International Journal of Research in Engineering, Science and Management
Volume 4, Issue 5, May 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Image Recognition and Voice Translation for Visually Impaired

Sandeep Pasupuleti[1*], Lahari Dadi[2], Manikumar Gadi[3], R. Krishnaveni[4]
[1,2,3,4]*Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Chennai, India*

***Abstract*: Image recognition and voice translation for the visually impaired is principally centered on aiding the visually impaired individuals which can guide them through changing the image into voice. Image captioning is that generating the caption of the given image. The arrival of image recognition and voice translation system can facilitate visually impaired individuals to visualize the globe. Image recognition needs to understand the numerous things, their properties, and their connections during an image. The "Flickr_8k" dataset is employed for this project with having 8000 pictures. During this project, we have a tendency to principally specialize in the eye mechanism, that extracts the options through VGG16 and converts the options into language through LSTM. During this project, we've projected a changed LSTM which can enhance the addition of data that have a tendency to get from a picture as further input to each unit of the block, which enhances the higher results on the recognition and translation. The sentence that's generated from the image is evaluated exploitation blue cheese Score. Moreover, the benefits and therefore the short comings of the prevailing ways and projected technique area unit mentioned and therefore the analysis criteria has been non commissioned. Hence, we have a tendency to area unit developing this model that generates a caption for the given image and that we show our model advances the state of art on tasks that need the joint process of Image and sentence description and finally converting it into voice.**

***Keywords*: image recognition, voice translation, Flickr_8k, VGG, LSTM.**

1. Introduction

Image recognition is one of the sub-method that generates the matter description of a picture. The appearance of an image captioning system can facilitate the visually impaired folks to ascertain the planet. Image captioning needs to ascertain the noteworthy articles, their properties, and their associations in an exceeding image. It equally must build syntactically and semantically right sentences. Profound learning-based systems area unit stocked with for managing the complexities and challenges of inscribing image. Image captioning could be an important piece of scene understanding, which mixes the information of pc vision and language process. what is more, the benefits and therefore the shortcomings of those ways area unit mentioned, providing the ordinarily used data set (Flickr 8k) and analysis criteria during this field? There area unit numerous kinds of ways employed in image captioning they're Encoder -decoder design, linguistics-based mostly, Novel object-based mostly. This project focuses on victimization attention-based image captioning.

Fig. 1. What do you see in the below picture?

Well, a number of you would possibly say "A white dog in an exceedingly grassy area", some might say "White dog with brown spots" and nevertheless some others would possibly say "A dog on grass and a few pink flowers", however, will machine say of these.

2. Motivation

The main motivation of this project is to develop an image recognition and voice translation system for the visually impaired folks which can facilitate them happening the streets while not the assistance of the other person. We are able to do that by 1st dynamical over the scene into content and afterward the content to voice. Each desires the image captioning method. Moreover, we have a tendency to principally focus on the scene to text and afterwards to voice. This project may also be employed in numerous fashionable area's issues for an easy answer like biomedicine, trade, the military, instruction, processed libraries, and internet wanting. Net-based mostly on life stages, as an example, Facebook and Twitter will licitly produce descriptions from photos. It may also be employed in the synthetic intelligence sector for knowledge intake. During this form of image captioning system, numerous challenges have been encountered in recent years like less accuracy and big demand for memory. This may enhance the system creating high accuracy and alternative numerous challenges are overcome by mistreatment deep neural network.

*Corresponding author: sandeeppasupuleti29@gmail.com

3. Literature Survey

This paper chiefly focuses on the novel attention-based model for automatic image captioning and projected by Marco Pedersoli. This system model the conditions between image districts, subtitle words, and also the condition of the associate RNN language model, utilizing 3 combine wise collaborations. Throughout coaching these affiliations area unit deduced from the image level subtitles. These affiliations facilitate boost subtitling throughout testing. For the bigger part of their trials, they utilize the MSCOCO dataset. It contains around 80K making ready photos and 40K improvement photos. Every image accompanies 5 mesmerizing subtitles.

This paper focuses on the automatic generation of language descriptions of a picture. Quan Zeng, You projected this paper. Existing methodologies are either top-down, which begins from the significance of the image and converts it into words or a base up. Microsoft COCO palm and Flickr30K datasets are used. CNN and therefore, the RNN deep learning techniques are utilized in this paper. The analysis metrics are METEOR, ROUGH-L, CIDEr. This model offers an accuracy of seventy-eight by checking through the human analysis.

