Image Caption Generator Using Deep Neural Networks

Rajan Puri[1] Ishu Singhal[2] Apoorva Jain[3] Ms. Garima Singh[4]
[1,2,3]Student [4]Assistant Professor
[1,2,3,4]Department of Computer Science & Engineering
[1,2,3,4]Dr Akhilesh Das Gupta Institute of Technology and Management, Shastri Park, Delhi, India

Abstract— In this era, image captioning has become one of the foremost wide needed tools. Moreover, there are intrinsic applications that generate and supply a caption for a particular image, all these things are done with the assistance of deep neural network models. The method of generating a description of an image is termed image captioning. It needs recognizing the important objects, their attributes, and therefore the relationships among the objects in a picture. It generates syntactically and semantically correct sentences. In this paper, we tend to present a deep learning model to describe an image and generate captions using computer vision and machine translation. So, we consistently analyze deep neural networks based mostly on image caption generation methodology. With a picture because of the input, the tactic will output an English sentence describing the content within the image. We analyze 3 elements of the method: convolutional neural network (CNN), recurrent neural network (RNN) and sentence generation. Initially, the input image is born-again to a grayscale image that's processed through the Convolution Neural Network (CNN) to properly determine the objects. Objects within the image area unit are properly identified, that is then converted to text messages.

Keywords: Image Captioning Generator, CNN, RNN, LSTM, Deep Learning, Neural Networks, Image, Caption, Xception

I. INTRODUCTION

Automatically describing the content of pictures using natural languages which could be an elementary and difficult task. It has a significant potential impact. For instance, it may facilitate visually impaired folks to perceive the content of pictures online. Also, it may give additional correct and compact info of images/videos in situations like image sharing in social networks or video police investigation systems. This project accomplishes this task by making use of deep neural networks. By learning data from image and caption pairs, the strategy will generate image captions that are typically semantically descriptive and grammatically correct.

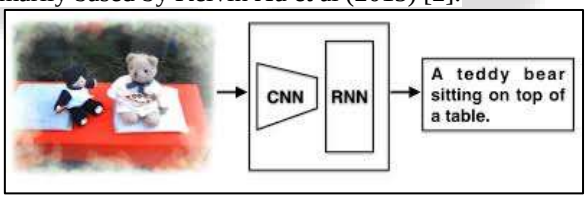
In fact, a blueprint should catch the things contained in an image, nevertheless it in addition ought to communicate however these articles establish with one another additionally as their traits and consequently the exercises they are related to, yet it should even be sufficiently clever to catch and categorical article's connections within the tongue. Its motivation is to mirror the human capability to grasp and method gigantic measures of visual information into an elucidating language, making it a perfect issue within the sphere of Artificial Intelligence.

Picture subtitle generator can be an enterprise that includes computer vision and tongue-reading ideas to acknowledge the setting of a picture and portray them throughout a tongue like English.

Convolutional Neural systems are specific profound neural systems which can method the information that has input shape reasonably a 2-dimensional framework. Infact, Human beings typically describe a scene by making use of natural languages which are aphoristic and compact. However, machine vision systems describe the scene by taking a picture that could be a 2 dimensional array. Images are effectively spoken as a 2-dimensional grid and CNN is inconceivably valuable in working with Images.

LSTM represents long momentary memory; they are a kind of RNN (recurrent neural network) that is all around coordinated for succession forecast problems. Uphold the past content, we'll foresee what the following word is. It substantiated itself powerful from the standard RNN by conquering the restrictions of RNN that had momentaneous memory. LSTM will perform important information at some stage in the handling of sources of information and with an overlooked entryway, it disposes of non-important information.

Many important tech-organizations are setting intensely in Deep Learning and AI cross-check, as a result of that the important issue of image subtitling is being learned at a number of associations by a number of distinctive teams. The Two principle assortments of labor that structure the rationale of this paper are Show and Tell by O.Vinyals et al (2015) [1] and furthermore the any developed, thought primarily based by Kelvin Xu et al (2015) [2].



II. PROJECT DESCRIPTION

The objective of our project is to be told the ideas of a CNN and LSTM model and build an operating model of Image caption generator by implementing CNN with LSTM.

In this Python project, we are going to be implementing the caption generator mistreatment CNN (Convolutional Neural Networks) and LSTM (Long short term memory). The image options are going to be extracted from Xception that may be a CNN model trained on the imagenet dataset so we have a tendency to feed the features into the LSTM model which is able to generate the image captions.

