International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 07 | July - 2021 ISSN: 2582-3930

Image to Text and Speech Converter

[1]Rahul Goyal , [2]Ajay, [3]Ketan Rana, [4]Sanjeev
Ms. Deepika Tyagi

[1,2,3,4]Student, Department of Information Technology, Dr. Akhilesh Das Gupta Institute of Technology & Management, Delhi, India.
Assistant Professor, Department of Information Technology, Dr. Akhilesh Das Gupta Institute of Technology & Management, Delhi, India.

Abstract

Processing an image is one of the most growing technology which enhances raw pictures received from various gadgets in normal day-to-day life for numerous applications. The purpose of this paper is to Convert an image into Text as well as Speech using the machine learning algorithms and python tools. An image-to-text model extracts text descriptions in natural language based on understanding of image using Convolutional Neural Network and Recurrent Neural Network. A text-to-speech module converts natural language to speech synthesis. As soon as the image is uploaded the text description and audio of the image is generated simultaneously. This system can be helpful to visually impaired as well as people with learning disabilities to understand the scenario from the images. The result shows that the accuracy of text generated is 75%.

Keywords : Machine Learning, Recurrent Neural Network, Convolutional Neural Network

Introduction

With the speedy development of medical care, there are a large quantity of pictures, attended with tons of connected texts . Automatic image text and sound has recently attracted a lot of analysis interest. The target of this is to come up with properly fashioned English sentences to explain the content of a picture mechanically

and also changing that text to audio, that is of nice impact in varied domains like virtual assistants, image compartmentalization, recommendation in written material applications, and therefore the facilitate of the disabled . Though it's a straightforward task for somebody's to explain a picture, it becomes terribly tough for a machine to perform such a task. So to make this task easy machine learning is used which perform tasks automatically and improve using the data and experience.

Image to text conjointly referred to as Image Captioning doesn't solely compelled to notice the objects contained in a picture but conjointly shows how these objects are associated with one another and their attributes . This is achieved by applying Recurrent Neural Network model, a machine learning algorithm in between the partial caption and the image vector and to convert the image in the form of vector Convolutional Neural Network is used . After generating the text of that image it is converted into audio format using gtts(Google Text To Speech) which is a Python library tool that takes text as an input and provide an audio format of it.

METHODOLOGY

1. Data Collection and Cleaning

In this model the Flickr 8k dataset is used . Training Set — 6000 images,Test Set — 2000 images.

While modifying text, some basic improvement like lower-casing is performed on all the words removing special tokens (like '%', '$', '#', etc.), eliminating words that contain numbers (like 'hey199', etc.). A vocabulary of all the different words available across all the 8000*5 (i.e. 40000) image captions (corpus) in the data set is created.

This means model has 8763 unique words across all the 40000 image captions. However, most of these words will occur very few times, say 1, 2 or 3 times. While creating a predictive model, it would not be good to have all the words present in the vocabulary but the words which are more likely to occur or which are common. Hence consider only those words which occur at least 10 times in the entire corpus. So now model has only 1651 unique words in vocabulary. However, model will append 0's (zero padding explained later) and thus total words = 1651+1 = 1652 (one index for the 0).

2. Loading the training set

The document "Flickr_8k.trainImages.txt" contains the names of the pictures that belong to the training set. therefore ,these names are loaded and listed into a listing "train".

'startseq' -> this can be a begin sequence token which can be additional at the beginning of each caption.

'endseq' -> this can be associate finish sequence token which can be additional at the top of each caption.

3 Data Preprocessing — Images

First it is needed to convert each image into a set sized vector which may then be fed as input to the neural network. For this purpose, the project take transfer learning by exploitation the InceptionV3 model (Convolutional Neural Network) created by Google analysis.

This model was trained on Imagenet dataset to perform image classification on one thousand totally different categories of pictures. However, this report purpose here isn't to classify the image however simply get fixed-length informative vector for every image. This method is termed automatic feature engineering. Hence, to simply this take away the last softmax layer from the model and extract a 2048 length vector .

4.Data Preprocessing — Captions

Captions are something that this model want to predict. But the entire caption of the given image is not predicted at once . It will predict the caption word by word. Create two Python Dictionaries namely "wordtoix" which means word to index and "ixtoword " which means index to word .Stating simply, it will represent every unique word in the vocabulary by an integer (index).As seen above, model has 1652 unique words in the corpus and thus each word will be represented by an integer index between 1 to 1652.

To do data processing , encipher every word into a set sized vector. These two Python dictionaries are often used as follows: wordtoix['text'] -> index of the word 'text' is returned.

5. Data Preparation using Generator Function

First convert the pictures to their corresponding 2048 length feature vector as mentioned on top of. Let "Image_1" and "Image_2" be the feature vectors of the primary two pictures severally.