Chetan Amritkar, Vaishali Jabade developed this project, and it presents work toward image captioning strategies pictures. The approaches use text-to-natural language generation techniques victimization AI and information processing that generates image captions by adapting text from captions of visually similar pictures. Pascal's knowledge set is employed containing data retrieved from Flickr30k, MSCOCO dataset that contains news articles, image captions, and pictures taken from the BBC News website. They followed Domain-specific image captioning that learns from weakly-supervised samples of antecedently captioned that is specifying a class into a specific or distinctive attribute. The disadvantage of this technique is that they performed on the untagged knowledge and provides the eightieth accuracy.

MD. MD. Zakir Hossain, Ferdous Sohel, Hamid Laga developed this project, and it presents work toward image captioning ways pictures. The approaches use text-to-text tongue generation techniques, that generate image captions by adapting a text from captions of visually similar pictures. Pascal's knowledge set is employed containing data retrieved from BBC data set that contains news articles, image captions, and pictures taken from the BBC News web site. UNT Data set is additionally used that consists of pictures and co-occurring text from web pages. They followed a domain-specific image captioning that learns from weakly-supervised samples of antecedently captioned that is specifying class into a specific or distinctive attribute. The disadvantage of this technique is that they performed on the untagged knowledge and provides sixty-two percent (62%) accuracy.

Shuang Bai, Tai developed the strategy for the synthetic image description with the CNN and LSTM models being in them. By mistreatment dense captioning mechanism, they did the project. The datasets used are Flickr30k and MSCOCO that have thousands of images and are employed for the info pre-processing and coaching the model with seventy-eight (78%) accuracy.

4. Methodology

A. Neural Network Techniques

Neural Network:

Neural networks are the set of procedures or algorithms, modeled that are defined to identify or recognize the patterns. Neural network act as similar to the human brain, that it collects and classifies information according to a specific algorithm or procedure or architecture. The network bears a high resemblance to methods such as curve cutting and regression analysis.

In our project we are using two types of neural networks.

1. Recurrent Neural Network
2. Convolutional Neural Network

1) Recurrent Neural Network

Before we get into the details of what recurrent neural network, let's see if we need a network especially for dealing with sequences in information. Also, we will see how the recurrent network will be used in image captioning. While we are dealing with RNNs, it has a great ability to deal with various inputs and outputs.

In image captioning, we have an image as an input and we get output as a textual description of an image, for generating the sequence of words with varying lengths as output we need an RNN.

A Recurrent Neural Network (RNN) is a kind of neural network where the output from the past is given as a contribution to the present. In customary neural systems, all the information sources and outputs are autonomous of one another, however, in cases like when it is required to anticipate the expression of a sentence, the past words are required and subsequently, there is a need to memorize the past words. Along these lines, RNN appeared, which settled this issue with the help of Hidden Layer. The fundamental and most significant element of RNN is the hidden state, which recalls some data about an arrangement.

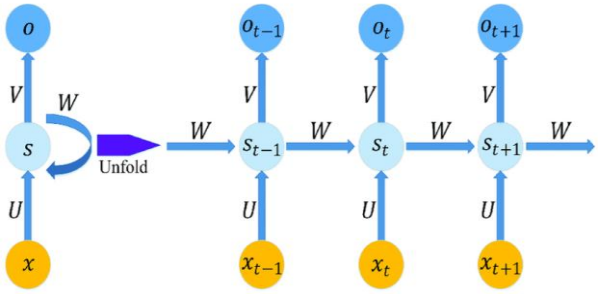

Fig. 2. Recurrent Neural Network

U, V, W are the weights of the hidden layer, an output layer, and therefore the hidden state.

X, t is the O, t is the input vector, and output result at time t.

RNN converts them into dependent activation by giving the same weights and biases to all layers, thus reducing the complexity of parameters and giving each output as input to the next hidden layer.

RNN works upon the fact that the result of information is dependent on its previous time stamps or previous t time in

which as we saw in the figure the t-1, t, t+1 are the timestamps that may produce sequence words based on the previous time stamps

As we saw in the figure the U, V, W are the hidden layers, weights of the network of which the input X passes through and we will get the output as OAs we seen in the figure the input of t-1 is passed through the network and then the output at t-1 is combined with input at t which may be passed through the network and got output at t in a similar way the neural network by depending on previous output.

