International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 289-293
Published Online June 2022 in IJEAST (http://www.ijeast.com)

IMAGE CAPTION GENERATOR USING DEEP LEARNING

Chaithra V, Charitra Rao, Deeksha K, Shreya,
Student
Dept. of Information Science and Engineering, CEC, Bantwal

Dr. Jagadisha N
Associate Professor & Head of Dept. of Information Science and Engineering, CEC, Bantwal

***Abstract*—In the last few years, the problem of generating descriptive sentences automatically for images have gained a rising interest in Natural language processing (NLP). Image captioning is a task where each image must be understood properly and are able generate suitable caption with proper grammatical structure. To describe a picture, you need well-structured English phrases. So to do this a computer system must use a strong algorithm and automatically defining image content is very helpful for visually impaired people to better understand the problem. Here, a hybrid system which uses multilayer CNN (Convolutional Neural Network) for features extraction which narrates given input image and LSTM (Long Short Term Memory)for constructing captions in a meaningful way with reference to features extracted in CNN. Convolution Neural Network (CNN) proven to be so effective that they are a way to get to any kind of estimating problem that includes image data as input. LSTM was developed to avoid the poor predictive problem which occurred while using traditional approaches. The model will be trained like when an image is given model produces captions that almost describe the image. The efficiency is demonstrated for the given model using Flickr8K data sets which contains 8000 images and captions for each image.**

***Keywords*— Convolutional Neural Network, Long Short Term Memory, Natural Language Processing, Deep Learning, VGG-16.**

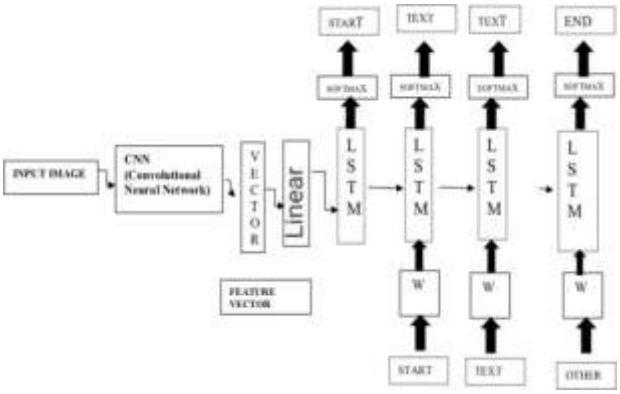
I. INTRODUCTION

Images are the ones that we will be encountering daily. Even though if there are no caption/description for the image, humans can easily understand those images. But for system/machine to perform similar operation is very difficult. To achieve this it includes a model from NLP (Natural Language Processing) for converting image to captions as a meaningful sentence. Image captioning contain various types of applications such as advising for editing applications, for image indexing, use in online personal assistants, for visually impaired people, in social media, and other applications of natural language processing.

Recently, Deep Learning models are able to achieve good results when it comes to predicting captions when an input image is given. Instead of need of composite data editing or a pipeline of specially designed models, one end-to-end model can define the captions, if an image is provided. The model here is being testing with the dataset called Flickr8k.

II. SYSTEM ARCHITECTURE

The figure below depicts the model for image caption generator. Firstly, an image is given as input which is processed by CNN (Convolutional Neural Network). This CNN will form a feature vector which will be dense in nature. Another name for dense vector is embedding. This embedding can be given as input to various other available algorithms. As output, a descriptive sentence that is caption is obtained which is suitable for inputted image. This output is obtained using Long Short Term Memory (LSTM) which will be helpful for obtaining suitable caption for the given inputted image using the feature vectors obtained in the previous stage.



III. IMPLEMENTATION

The image caption generation as the name suggests is the process of understanding the context of the image and using deep learning techniques relevant captions for the image can

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 289-293
Published Online June 2022 in IJEAST (http://www.ijeast.com)

be generated. During training of the model, input image need to be labelled with English keywords using the dataset. In order to generate suitable caption for the inputted image, CNN and LSTM has been used.