III. METHODOLOGY

Here we have a tendency to use CNN and LSTM model to achieve our goal (image caption generator)

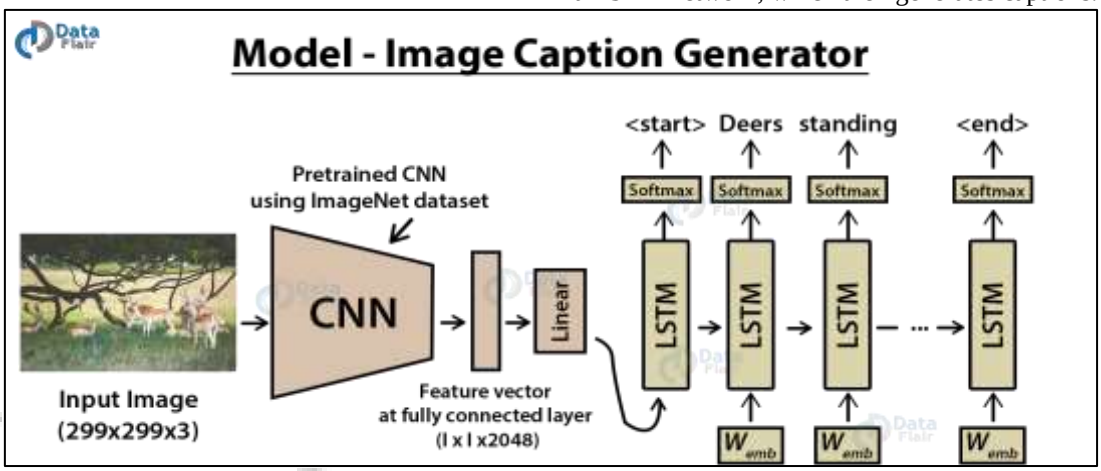
– CNN-Convolutional Neural Network is a man-made deep learning neural network. It is used for image classifications, computer vision, image recognition and

All rights reserved by www.ijsrd.com

Object detection. CNN image classifications take an input image, process it and classify it underneath certain classes (Eg., Dog, Cat,etc). It scans pictures from left to right and high to bottom to tug out vital options from the image and combines the feature to classify pictures.

– LSTM stands for Long short term memory, they're a sort of RNN (recurrent neural network) that has a similar temperament for sequence prediction problems. Supported by the previous text, we are able to predict what succeeding words are. it's proved itself effective from the standard RNN by overcoming the constraints of RNN that had short term memory. LSTM will do relevant data throughout the process of inputs and with a forget gate, it discards non-relevant data.

We merged these two architectures in one model referred to as a CNN-RNN model.in general Our approach attracts the success of the top-down image generation models listed above. We have a tendency to use a deep convolutional neural network to extract the visual image features and semantic features are extracted from the semantic tagging model. Visual options from CNN and semantic features from tagging model are concatenated and feed because the input to a LSTM network, which then generates captions.



IV. MODEL

Our model comprise of three main phases:

1) Image Feature Extraction: The options of the pictures from the Flickr 8K dataset are extracted by applying the Xception model because of the performance of the model in object identification. The Xception is a convolutional neural network that consists of thirty six, as this model configuration learns very quickly. These are processed by a Dense layer to produce a 2048 vector part representation of the image and passed on to the LSTM layer.
2) Sequence processor: The role 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 specified features of the text and consists of a mask to ignore cushiony values. The network is then connected to a LSTM for the ultimate stage of the image captioning.
3) Decoder: The final section of the model combines the input from the Image extractor section and therefore the sequence processor stage using further operation then fed to a 256 neuron layer then to a final output Dense layer that produces a softmax prediction of subsequent word within the caption over the whole vocabulary that was shaped from the text information that was processed within the sequence processor section.

V. EXPERIMENT

A. Image Captioning Datasets:

For the image caption generator, we are going to use the Flickr_8K dataset. To fulfill our purpose there are also other and huge datasets like Flickr_30K and MSCOCO dataset, however it can take weeks simply to train the network thus we are going to be employing a small Flickr8k dataset. The advantage of a large dataset is that we are able to build better models.

Flickr8k dataset contains a range of pictures depicting scenes and things. The dataset consists of 8000 images with various dimensions and every image has five corresponding descriptions. We divide the data into 6000, 1000, & 1000 images as training, validation and testing sets respectively.