2) Convolutional Neural Network

Convolutional Neural Networks (CNNs) are most famous for their performance in image processing, especially image-classification and image-detection tasks. A CNN just associates a specific number of hubs to one of its succeeding hubs. This is like applying a channel or a convolution to a picture; along these lines the name CNN. This methodology spares a great deal of computational time, from one perspective, then again, the nearby setting turns out to be progressively significant in contrast with an FFNN, where the entire information is considered without a moment's delay. Between the convolutional layers, a CNN regularly has pooling layers, diminishing the size of the picture portrayal. After the last convolutional layer, completely connected layers are frequently included. Fig. 3 shows schematically, how a CNN for picture characterization can resemble. The information is classified by passing through different convolutional, pooling, and completely connected layers.

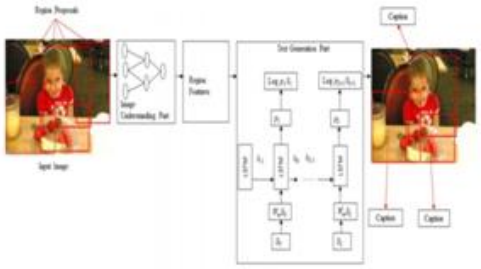

Fig. 3. CNN (Convolutional Neural Network)

Convolution is performed on the input data with the use of a filter or kernel to then produce a feature map that contains the features of an image.

This pooling layer is obtained by the operator to aggregate information within every short region of the features of input channels and then down-sampling the results

As we see in the figure 3 of a deer image has been sent as an input to the network in which it undergoes the process of convolutions and pooling and we got an extracted feature of an object in which we check the precision of the object from that which animal has high similarity have high precision score we conclude that is an object.

The main objectives of the project are divided into three modules.

- To perform object classification using VGG, that is first we need to detect the object from the image and second, we classify the object.
- Generation of caption from the classified object that is the output of VGG is taken as input in the LSTM and convert the features into caption in natural language.

Final step is to convert the text sentence into voice.

5. The Proposed Image Captioning System

In the proposed system, we are developing the image captioning using VGG16 and Modified LSTM. To consume less time while loading the vocabulary we use the GLSTM (GUIDED)model which decreases the loading time. Also, speech synthesis techniques are added to the captioned image which can generate a voice for the caption. In our proposed model we are using the Flick8k dataset which is memory efficient.

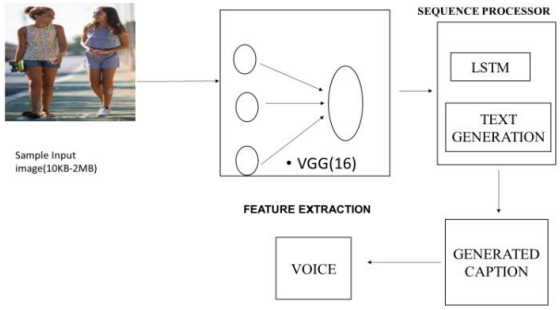

Fig. 4. Architecture of image captioning

The sample image taken in the above architecture is of size 346kb with a resolution of 640*840.There can be one input image at a time but there is room for improving it to a number of images and that will be future work. Training of the data is done in a way that the features of girl will be trained like a girl has long hair and etc. and for the testing, we will check by giving the object featured as a girl.

1) Image Feature Extraction

The Image feature extraction has done through 16-layer VGG. VGG is a convolutional neural network model for image recognition proposed by the Visual Geometry Group at the University of Oxford, where VGG19 refers to a VGG model with 19 weight layers, and VGG16 refers to a VGG model with 16 weight layers. The figure 4 illustrates the architecture of VGG16, the input layer takes an image in the size of(224 x224 x3), and the output layer is a SoftMax prediction on 1000 classes. From the input layer to the last max-pooling layer (labeled by 7x7 x 512) is taken into account the feature extraction a neighborhood of the model, while the rest of the network is referred as the classification part of the model.

2) Preprocessing Data Set

The pre-processing of the data set is the important step in every data mining or data-related project which enhances the medium of cleaning the datasets and transforming the datasets to the required algorithm. This pre-processing may fulfill any data missing or data repetition that happened, this pre-processing step can enhance the efficiency of data results which may increase the success rate of a project, by cleaning the noise data and replacing the missing attribute values.

In the pre-processing part after acquiring the data set and

importing all the crucial libraries and the pivotal part is identified and correctly handle the missing values will increase the accuracy which is training the system that how an object looks like in order to identify it quickly and finally the feature extraction.