Data: The dataset which we have used is Flickr8k which is a standard dataset downloaded from Kaggle. The dataset has totally 8000 images where each of the image has 5 captions which makes it possible to cover all possible scenarios. The dataset containsFlickr_8k.trainImages.txt file which has 6000 images which is training dataset, Flickr_8k.devImages.txt file which has 1000 images which is validation\development dataset, and at last Flickr_8k.testImages.txt file which has 1000 images which is the test dataset. The dataset images are taken from six groups of Flickr and it doesn't contain any popular personalities, or even places. The dataset images areof different scenes that are selected manually. The dataset size is 1GB which is available and can be freely downloaded inKaggle.

Model: Machine learning process is carried out by deep learning using artificial neural network that consist of many levels arranged in a structural order. In the model the flow of data starts from the initial level, here the model learns something simple and the result is sent to layer two of the network. Later the input is combined into something which is a bit more difficult/complicated and finally sent to the third level. This procedure is continued until the network gives something more complicated from the input.

Convolutional neural network: Convolutional neural network (CNN)are the ones that are used to process the data that consists of input image shape such as 2D matrix. CNN is one of the best technique that is suitable for image classification and identification. CNN will take the input image and for those input image it will assign the weights and biases to various features in the input image. Firstly an input image is given which is processed by CNN. This CNN will form a feature vector which will be dense in nature. This dense vector is also called as embedding. This embedding can be given as input to various other available algorithms. As output, a descriptive sentence that is caption which is suitable for inputted image is obtained.

Long Short Term memory: LSTM helps to maintain the connectivity with the previous word and will be able to remember the long term dependencies. Long short term memory will observe the previous word and predict what the next word will be. Remembering information/data for longer intervals is one of their behavior which is default, and this behavior is mainly controlled with the help of gates present. LSTM consists of input gate, output gate and forget gate. Here LSTM translates the features and objects extracted by CNN to a natural sentence that is caption that describes the image in a suitable way.

IV.EVALUTION

A. Data Cleaning and Pre-processing:

The program here starts by loading both the text file, and the image file in separate variables. Here the text file will be stored in a string, later this string will be used and manipulated in such a manner for the creation of a dictionary that will map every image to a list of five descriptions. The main process of data cleaning is not using punctuation marks in the captions and if there is a text that consists of capital letters that should be eliminated because only lowercase letters are allowed and the words containing numbers should be removed.Further, creation of vocabulary which consists of unique words from each of the descriptions takes place, which will be used in caption generation for the test images.We have to append the tags <start> and <end>identifier for each caption since these both tags will act as indicators for LSTM to recognize where a caption is starting and ending.

B. Extraction of feature vectors:

A feature vector which is simply nothing but a feature, is a numerical value in the matrix form, consisting information about an object's essential features. Example: in our case each pixel of image has a distinct intensity value. And pickle file is used to store the feature vectors. In this model, Transfer Learning has been used which is simply a pre-trained model called VGG-16 used for feature extraction. The VGG-16 model is of 16 layers deep. VGG-16 model has been trained on Image net dataset. Image net dataset has more than 1000 divergent classes and up to millions of images. We can directly import this model from the keras. applications. Python helps using this model extremely easy in our code with keras.applications.vgg16 module. Since the VGG16 model was originally utilized/built for Image net dataset, little changes for integrating with our model is done. One main thing that should be noticed here is that the VGG16 model takes 224*224*3 size of image as input. We removed the last classification layer and get the 2048 feature vector. For each image the weight will be downloaded and each image name is mapped with their respective feature array. Based on processor this process will take some time.

C. Layering the CNN-RNN model:

1.**Feature Extractor**: This is used to minimize the dimensions from 2048 to 256.We have used Dropout layer where one of this will be added in CNN and LSTM. Pre-processing of the input images is performed by VGG16 model in the absence of output layer and then make use of the extracted features which is predicted by this model as an input.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 289-293
Published Online June 2022 in IJEAST (http://www.ijeast.com)

2.**Sequence Processor**: The Embedding layer present here is used to handle textual input, which is next followed by the layer of LSTM.

3.**Decoder**: Merging of the above two layers takes place, and to make the final predictions Dense layer is used. The Feature Extractor and Sequence Processor produces an output which is fixed-length vector which will be together merged and later processed by Dense layer. The number of nodes that will be present in the final layer will be same as the vocabulary size.

