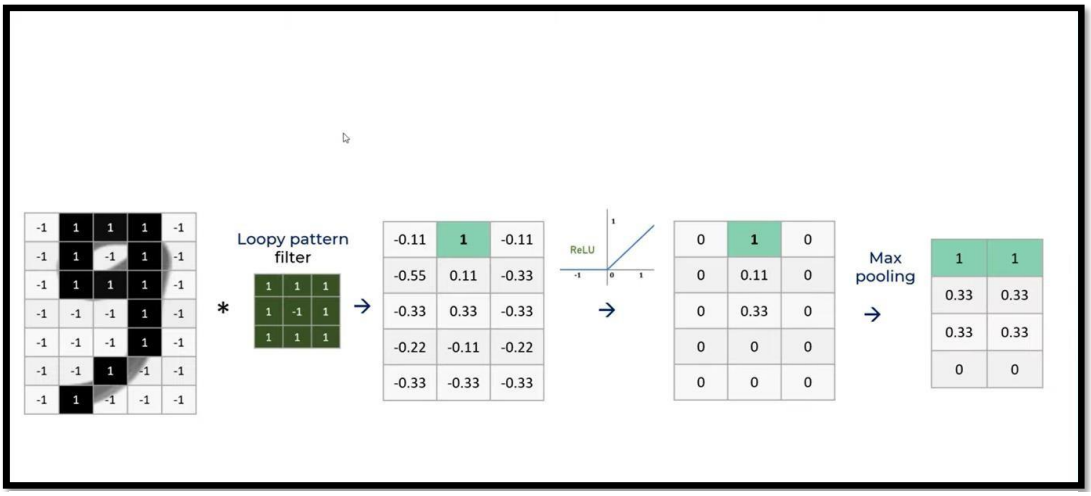

Fig. 1: Digit '9' going through its loopy circle pattern filter, the ReLU function and at last through Max pooling function

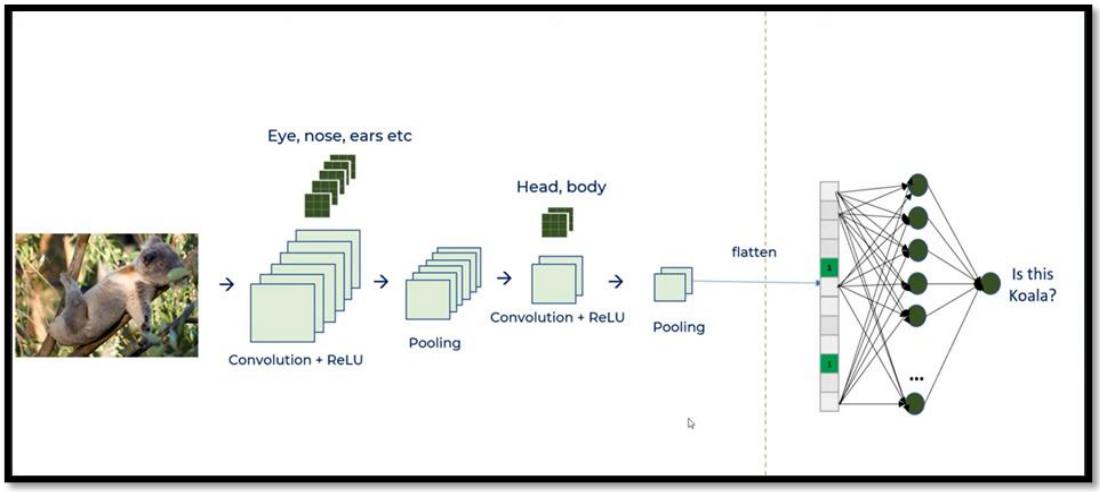

Fig. 2: Identifying Koala with CNN, with its Convolution part at the left-hand side of the dotted line and classification part at the right

Recurrent Neural Networks (RNN) is used in the project to predict the caption for an image, processed through CNN, in a natural language like English [5]. In RNN the output from the previous step is fed as an input to the current step. It remembers few words back to it or just a single word and to improve its short-term memory LSTM came into picture. RNN is used in auto completing sentences, like in Gmail; in translating a sentence from one language to another; in Sentiment Analysis, like that of Twitter; and in Named Entity Recognition (NER).

