Building A Voice Based Image Caption Generator with Deep Learning

Mohana priya R
U.G Scholar
Department of Information Technology
Sathyabama Institute of science and Technology
Chennai, India
amiraakash99@gmail.com

Dr.Maria Anu
Professor
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, India
maria.it@sathyabama.ac.in

Divya S
U.G Scholar
Department of Information Technology
Sathyabama Institute of science and Technology
Chennai, India
sureshdivya821@gmail.com

2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432091

Abstract—. ***Image processing is used in various industries and it is remaining as one of the most advanced technologies used in Google, medical field etc. Recently, this technology has also attracted many programmers and developers due to its free and open source tool, which every developer can afford it. Image processing also helps in finding out lot of information from a single image since it is currently utilized as a primary method for collecting the information from image and processing it for some purpose and some operations will also be performed on the image. A voice based image caption generation is a task that involves the NLP (natural language processing) concept for understanding the description of an image. The combination of CNN and LSTM is considered as the best solution for this project; the main target of the proposed research work is to obtain the perfect caption for an image. After obtaining the description, it will be converted into text and the text into a voice. Image description is a best solution used for a visually impaired people who are unable to comprehend visuals. With the use of a voice based image caption generator, the descriptions can be obtained as a voice output, if their vision can't be resorted. In future, image processing will emerge as a significant research topic, which will be primarily utilized to save human lives.***

Keywords: NLP (natural language processing), CNN (Convolutional neural network), LSTM (Long short term memory) ,RNN(recurrent neural network)

INTRODUCTION

In recent years Deep learning is one of the most used trends in Machine Learning and artificial intelligence, it is a machine learning Technique inspired by the Human brain, it uses the algorithm like convolutional neural network, recurrent neural network, long short term memory etc., where there are many developments had already made for visually impaired people, voice based Image caption generator is used to identify the objects and information present in the image, which could improve the lives of Visually impaired people, Using CNN and LSTM together can be best fit for this project because LSTM is similar to RNN, and the RNN algorithm is depending on the LSTM because its having the feedback connectivity and also LSTM process the entire sequence of data. . the main challenge of deep learning is when we deal with large data we need to go deeper that is analyzing the huge data need to done thoroughly, The structure of text descriptions should be relevant to the objects present in the image, and the relationship between the objects and it's descriptions need to be clarified, Our ultimate aim of the project is to train the dataset with the good result and with the high accuracy. Flicker dataset is utilized with the huge collection of photographs used for computer vision and image processing algorithms. So this voice based caption generator act as a eyes for the people don't have the ability to conceptualize the scene happen around themselves, they can roam anywhere without the support of anyone else.

2. LITERATURE REVIEW

In image process [1] used support vector machine and JEC to extracting the depth feature for an image by applying the Gaussian effect to get a better idea of a user given image. In [1] they used JEC for image feature extraction method. It will create a feature vector for annotation an image in various dimensions. It's nothing but processing the image with various model. After extracting the features from the JEC it applied to the SVM model to performing various operations like, rotating image in flat wise, axis wise and position wise manner to map the important features it contains keyword while annotating the image. It helpful for us to identify or recognize the object.

And in [2] CNN and RNN algorithm is used to performing the caption generation process by including the attention model to predict the proper words LSTM is more effective when compared to RNN that's why we have used long short term memory for our model . For recognizing the unusual objects they have used CNN after it will pass on the obtained information of the image is fed to the first step of long short term memory by using this it will predict the only starting word of the sentence correctly to overcome this they used the Guide long short term memory. In this method the guide carried out through the entire process, with the previously obtained word and it does not changed during the process

In [3] Based on the transfer learning approach to develop automated image captioning for user given image. Here they have using the VCG16 for encoding process. And then recurrent neural network to encode the input to produce constant dimensional vector for getting the proper description. They used various algorithm like VCG16, RESNET and inception model to compare the accuracy that are obtained between them to use more effective one. Next appropriate caption is created for a user given image.

