WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Review on Automated Generation of Audio Captions from Images

[1]Ramesh Lavhe, [2] Gaurav Sukre,[2] Ashwin Tambe,[2] Aman Somwanshi, [2] Aniket Bhadale

[1]*Professor,* [2] *Students, Department of Information Technology, Anantrao Pawar College of Engineering and Research, Pune* , India

[1]ramesh.lavhe@abmspcoer.org, [2]gsukre@gmail.com, [3]ashwintambe213@gmail.com, [4]amansomwanshi12@gmail.com, [5]bhadaleaniket145@gmail.com

ABSTRACT

Visually impaired people face innumerable types of hindrances owing to their inability to visualize the natural environment. To overcome this problem, this proposed system will automatically generate captions after receiving the image as an input and convert the generated caption to an audio format so that visually impaired individuals can listen to the natural language captions. Caption generation is performed using Deep Learning algorithm Convolution Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory.

Keywords: *Deep Learning, RNN, LSTM, CNN.*

INTRODUCTION

Vision deficiency, widely known as protanopia, is nothing but a vision disorder that has no immediate cure of any kind.World Health Organisation (WHO) has reported 2.2 billion humans have some sort of visual impairment.It is more than strenuous to persevere in an environment without being able to use your vision substantially. There is a lot of room for developing technologies for visually impaired people. One approach is to scan the image and identify items that will produce a relevant caption in an audio format which will enable visually impaired users to syndicate all elements of a picture. Composing a caption for an image employs a series of tasks, such as in depth understanding of semantics, followed by employing the semantics in an understandable order of a sentence. Human beings communicate with the help of natural language, thus producing output that makes sense to humans is one more task. This paper aims to employ Deep Learning Neural Networks to detect and identify objects and produce plausible captions.

LITERATURE REVIEW

Generating natural language captions from an image is a reasonably modish idea without denying the substantial research done in this area. The following section will draw out an understanding of work done hitherto.

In this [1], the system is seen prevising captions automatically for a news article. Resulting caption is from the dataset of numerous news articles along with an image syndicated. This model consists of two phases of collection of material and surface realization. Selection of content defines what the image and the corresponding article are about, while surface knowledge determines how to convey the chosen content.

The following paper [2], propounds a model that effectuates captions for a news story with an embedded image in it. These captions are composed by employing frequency ranking calculation and the stemming algorithm.

The[3] substructure is designed to enable mobile device users to compose captions out of images. Users upload a photo that travels to a cloud system where several parallel modules, such as face recognition, scene recognition, GPS, date-time are used to identify a variety of entities and relationships. Module outputs are integrated to compose a huge multitude of captions.

Deep Visual-Semantic Alignments for prevising Image Descriptions model can automatically propound captions for proffered images using multimodal Recurring Neural Network [4]. Subjected model orients sentence snippets to the visual regions that they expound via multimodal embedding before using it as an input to multimodal Recurring Neural Network that will formulate snippets.

Here [5] set forth model is built on the theory 'Show and Tell'. This model had used Long Short Term Memory and Convolutional Neural Network to put forward a caption. This model has been executed by using several datasets to acquire more precise results.

The following proposed model [6], accepts input as an audio file and permutes it into text. The conversion is achieved by implementing an encoder-decoder system with a layer in between them. This tack was assessed enforcing a dataset of recordings, each of which was incorporated with a description (caption) within the dataset.

An Auto caption generating model has been implemented by enacting Recurring Neural Network and Long Short Term Memory Algorithm with additional Read-Only units [7]. This model is trained with the MSCOCO dataset which produces more faultless captions by synergizing all these technologies

III. METHODOLOGY USED

Image caption generation problem can be solved by using a conventional encoder-decoder Recurring Neural Network framework involving two elements,the encoder and the decoder. The encoder reads the input image and produces a fixed-length vector using an internal representation, and the decoder transforms the input into a text summary.

To sum up, the produced output is surmised to narrate in one sentence what is present in the image such as the objects present, their properties, identifying actions performed and interaction between objects, etc. A pre-trained convolutional neural network (CNN) algorithm is able to encode the images and RNN, such as a Long Short Term Memory network is used to encode the text sequence and formualate the next word in the series. Technical approaches which are used to build a model mentioned below are in detail

A. CONVOLUTIONNEURAL NETWORK

A Convolution Neural Network is a deep learning algorithm capable of taking input in the form an image, assigning the learnable weights and angles to several objects in an image and then distinguish objects. This algorithm can work with substantially less preprocessing. The framework of this network is similar to that of the design of neuron conveyance in the human brain.

Individually, neurons can answer to incitation only in the confined area of the perceptible field known as the Receptive Field. The series of these fields is then imbricated to fill the total visual region. An accumulation of these fields later overlaps to mantle the whole visual area. Also,there is an "observation window" from which the model adapts the weights of the filter. The weights are ssplit for every convolution ensuring translational equivariance by authorizing the model to identify objects in the image.