D. Long Short-Term Memory Units (LSTM)

LSTM or Long Short-Term memory units were developed to remember key words from an entire sentence which can prove to be the result generating key factor [6]. LSTM is one of the Recurrent Neural Networks but slightly better than them in terms of memory. They have a good hold over recognizing certain patterns because of which they work better.

III. MATERIALS AND METHODS

A. System Requirements

This Project has been implemented on a Windows 10 (64-bit Operating System) machine running on a Core i3 processor having a RAM of 4 GB.

B. Software Requirements

Below mentioned are the primary software and Python libraries required for developing the project:

- Python 3.7.9,
- Google Chrome,
- Kaggle Notebook (For the training of the neural network model and the validation steps),
- Computer Vision Packages (OpenCV),
- Deep Learning Neural Network Packages (Keras) [7],
- Data Analytics Packages (NumPy, Pandas, TensorFlow),
- Socket Programming Framework (Flask).

IV. METHODOLOGY

To develop the Web Application the following three phases were followed:

A. Training Phase: Kaggle Notebook

The training of the model was accomplished in Kaggle Notebook using images and caption dataset of the **"*flickr8k_sau dataset*"** database. The following figure (*Fig.3*) represents the description of the model developed.

The image pre-processing and text pre-processing have been done in this Kaggle Notebook and the vocabulary is created here. Model has been trained with 72.26% accuracy and 1% loss and is able to predict the captions only with that much accuracy with loss percentage being mentioned.

B. Python Scripting

Script for the Web Application was written in Python 3.7.9 using the PyCharm IDE Community Edition. To handle the client-server interaction Flask framework has been employed. Here, the client application is the website that serves the image (whose caption is to be generated) to the server application at the back end. The Server on the other hand receives the input image from the client and feeds it to the neural network model trained in the first phase. The model by virtue of CNN and RNN generates the words that make up for the caption. The recursive mechanism of LSTM combines the words generated to frame an appropriate caption. This caption generated is fed back to the client as a response to the query generated by the server.

A. Design Phase: Web Designing and Styling

HTML 5.0 has been used to generate the basic template for the webpage this includes the title mechanism for opening file explorer of user's local machine and the necessary buttons. CSS serves the purpose of styling the webpage to give a pleasant look and feel. Bootstrapping enables the dynamic content generation and delivery.