In[4] based on the image detection, like how the face detection ,object detection in the self driving machines, it detects each and every object that where present in front of the cameras and predict the proper word about the object that is box,pen,bottles,it combined with the previous pretrained model also and added the new created unknown words to the previous dataset values.By using this methodology it will become some time consuming process and results in the irrelevant description creation.

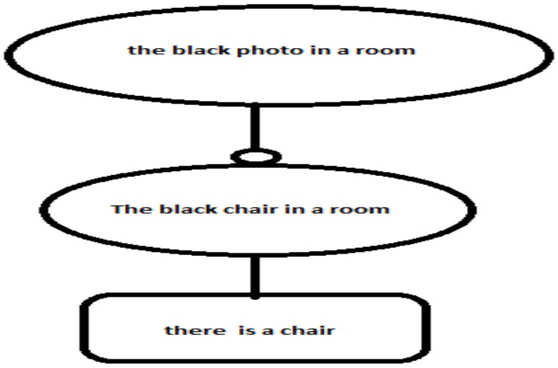


Fig 1.1 Image Detection Method Output

In[5] entire world is focusing on the image caption generation, but no one is interested in predicting the emotions and sentiments that present in the image, for that they created the model that defines both the description and emotion that included in the image .by using CNN to extract the feature and CAST to predict emotions.

3. PROPOSED SYSTEM

The Proposed methodology for voice based captions which is not only deals with internal images but also give a descriptions for external input images .once a description is created that text description will be read out as voice outputs then the audio is saved in the separate folder that contains all the audio files for the future references. For developing this model we have used convolutional neural network and long short term memory. Convolutional neural network for indentify the various features or objects that are present in the image. It will be helpful for the entire system predict the proper result then it will fed into the long short term memory to produce the sequence of words that properly describe about the image .

4. SYSTEM ARCHITECTURE

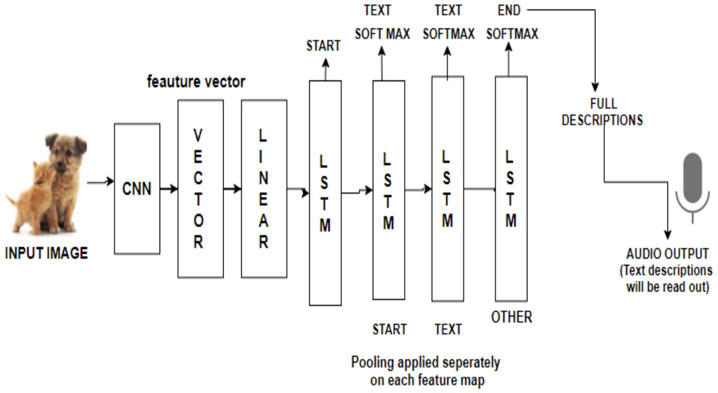


Fig 1.2 Architecture of Proposed Model

Input image taken from the user side convolutional neural network identify the objects that present in the image it extract the important features of an image and store those feature vector values, using pooling functions it will predict the features. Once the process completed it will move on to the long short term memory layer for the sequence sentence prediction based on the previous one, here softmax function is used to predict the output accurately and for overcoming the over fitting problem ,when we are working with the neural network most of the nodes having the output that are related to the previous one its results in overfitting, to avoid those problem softmax layer is used .If the output of this layer is between the range of zero to one ,if the range is greater are lesser it results in the error .and system will not predict the correct description for an image

$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}}$$

5. ALGORITHM

5.1. Convolutional Neural Network:

A convolutional or CNN is a class of deep neural network it is mostly used for analyzing visual images and classification, and also it is used in various field like image recognition , NLP and speech recognition. It has three layers namely, convolutional layer, pooling layer, and fully connected layer. The main advantage of using convolutional neural network is, it can identify the objects and faces present in the image. If a picture is indexed with cats and dogs, it will identify the key attributes and relationship between the two, and also behavioral patterns of the objects will be noted by CNN. It is truly efficient than the other algorithms because it has the ability to predict the image with highest accuracy.