B. Preprocessing for captions

(Description of Image):

Each Image has five descriptions (captions), The main role here is of Clean() function that takes all descriptions and performs a basic data clean :

1) Removing punctuations
2) Removing Words that contain numbers
3) Converting all description in lowercase
4) Removing special characters (like '%', '$', '#', etc.)
5) Applied tokenization to our dataset and fixed vocabulary size.

C. Preprocessing for Images:

Some preprocessing of images are required before feeding it to our model:

1) Resize each image to (299 * 299) size as Xception model takes
2) Flatten it
3) Scaling image pixels (normalization)

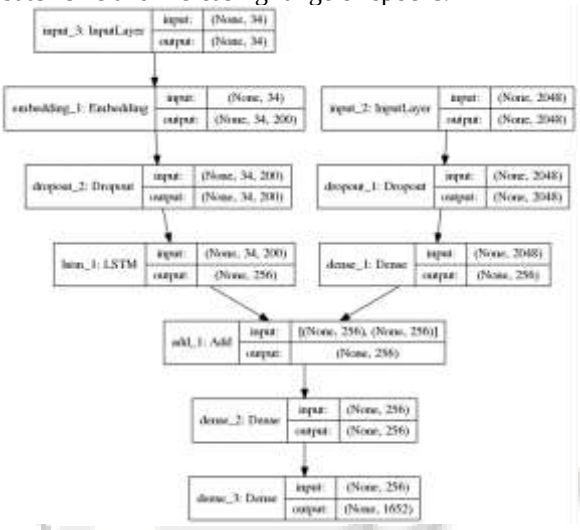
All rights reserved by *www.ijsrd.com*

D. Training Our Model

We have used GloVe stands for world vectors for word representation. It's an unsupervised learning algorithmic program developed by Stanford for generating word embeddings by aggregating global word-word co-occurrence matrix from a corpus. Also, we've 6000 pictures and every image has five captions related to it. It suggests that we've 30000 examples for training our model.

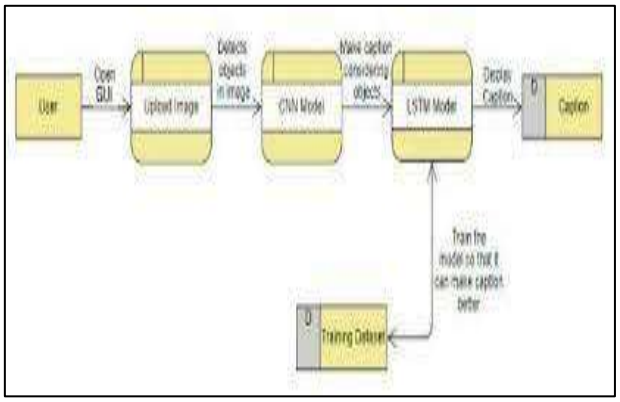
We used Keras Model from the functional API. it's 3 major steps: Processing the sequence from the text, Extracting the feature vector from the image and decoding the output by concatenating the on top of 2 layers.

For training our model, we employ Adam's optimizer and loss operate as categorical cross-entropy. I'm training the model for fifty epochs which can be enough for predicting the output. just in case if you have got additional machine power (no. of GPU's) you'll train it by decreasing batch size and increasing range of epochs.