Secondly, build the vocabulary for the primary two (train) captions by adding the two tokens "startseq" and "endseq" in each of them: Generate One word at a time using iteration. This is stopped once either of the below two condition is met :

• Encountered an associate 'endseq' token which implies the model thinks that this can be the top of the caption. (You ought to currently perceive the importance of the 'endseq' token)

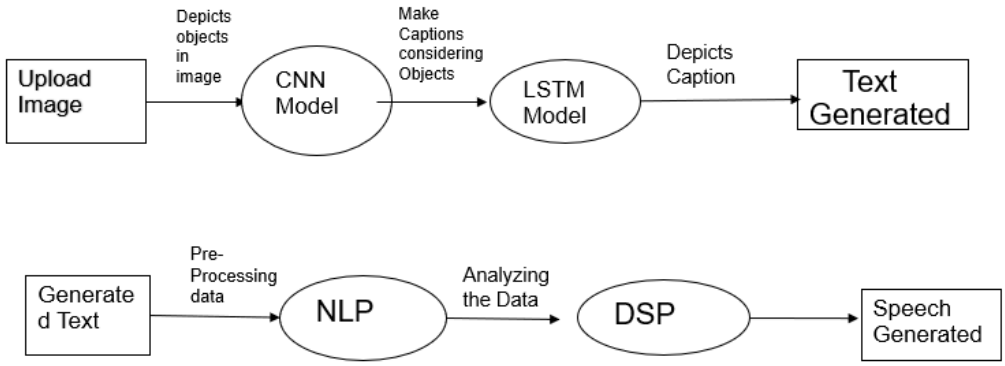
• Reached a most threshold of the quantity of words generated by the model.

If any of the above conditions is met, break the loop and report the generated caption as the output of the model for the given image.

6. Text To Audio

There square measure many Api's obtainable to convert text to speech in Python. One amongst such Api's is that the Google Text to Speech API unremarkably referred to as the gTTS API. gTTS could be a terribly

simple to use tool that converts the text entered, into audio which may be saved as a mp3 file. The gTTS API supports many languages together with English, Hindi, Tamil, French, German and lots of a lot of. The speech are often delivered in anyone of the two obtainable audio speeds, quick or slow. However, as of the newest update, it's impracticable to vary the voice of the generated audio.



System Architecture of Text well As Speech Conversion

RESULTS AND DISCUSSIONS

Based on the experimental analysis that we performed we found out that the proposed method can accurately detect regions from images which have different text sizes, styles and color and convert the extracted regions into text based on our training model. Moreover, the text to audio method also works successfully and audio is played as soon as the text is extracted.

Below figure shows the accuracy of the model done on 1600 training samples and 400 testing samples:-

```
Train on 1600 samples, validate on 400 samples
Epoch 1/10
1600/1600 [==============================] - 1s 410us/step - loss: 0.0741 - acc: 0.9856 - val_loss: 0.7261 - val_acc: 0.7825
Epoch 2/10
1600/1600 [==============================] - ETA: 0s - loss: 0.0631 - acc: 0.984 - 1s 453us/step - loss: 0.0610 - acc: 0.9856
val_loss: 0.8409 - val_acc: 0.7800
Epoch 3/10
1600/1600 [==============================] - 1s 418us/step - loss: 0.0636 - acc: 0.9888 - val_loss: 0.6447 - val_acc: 0.8025
Epoch 4/10
1600/1600 [==============================] - 1s 391us/step - loss: 0.0616 - acc: 0.9888 - val_loss: 0.9651 - val_acc: 0.7900
Epoch 5/10
1600/1600 [==============================] - 1s 386us/step - loss: 0.0493 - acc: 0.9925 - val_loss: 0.7148 - val_acc: 0.8125
Epoch 6/10
1600/1600 [==============================] - 1s 388us/step - loss: 0.0382 - acc: 0.9956 - val_loss: 1.0710 - val_acc: 0.7825
Epoch 7/10
1600/1600 [==============================] - 1s 411us/step - loss: 0.0425 - acc: 0.9944 - val_loss: 0.9285 - val_acc: 0.7875
Epoch 8/10
1600/1600 [==============================] - 1s 411us/step - loss: 0.0417 - acc: 0.9944 - val_loss: 0.7851 - val_acc: 0.7825
Epoch 9/10
1600/1600 [==============================] - 1s 390us/step - loss: 0.0924 - acc: 0.9800 - val_loss: 1.0916 - val_acc: 0.7775
Epoch 10/10
1600/1600 [==============================] - 1s 388us/step - loss: 0.0575 - acc: 0.9925 - val_loss: 0.9009 - val_acc: 0.7925
```

CONCLUSION

In this way we presented the techniques and implemented image-to-text as well as speech system. We used Convolutional Neural Network, Recurrent Neural Network, and sentence generation to understand the image to text method deeply .Then we generated audio of that text generated. Our contribution towards this work will surely be helpful for blind as well as physically disabled people of our society. This will help such people to make them more interact able with real world. Main focus is on identification of object in an image which will help in identifying important objects from an image. This report also contains an abstract view of various technique proposed in recent past year for image to text conversion and text to speech conversion.

But there are some drawbacks of this model also as it does not work properly for blur images or if the image region is small.

FUTURE WORK:

Although this approach overcomes most of the challenges faced by other algorithms, it still suffers to work on images where the image regions are very small or if the image regions are blur. In such case the accuracy is pretty low and text generated is not accurate. So, to avoid such situation one thing that could be done is that image quality can be improved before putting it in the model so that accurate text is generated. Moreover, audio is also in one language so different language can also be provided which can help that people who do not know this language.

REFERENCES

[1]Ramesh M. Kagalkar, "Conversion Of Image To Text As Well As Speech Using Edge Detection And Image Segmentation", 11,2014, International Journal of Science and Research (IJSR).

[2]H Rithika; B. Nithya Santhoshi, "Image text to speech conversion in the desired language " , 2016, IEEE.

[3] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares, "Image Captioning: Transforming Objects into Words",2019,NIPS.

[4] Shuang Liu, Liang Bai, Yanli Hu, "Image Captioning Based on Deep Neural Networks", 2018,MATEC web conference.

[5] Lakshminarasimhan Srinivasan1, Dinesh Sreekanthan2, Amutha A.L3, "Image Captioning - A Deep Learning Approach", 2018, International Journal of Applied Engineering Research.