5.2.Long Short Term Memory:

Long short term memory is a type of RNN, it is used for sequence prediction problems. The non-relevant information will be removed by using LSTM, and long short term memory have the efficient performance when compared to the RNN, it can be sustainable get the information with the long duration of time. It can be able to predict the information from the next data or previous data. The main challenge in LSTM is it will take more time to drain the data depending on the size of the dataset. CNN will be used for extracting information's from the image and LSTM will generate captions for the input image.

6. METHODOLOGY

a. Dataset Collection and Data Cleaning

b. Extracting Feature Vector

c. Loading dataset for Training the model

d. Tokenizing Vocabulary

e. Creation of Data Generator

DATASET COLLECTION AND DATA CLEANING

We are using flicker dataset, that contains images and descriptions that descriptions are in the form of dictionary with keys and values ,it's a easy way to map the description with input images . Every text dataset needs to be done with the data cleaning process. That involves clearing the symbols like special characters like asterisk, semicolon, colon, double quotes. Then the keywords starts with digits or ends with digits will be cleaned in this module. Compressing the long sentence which contains the inappropriate words.

EXTRACTING FEATURE VECTOR

Extracting features of an image with the help of exception , model and it is a pre-trained model, this model is utilized for the implementation process by using this model we can't do anything the trained model will do everything because its already trained by the large image net dataset it will classify the various difference in the image. It will take 299*299*3 as an input image and removing the end classification layers for getting the 2048 feature vector. It can accept any image format including PNG, JPG, and others. The neural network reduces large set of features extracts from the original input into smaller recurrent neural network-compatible feature vector. It is the major reason for calling CNN as 'Encoder'.

LOADING DATASET FOR TRAINING THE MODEL

In our flicker dataset folder, we have image file, it holds five thousand images for training and also it includes a function name called open_input, where it will print the image name in a string format. Next step is to be func_clean function, which will have captions in the form of dictionary. Also, the Begin and End keyword is added at the starting and ending point of the description. It will be given to long short term memory to forecast the caption. End keyword used to halt the looping process.
Training our model generally about to test the length of the training images, training the descriptions and it will include the features too. Best description is obtained only by the training the developed model it especially based on the epoch values what we have given during the process of training. It results in the forecast of accurate description for an input. During the process of training the loss value should be reduced in every iteration, how less the loss value will result in the good model.

TOKENIZING VOCABULARY

When language processing is used, it is required to tokenize the data, i.e. segregating a data like they've into "they" ,"have" ,compressing huge content into small readable unique content. Basically token means arranging the data into smaller blocks. In this module, keras library is used for tokenizing the text data and it can be saved into a separate file, which contains the index value of the text.

CREATION OF DATA GENERATOR

In this module it follows the supervised and un supervised learning model, having the internal image and their accurate output description is comes under the supervised learning ,if user is giving the external source it will show the output with the help of learning pattern from the trained data and it's an un supervised learning. when we using the data generator it first going through the CNN layer and performing some process like pooling ,next passes through the LSTM model it taken the output of CNN model and fit the first input with the second generated word with the help of dense. Comparing each pixel of an image long term short memory will forecast the suited description.