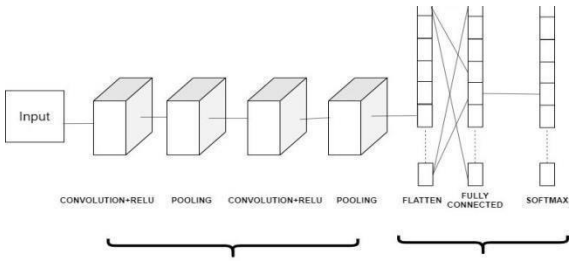

Fig. 1 Progression of transformations involved in the convolutional neural network [11].

B. RECURRENT NEURAL NETWORK

The recurrent neural network framework progresses in being capable to apprehend information about past states to inform predictions through its memory cell status [1]. Conventionally neural networks don not require inputs and outputs to be interdependent. The issue with feed-forward neural networks is that they need a set-sized input and give a fixed-sized output. They do not capture sequences or information about time series nor do the account for memory.Thus, RNN countered this issue. It takes variable-sized input and delivers variable-sized output as it works competently with time series data as well. One of the most important features of RNN is the concealed state which grasps some information about a sequence. RNN sports a 'memory' which enables it to remember all the information about what is calculated. For each input, it uses the similar parameters as it executes the same task on all the inputs to generate the output. This brings down complexity of parameters

The intuitive representation of RNN is shown in Fig.

2. The basic formula of RNN is the recursive formula which is as follows: $St=F_w (S_t -1, X_t)$ where St is the new state at time t, F_w is the recursive function, St is the state at time t-1 and X_t is the input at time t.

To deal with the vanishing gradient problem faced by RNN, Long Short-Term Memory (LSTM) was introduced.

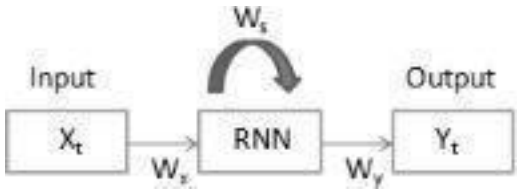

Fig. 2 Representation of RNN

If there is a deeper network with single input layer, more multiple hidden layer and single output layer then RNN basically executes the following actions:

1. The transformation of independent activation into dependent activation is executed by providing the same weights and biases to all the layers. It helps to reduce the difficulty of more parameters and memorize every previous output by giving output as input to the next hidden layer.

2.Thus all the hidden layers can be combined such that the weights and biases of all the hidden layers are the same into a single recurrent layer. Fig. 3 shows a simplified version of RNN model.

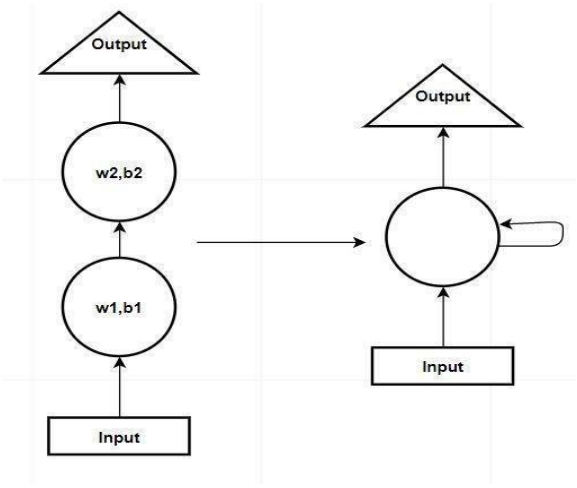

Fig. 3 Simplified form of RNN model

DATASET

Flicker 30k dataset is used to train this model. This dataset houses 31000 images 15586 captions. The images have JPEG format and different shapes as well as size. 18600 images will be used for training the model,6200 for validation and rest for testing. In that dataset, Fliker30k.token.txt file contains 5 captions for each image i.e., 155000 captions.

TEXT TO SPEECH CONVERSION

For converting generated captions into audio Google text to speech has been API has been used. This API enables us to convert text into audio in a wide range of languages. This API can be initiated in the outset and has to be installed on cloud.

IMPLEMENTATION

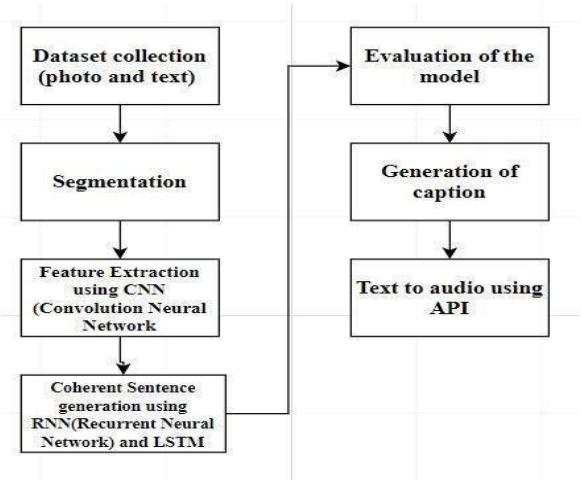

Fig 4. Block Diagram of the System

The implementation is initiated by importing all the necessary packages. The implementation of the model as shown in Fig. 4 is as follows.