D. Training the model:

For training the model we are using 6000 images which has feature vector of 2048 long. Memory present is not enough to hold all this data, so Data Generator is used. This helps in the creation of batches of data and later improve in terms of speed. Next defining the number of epochs i.e number of iterations of the training dataset. The epochs should be selected such that the model is neither over fitted nor under fitted. For this process model. fit_ generator() will be used. This may take a while as it depends completely on the processor. The sequence creator is created, based on the previous word/image feature vectors it will predict the next word. Development dataset consists of 1000 images and can be used while training the model for monitoring the model performance and to decide the when the model version will be saved to the file. Several models will be saved, in which any of the model can be used for testing. The final step after training the model is testing the model.

E. Testing the model:

The model that is saved in the previous step will be loaded now to generate the final predictions. A sequence generator is used along with the tokenizer file. Here for generating the captions, the initial step of extracting feature for the given image will be performed. The path of the image from the testing dataset can be given in the function. We also can iterate through the entire test dataset and later the predictions can be stored for every image in a list or a dictionary. The proper functioning of caption generation involves using the start sequence and the end sequence and to call the model iteratively to generate meaningful captions for given input image.

V.RESULTS

Fig. Home Page

The above figure shows the first view of home page where we click a button “Get Started” to get the caption for the inputted image.

Fig. Model Details

The above figure shows the second view of home page which shows the details of dataset used, how the dataset is being split, Algorithms used, IDE/Tools/Libraries Used, Optimizers Used for proposed model and Compared with existing model with their accuracy.

Fig. Training Details

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 289-293
Published Online June 2022 in IJEAST (http://www.ijeast.com)

The above figure shows the third view of home page which displays the model training details such as number of epochs, accuracy obtained for the proposed model, optimizer used. We have also included window snip of training the model with 1000 epochs where the accuracy is obtained as 78% after the completion of iterations.

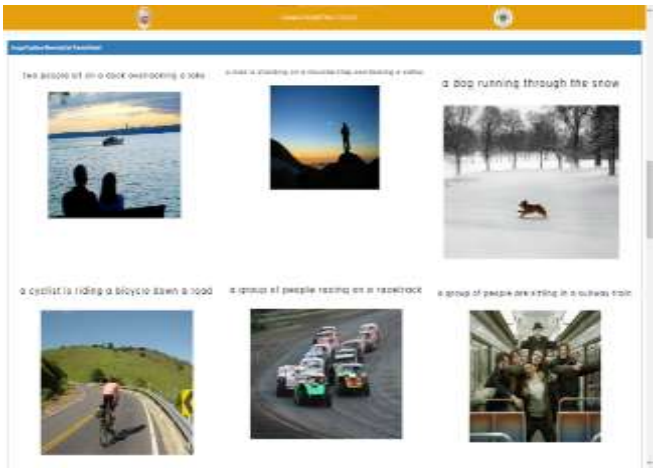


Fig. Image Caption Generated

The above figure shows the fourth view of home page where we can see how the caption is being produced for some of the sample input images.

Fig. Normal image vs Blurred image

The above figure is the fifth view of home page where we can see the comparison between Normal image and Blurred image. We can see how the caption is being generated for both the images. In both cases of images, the caption generated was appropriate.

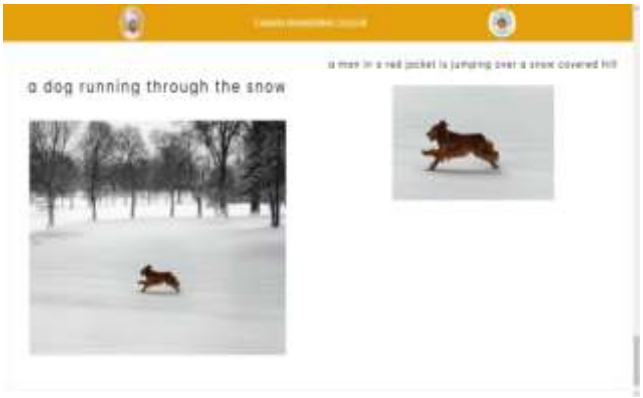


Fig. Normal image vs Cropped image

The above figure is the last view of home page where we can see the comparison between Normal image and Cropped image. This picture depicts how the caption is generated for both the images. In first cropped image, the caption predicted was suitable for the image where as in second cropped image the caption obtained was not suitable.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 289-293
Published Online June 2022 in IJEAST (http://www.ijeast.com)

Fig. Input for Caption Generation

The above figure is the screen obtained when we click on "Get Started" which was on the first view of home page. In order to get the caption for the image user require, he has to click on Choose file and select that jpg/png image. Later click on Generate Caption for the caption to be seen on the screen for the inputted image.