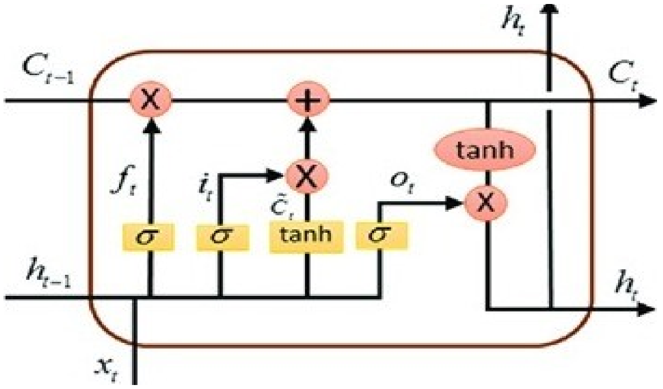


Fig 1.3 LSTM Structure

7. MODEL ARCHITECTURE

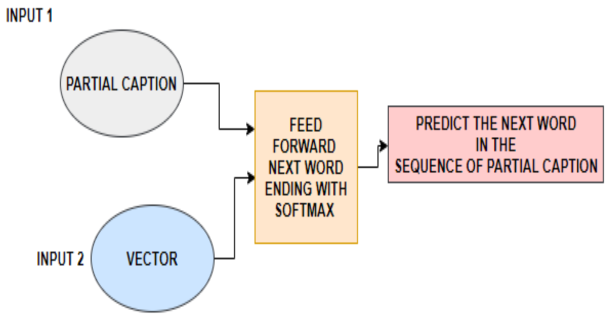
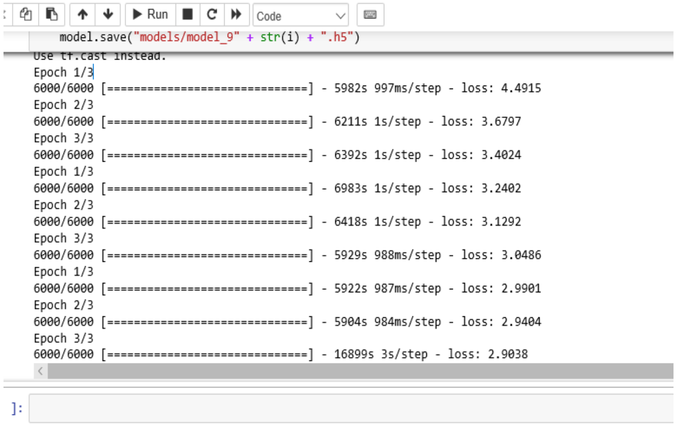

INPUT 1
PARTIAL CAPTION
FEED FORWARD NEXT WORD ENDING WITH SOFTMAX
PREDICT THE NEXT WORD IN THE SEQUENCE OF PARTIAL CAPTION
INPUT 2
VECTOR

Fig 1.4 RNN Flow

8. TIME TAKEN AND PARAMETER CALCULATION DURING TRAINING

We used 8GB RAM and jupyter notebook for training the model. And checked giving various epoch values like (2,3,8) it taken 8hours to complete entire training process. It consume more time or less time based on the system compatibility ,by using those epochs value itself the model produces a ninety percentage of description correctly

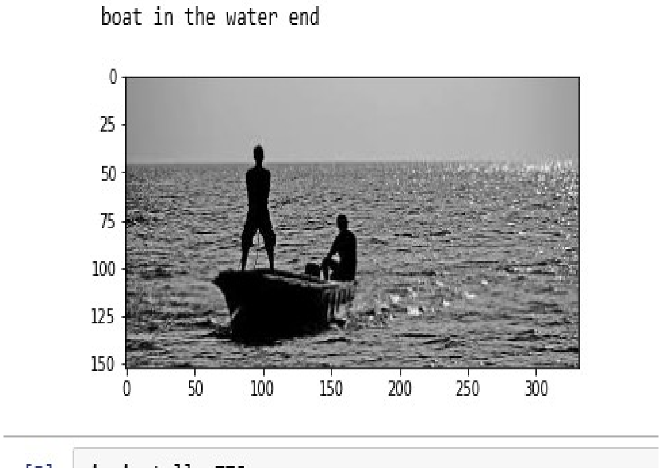


```
    model.save("models/model_9" + str(i) + ".h5")
Use tf.cast instead.
Epoch 1/3
6000/6000 [==============================] - 5982s 997ms/step - loss: 4.4915
Epoch 2/3
6000/6000 [==============================] - 6211s 1s/step - loss: 3.6797
Epoch 3/3
6000/6000 [==============================] - 6392s 1s/step - loss: 3.4024
Epoch 1/3
6000/6000 [==============================] - 6983s 1s/step - loss: 3.2402
Epoch 2/3
6000/6000 [==============================] - 6418s 1s/step - loss: 3.1292
Epoch 3/3
6000/6000 [==============================] - 5929s 988ms/step - loss: 3.0486
Epoch 1/3
6000/6000 [==============================] - 5922s 987ms/step - loss: 2.9901
Epoch 2/3
6000/6000 [==============================] - 5904s 984ms/step - loss: 2.9404
Epoch 3/3
6000/6000 [==============================] - 16899s 3s/step - loss: 2.9038
]:
```