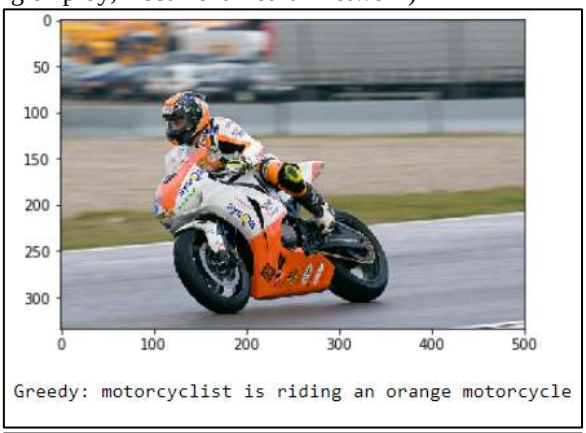
E. Implementation

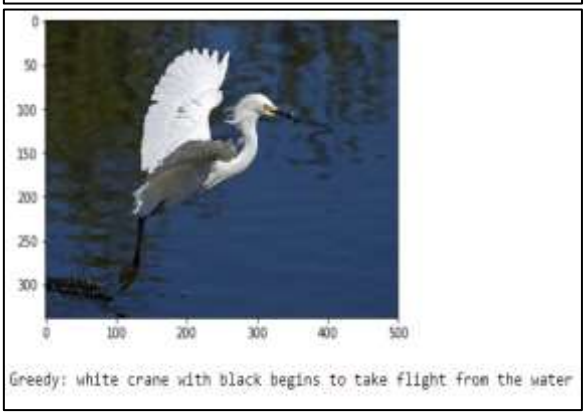
The implementation of the model was done in Python. Keras 2.0 was accustomed to implementing the deep learning model. Tensorflow library is put in as a backend for the Keras framework for making and training deep neural networks. TensorFlow is a deep learning library developed by Google. The neural network was trained on google colab .It provides a heterogeneous platform for execution of algorithms i.e. It may be run on low power devices like mobile furthermore as a massive scale distributed system containing thousands of GPUs. The neural network was trained on the Nvidia Geforce 1050 graphics process unit which has 640 Cuda cores. So as to outline the structure of our network, TensorFlow uses graph definition. Once the graph is defined it may be run on any supported devices. The photo features are pre-computed using the pretrained model and saved. These features are then loaded into our model as the interpretation of a given image within the dataset to scale back the redundancy of running every image through the network every time we would like to check a replacement language model configuration. The preloading of the image features is additionally finished with a real time implementation of the image captioning model.



VI. RESULT

Our model supported multi label classification employed on fast Text and CNN, is beneficial in sleuthing and extracting objects from image and generates caption consistent with the provided datasets. We've given multiple approaches for Image caption Generator like (Convolution neural network, Long employ, Recurrent Neural Network)





All rights reserved by *www.ijsrd.com*

VII. CONCLUSION AND FUTURE SCOPE

Picture inscribing might be associated with energizing activity and raises extreme contention among specialists. There is an ever increasing variety of researchers who are selecting to investigate this examination field that the measure of information is consistently increasing. It utterly was seen that the outcomes are ordinarily contrasted, in spite of the actual fact that there are several late ones, with abundant higher outcomes and new thoughts for upgrades. Additionally it will in any case not be satisfactory that if Flick8k datasets are useful and enough for our model by checking and assessment and within the event that they will probably by serve very little and adequately well and whereas having as a primary concern wandered conditions.

In this paper, we've got an actual CNN-RNN model by building a picture subtitle generator. We have a tendency to use a touch low dataset comprising 6000 images. For creation level models, we'd like better to mentor on datasets larger than 100,000 images which can deliver higher exactitude models

ACKNOWLEDGEMENT

The writers are very grateful to Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi, India, for providing eminent computation amenities in the College campus. Authors would also like to pay regards to the Director of College, Department HOD and colleagues for giving their ethical guidance and assistance in this research work.

REFERENCES

[1] O. Vinyals, A deep convolutional activation feature for generic visual recognition. In ICML, 2015.
[2] K. Xu (2015) Show and tell: image caption generation with all the visual data attention. inProc. Int. Conf. Mach.
[3] Farhadi A. et al. (2010). Every Image Tells a Story: Generating description from Images. Lecture Notes in Computer Science, vol 6314. Springer, Berlin, Heidelberg
[4] S. Li, Composing simple image descriptions using web scale n-grams.
[5] Kulkarni G, Li S,Baby Talk: Understanding and Generating Image Descriptions. IEEE Conference on Big Computer Vision and small Pattern Recognition (CVPR) (20-25 June 2011).
[6] Show and Tell: A neural image caption generator, every picture tells a story: Generating sentences from images. In ECCV, 2010.
[7] M-RNN: Explain images with multimodal recurrent neural networks. In arXiv: 1410.1090, 2014.
[8] https://practice.geeksforgeeks.org/problems/number-of-pairs/0/
[9] B.Krishnakumar,K.Kousalya,S.Gokul,R.Karthikeyan, D.Kaviyarasu (IJAST)Vol.29 NO.3s(2020) :Image Caption Generation Using Deep Learning.
[10] Seung-Ho Han, Ho-Jin Choi (2020) IEEE International Conference on Big Data and Smart Computing (BigComp) : Domain-Specific Image Caption Generator with Semantic Ontology

All rights reserved by *www.ijsrd.com*