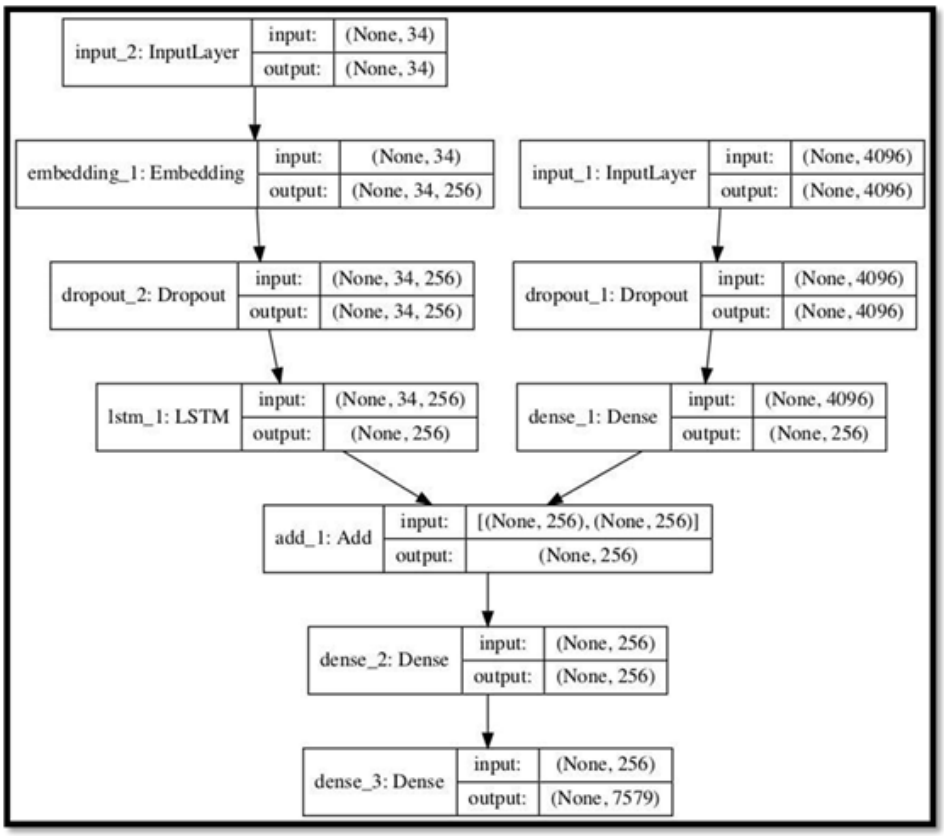

Fig. 3: Description of the Model developed using Kaggle Notebook

V. Results And Discussion

In this section we have provided few screenshots of the website developed in *Figs. 4, 5, 6 & 7*. Model has been trained for 50 epochs with 189 steps per each epoch to give a 72.26% accuracy and with 1% loss and is able to predict the captions only with that much accuracy.

Hence, we observe some words getting repeated in the captions generated. Also, for some set of images the caption produced is not much grammatically correct but gives some meaningful insights of the theme visualized.