Fig 1.5 During Training Process

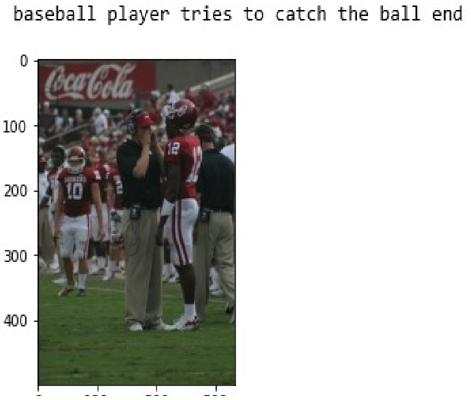
9. TESTING THE MODEL

First process is to uploading the image it may be from the dataset which we have gathered or else it may be user own image. After that step it enter into various module then it will print the related description for an user given input, once the captions is created then it will play the audio of an caption generated



```
boat in the water end
```

Fig 1.6 Description Generated



```
baseball player tries to catch the ball end
```

Fig 1.7 second output

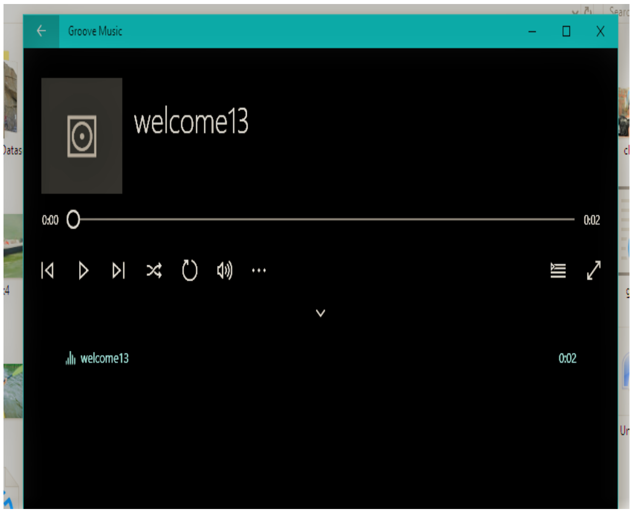


Fig 1.8 Voice Output

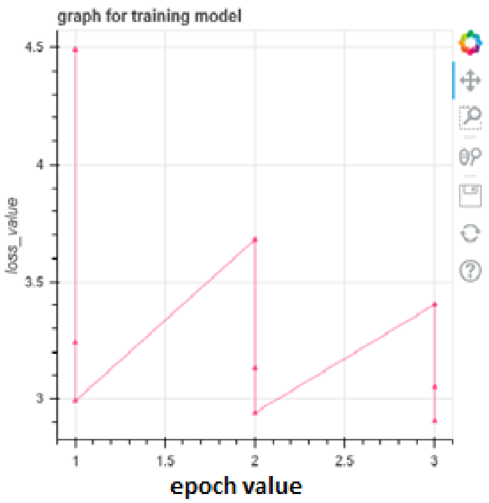


Fig 1.9 Graph for Epoch with Loss Value

10. CONCLUSION

Voice based image caption generator has been developed using a CNN-LSTM model. Main key aspects of our project to note, the proposed model not only depends on the dataset, the proposed model is trained for testing the user input, so that it can predict the descriptions from the external images. Out dataset consists of 8091 images. The proposed model is required to be trained on huge dataset that contains more than 10,000 images for achieving a better accuracy. This model is not applicable for the exact representation of an image that work finely for some sort of pictures.

REFERENCES

[1] Sumathi, T., Hemalatha, M, “A combined hierarchical model for automatic image annotation and retrieval.” In: International Conference on Advanced Computing (ICAC)- (2011).