1.*Importing and performing data cleaning*: The Flickr30k text dataset is used for training. This document is read to read the contents as string. A description dictionary is c that maps with the images in the Flickr30k dataset with a dataset of captions is set. Then data wrangling is conducted on all the descriptions which involve removing redundant characters,numbers and converting all text to lowercase. Later, all the distinct words are separated and the vocabulary is formulated from all the descriptions. Finally, list of preprocessed descriptions is stored in a file.

2.*Acquiring the feature vector from image dataset*: Using conventional transfer learning where pretrained model is used to derive features. Here VGG16

imaging model is employed for the same.

3.*Loading dataset for Training*: The module has various functions:

a. Load a text file having a list of 31000 image names that are applied for string training and return an array of image names.

b. Create a dictionary with the captions for each image from the dataset of images. For each caption, the<start> and <end> Identifiers are appended. Necessited for the Long Short Term Memory model to demarcate the start and end of the caption.

c.Use the maximum length (max_length) function to determine model structure parameters

Here the maximum length is 34.

4.*Tokenizing the vocabulary and creating data generator*: Every word of the vocabulary with distinct index value and saved to a pickle file. The model is trained on 31000 images. Summation of data for 31000 images cannot be stored in memory such that a generator process is used to produce batches.

TABLE 1.
INPUT & OUTPUT EXAMPLE TO THE MODEL

X1(feature vector)	*X2(text sequence)*	*Y(word to predict)*
feature	start,	two
feature	start, two	dogs
feature	start, two, dogs	drink
feature	start, two, dogs, drink	water
feature	start, two, dogs, drink, water	end

Table 1 presents one example where the input to is [x1, x2] and the output is y, here x1 = 2048 feature vector of that image, x2 is input text sequence though y is the output text sequence the model will predict.

5.*Defining the CNN-RNN model*:

The structure of the model consists of three major parts:

1.Feature Extractor – The element obtained from the image has a dimension of 2048 with a broad layer, which limits the size to 256 nodes.

2.Sequence Processor – Embedding layer controls the text input, before the LSTM layer.

3.Decoder –By merging the output from the two layers over, we're going to process the broader layer to compose the final estimate. The last layer shall contain the number of nodes identical to the scale of the vocabulary.

6. *Training the model and evaluation*:

To train the model, 3000 training images have been utilized by propounding the input and output cycles in batches and embedding them to the model using model.fit_ generator() function. The BLEU metric is applied for evaluating the functioning of the caption generator.

CONCLUSION

The proffered system generates natural language captions for photos taken using a web application in any mobile device. This system is specifically modeled for assisting visually impaired people in understanding what is occurring around them. This model executes by sending an image to a server where all the processing is done by using deep learning algorithms like CNN, RNN, and LSTM. The user will receive an audible output from this application and provide a better understanding of their surroundings. This model will be further improved to increment the precision. This shall comprise the feature of converting continuous captions into the video and many more. Other potential attributes include the ability to use user feedback to improvise the captions.

REFERENCES

1. Saha Sumit, (2015, Dec). A Comprehensive Guide to Convolutional Neural Networks theELI5 way.towards data science.Retrieved From Https://towards data science.com/
2. Xu, N., Liu, A.-A., Wong, Y., Zhang, Y., Nie, Y., & Kankanhalli, M. (2018). Dual-Stream Recurrent Neural Network for Video Captioning. IEEE Transactions on Circuits and Systems for Video Technology, 1–1.
3. Galvez, R. L., Bandala, A. A., Dadios, E. P., Vicerra, R. R. P., & Maningo, J. M. Z. (2018). Object Detection Using Convolutional Neural Networks. TENCON 2018 - 2018 IEEE Region 10 Conference.
4. Vishwash Batra, Yulan He, Neural Caption Generation for News Images (2018) George Vogiatzis School of Engineering and Applied Science, Aston University
5. Poghosyan, A., & Sarukhanyan, H. (2017). Short Term memory with read-only unit in neural image caption generator. 2017 Computer Science and Information Technologies (CSIT).
6. Drossos, K., Adavanne, S., & Virtanen, T. (2017). Automated audio captioning with recurrent neural networks. 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

7. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
8. Karpathy, A., & Fei-Fei, L. (2017). Deep Visual-Semantic Alignments for Generating Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4)664–676.
9. El-Saban, M., Sinha, S. N., Kannan, A., Torresani, L. (2014). AutoCaption: Automatic caption generation for personal photos. IEEE Winter Conference on Applications of Computer Vision.
10. Vijay, K., & Ramya, D. (2015). Generation of caption selection for news images using a stemming algorithm. 2015 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC).
11. Feng, Y., & Lapata, M. (2013). Automatic Caption Generation for News Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(4), 797– 812.