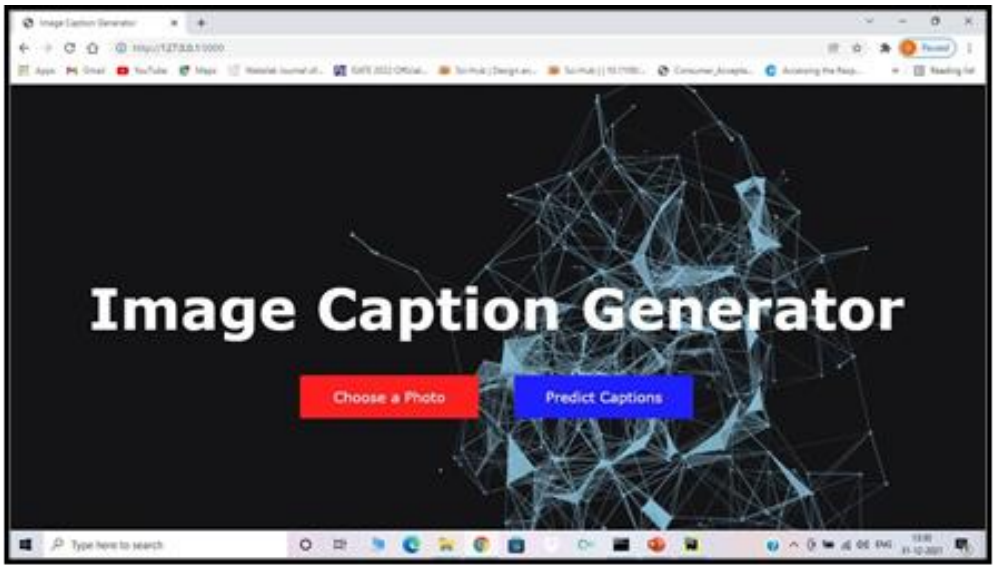

Fig. 4: The Welcome Page of the Caption Generator

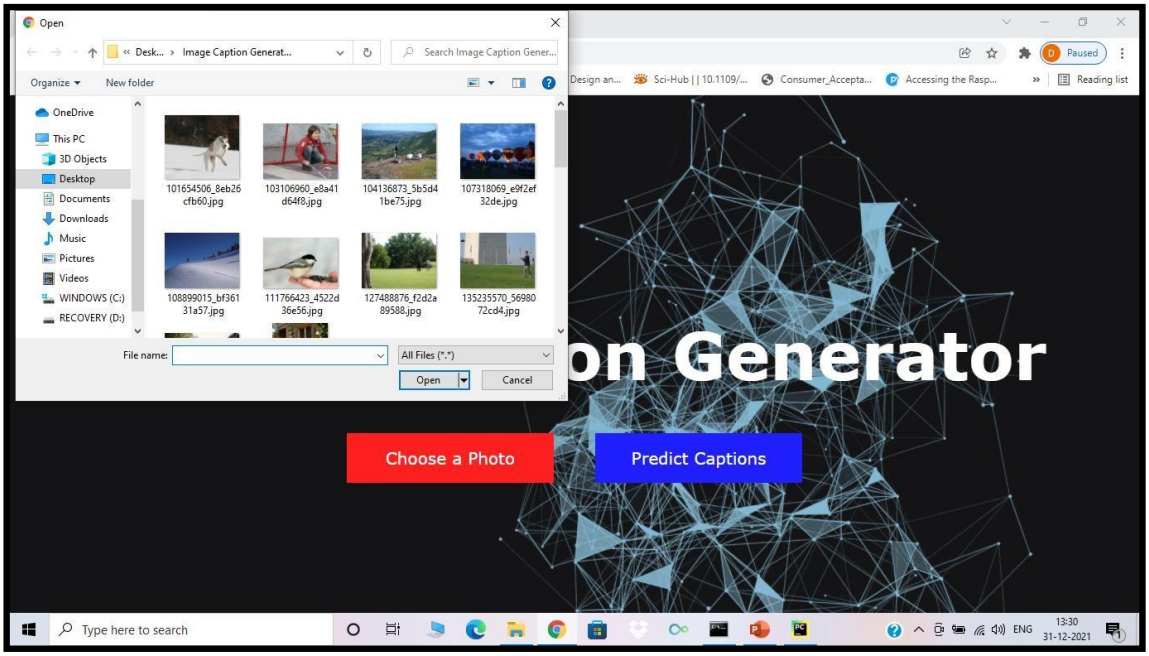

Fig. 5: Selecting Images for Caption Generation

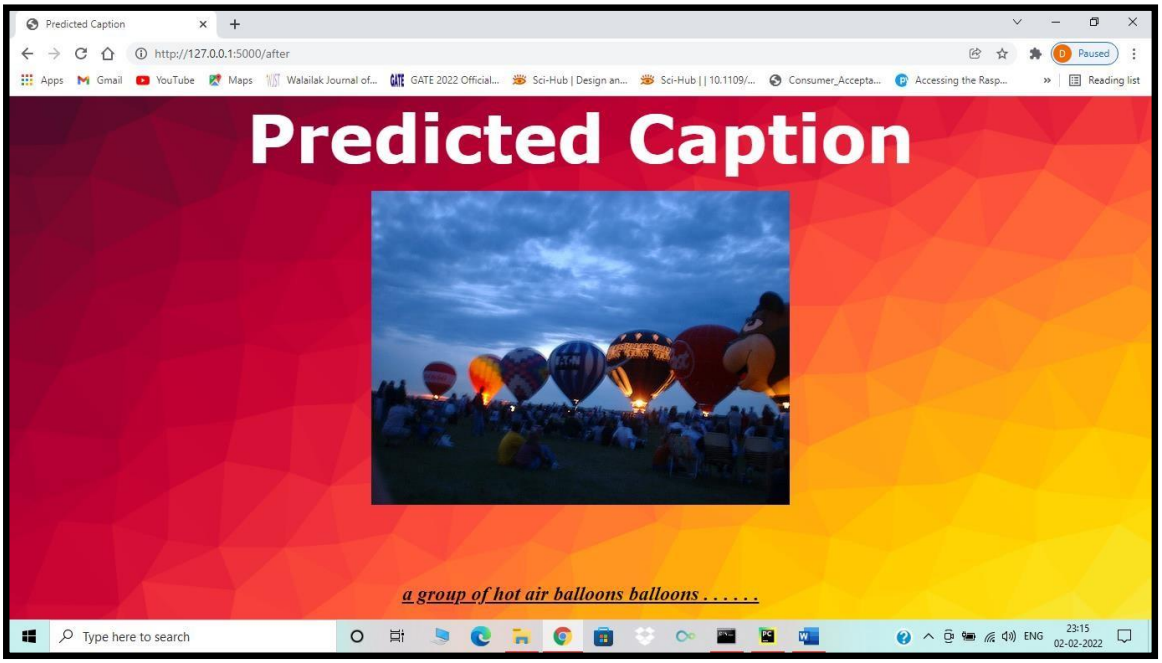

Fig. 6: Predicted Caption- “A Group of Hot Air Balloons”