[2] Dong-Jin Kim, Donggeun Yoo, Bonggeun Sim, In So Kweon, “Senetence Learning Deep convolutional neural Network for Image Caption Generation “, In : 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)-2016

[3] Varsha Kesavan, Vaidehi Muley,Megha kolhekar, “Deep Learning based Image Caption Generation” Global Conference for Advancement in Technology (GCAT)-2019

[4] Ren C. Luo, Yu-Ting, Hsu, Yu-Cheng, Wen, Huan-Jun, Ye, “Visual Image Caption Generation for Service Robotics and Industrial Application” IEEE-2019

[5] Yu, M.T., Sein, M.M.: “Automatic image captioning system using integration of N cut and color-based segmentation method”. In: Society of Instrument and Control Engineers Annual Conference(SCCEAC)- (2011).

[6] Ushiku, Y., Harada, T., Kuniyoshi, Y.: “Automatic sentence generation from images”. In: (ACM)Multimedia (2011)

[7] Federico, M., Furini, M.: “Enhancing learning accessibility through fully automatic captioning.” In: International Cross-Disciplinary Conference on Web Accessibility(ICDCWA)- (2011)

[8] Xi, S.M., Im Cho, Y.: “Image caption automatic generation method based on weighted feature”. In: International Conference on Control, Automation and Systems(ICCAS) (2013)

[9] Horiuchi, S., Moriguchi, H., Shengbo, X., Honiden, S.: “Automatic image description by using word-level features”. In: International Conference on Internet Multimedia Computing and Service(ICIMCS)- (2013)

[10] Ramnath, K., Vanderwende, L., El-Saban, M., Sinha, S.N., Kannan, A., Hassan, N., Galley, M.: “AutoCaption: automatic caption generation for personal photos”. In: IEEE Winter Conference on Applications of Computer Vision (2014)

[11] Sivakrishna Reddy, A., Monolisa, N., Nathiya, M., Anjugam, D.: “A combined hierarchical model for automatic image annotation and retrieval” . In: International Conference on Innovations in Information Embedded and Communication Systems(ICIIECS)- (2015)

[12] Shivdikar, K., Kak, A., Marwah, K.: “Automatic image annotation using a hybrid engine”. In: IEEE India Conference (2015)

[13] Mathews, A.: “Captioning images using different styles”. In: ACM Multimedia Conference (2015)

[14] Mathews, A., Xie, L., He, X.: “Choosing basic-level concept names using visual and language context”. In: IEEE Winter Conference on Applications of Computer Vision (2015)

[15] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: “Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models”. In: International Conference on Computer Vision(ICCV)- (2015)

[16] 14. Vijay, K., Ramya, D.: “Generation of caption selection for news images using stemming algorithm”. In: International Conference on Computation of Power, Energy, Information and Communication(ICCPEIC)- (2015)

[17] Shahaf, D., Horvitz, E., Mankof, R.: “Inside jokes: identifying humorous cartoon captions”. In: International Conference on Knowledge Discovery and Data Mining (ICKDDM-)(2015)

[18] Li, X., Lan, W., Dong, J., Liu, H.: Adding Chinese captions to images. In: International Conference in Multimedia Retrieval(ICMR)- (2016)

[19] Jin, J., Nakayama, H.: Annotation order matters: recurrent image annotator for arbitrary length image tagging. In: International Conference on Pattern Recognition(ICPR) (2016)

[20] Shi, Z., Zou, Z.: Can a machine generate humanlike language descriptions for a remote sensing image? IEEE Trans. Geosci. Remote Sens. 55(6), 3623–3634 (2016)

[21] Shetty, R., Tavakoli, H.R., Laaksonen, J.: “Exploiting scene context for image captioning”. In: Vision and Language Integration Meets Multimedia Fusion (2016)

[22] Li, X., Song, X., Herranz, L., Zhu, Y., Jiang, S.: “Image captioning with both object and scene information”. In: ACM Multimedia (2016)

[23] Wang, C., Yang, H., Bartz, C., Meinel, C.: “Image captioning with deep bidirectional LSTMs”. In: ACM Multimedia (2016)

[24] Liu, C., Wang, C., Sun, F., Rui, Y.: Image2Text: a multimodal caption generator. In: ACM Multimedia (2016)