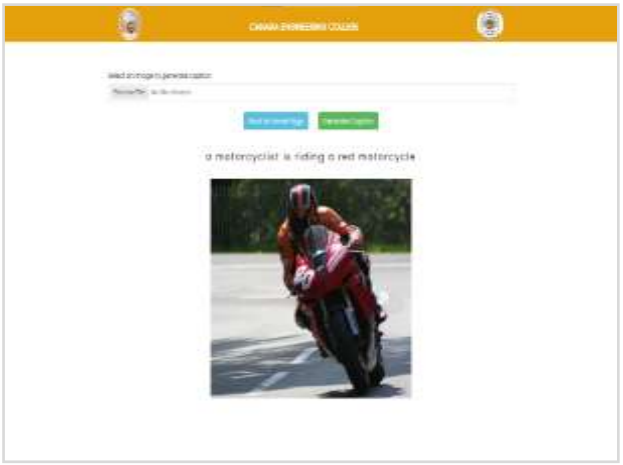


Fig. Caption Generated

The above figure shows the caption generated along with the input image selected by the user. User can choose image again and generate the caption else can go back to home page.

VI. CONCLUTION

Image caption methods based on deep learning have made remarkable progress in recent years and it produces high quality captions for every image to be achieved. With boom of novel deep learning network architectures, automatically captioning an input image will remain as a functioning study area for some time. The goal of image captioning is very huge in the future since use of social media is increasing day by day to post photos and so on. So this model can be used in such cases and will be available for help in greater extent. The proposed model automatically generates captions for an image using Neural Network and Natural Language Processing techniques in VGG 16 model. CNN and LSTM have been combined to work well together and were able to find a connection between objects in images to generate the right caption. The dataset used for training the model is Flickr8k. The Flickr8k dataset includes about 8000 images, and suitable captions are also saved in a text file. The main point of our model is it is dependent on data, the prediction cannot be done for the words, if that word is not present in vocabulary. The model performance can be expected to increase once it is trained with larger datasets.

VII. REFERENCES

[1]. T Swarnim ,S Ravi. (2021) Image Caption Generator Using CNN and LSTM.
[2]. S P Sreejith, A Vijayakumar. (2021) Image Captioning Generator using Deep Machine Learning. (Pg 832-834)
[3]. C Sneha, B Premanvitha , B Shanmukh, C Kavitha. (2021) Image Caption Generator Using Deep Learning. (Pg 1015-1031)
[4]. J P Megha, U Vikas, S Gunjan, M Vrinda. (2021) Image Caption Generator.(DOI: 10.35940/ijitee.C8383.0110321)
[5]. K Preksha, D Vishal , K Aishwarya, K Prachi. (2021) Image Caption Generator using CNN-LSTM.(Pg 4100-4105)
[6]. Mohamed Ashraf Ali. (2020) Image Caption Using CNN and LSTM.
[7]. K Varsha, M Vaidehi, K Megha. (2019) Deep Learning based Automatic Image Caption Generation.(Global Conference for Advancement in Technology)
[8]. A Chetan , J Vaishali. (2018) Image caption Generation Using Deep Learning Technique.(Fourth International Conference on Computing Communication Control and Automation)
[9]. S Lakshminarasimhan, Sreekanthan Dinesh, Amutha A.L. (2018) Image Captioning - A Deep Learning Approach.(Pg 7239-7242)
[10]. D Subrata, J Lalit, D Arup. (2018) Deep Learning for Military Image Captioning.(Pg 2165-2171)
[11]. M Pranay, G Aman, Y Aayush, M Anurag, B NandKumar. (2017) Camera2Caption:A Real-Time Caption Generator.(International Conference on Computational Intelligence in Data Science)
[12]. C Jianhui,D Wenqiang, L Minchen. (2015) Image Caption Generator Based On Deep Neural Networks.
[13]. V Oriol, T Alexander, B Samy, E Dumitru. (2015) Show and Tell: A Neural Image Caption Generator.(Pg 3156-3164)