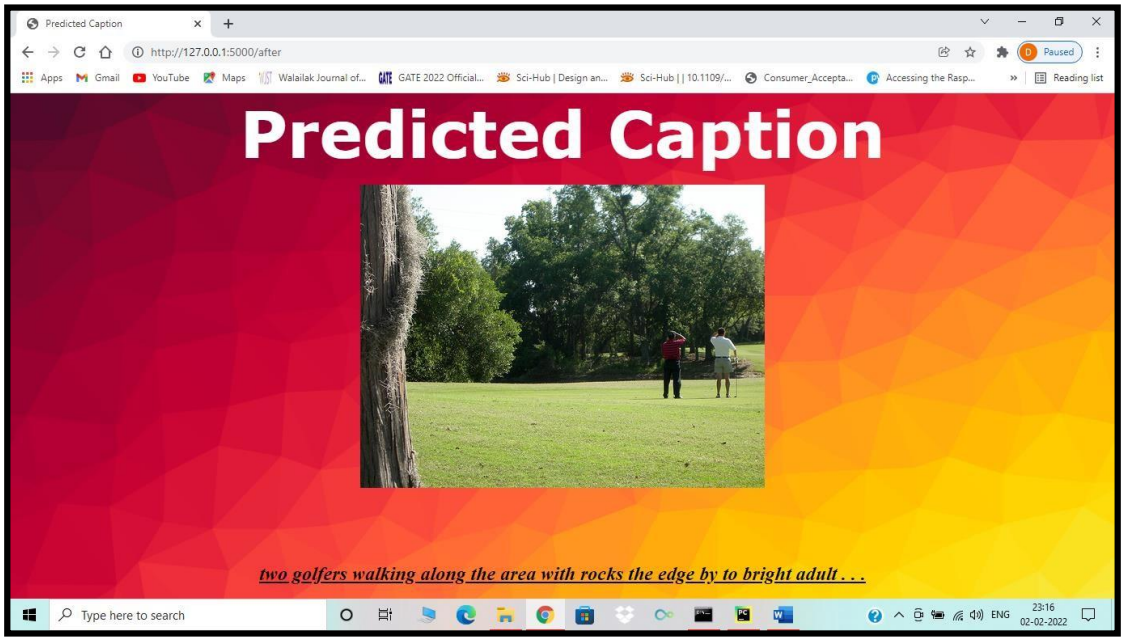

Fig. 7: Predicted Caption- "Two Golfers walking along the area"

VI. Conclusion And Future Scope

This Web-Based Application will help predict captions for a set of 1500 images only from the dataset which contains over 8000 images. This project displays the use of CNN to extract features from an image and predict the caption for that image using RNN embedded with LSTM. The future scope of this project could be the changes made be incorporated so that the model not only translates but catches attention from the images. Also, this project with a well wider scope could prove helpful for the visually impaired to see world from a different perspective.

References

[1] J. D. Kelleher, Deep Learning, The MIT Press, 2019.

[2] P. Kim, "Convolutional Neural Network," MATLAB Deep Learning, p. 121–147, 2017.

[3] U. Karn, "An Intuitive Explanation of Convolutional Neural Networks," 2016. [Online]. Available: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/.

[4] M. Glossary, "Activation Functions," 2017. [Online]. Available:https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html

[5] A. L. Caterini and D. Chang, "Recurrent Neural Networks," Deep Neural Networks in a Mathematical Framework, pp. 59-79, 2018.

[6] G. V. Houdt, C. Mosquera and G. Nápoles, "A review on the long short-term memory model," Artificial Intelligence Review, vol. 53, p. 5929–5955, 2020.

[7] A. Gulli and S. Pal, Deep Learning with Keras, Packt Publishing Ltd., 2017.