3) Visual Geometry Group

VGG stands for visual geometry group. VGG structures make the express assumption that the wellsprings of information are pictures, which grants us to encode certain properties into the design. It uses the VGG model which evaluates the features of the picture in 16 layers. VGG is moreover computationally capable. It uses remarkable convolution and pooling exercises and performs parameter sharing. It is viewed as one of the incredible vision model engineering's to date. The most one-of-a-kind thing about VGG16 is that as opposed to having countless hyper-parameters they concentrated on having convolution layers of 3x3 channel with a step 1 and consistently utilized the same cushioning and max pool layer of 2x2 channel of step 2. It follows this plan of convolution and max pool layers reliably during the whole engineering. At long last, it has 2 FC (fully associated layers) trailed by a SoftMax for yield. The 16 in VGG16 alludes that it has 16 layers that have loads. This system may be a really enormous system and it's around 138 million (approx.) parameters.

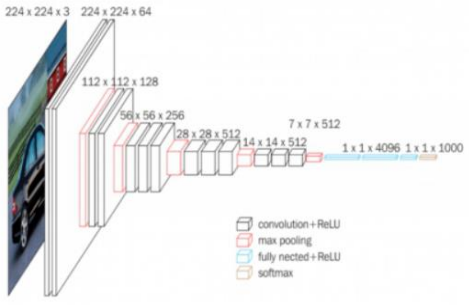

Fig. 5. Visual Geometry Group

In convolutional layer, it takes a volume of size w * h* d requires four parameters:

No of filters – K
Spatial Extent – Fw, Fh(Filter-width, Filter-height)
Stride - SW, Sh(Stride width, Stride height)
Padding -P
The formula to calculate receptive field
Output Width = (W – Fw + 2P/Sw) + 1 ----------(1)
Output Height = (H – Fh +2P/Sh) +1 ----------(2)
The formula to calculate pooling layer
OM = (IM +2P – F/S) + 1 ----------(3)
OM – Output Matrix
IM – Input Matrix
P – Padding
F – Filter
S – Stride

By applying the above receptive and pooling formulas, the convolutions, pooling, and feature maps outputs are derived

4) Text generation using long short-term memory

LSTM is an irregular neural framework and it is liable for delivering captions. LSTM was made from RNN with the mean to work with successive data. It is convincing in holding long stretch conditions through the memory cells. This is conduct required in complex issue areas like machine interpretation, discourse acknowledgment, and more. LSTMs are an unpredictable region of profound learning. It tends to be difficult to get your hands around what LSTMs are, and how terms like bidirectional and arrangement to-grouping identify with the field. It handled the issue of long-haul conditions of RNN in which the RNN can't anticipate the word put away in the drawn-out memory yet can give increasingly precise expectations from the ongoing data. As the whole doesn't give efficient execution. LSTM can as a matter of course holds the data for an extensive stretch of time. It is utilized for handling, foreseeing, and characterizing based on time arrangement information.

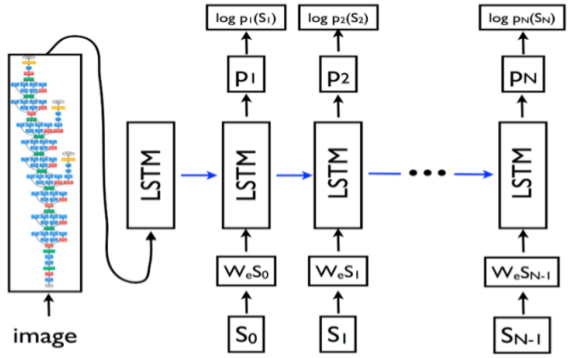

Fig. 6. Long short term memory

S, p is the inputs and outputs of the network, W is the weights

In the figure 5.2, we have seen the process that undergoes text generation using LSTM. From the extracted features of the image it has sent to the LSTM part which may take the classified object and convert those objects into the text and once all the classified objects are converted it then make the sentence.

5) Speech synthesis

In this project we also advent a new feature, that is connecting the caption to audio which may be mainly helpful for visually impaired people. For this image captioning we have added this sound to make this application useful for visually impaired people.

This speech synthesis has done by importing the libraries gTTS (google text to sound) which may make any text play sound.

gTTS

- gTTS stands for Google Text to Speech.
- gTTS library is used to convert the text into speech in python.
- It is a useful tool that changes text into audio which can be saved as an mp3 file
- It supports several languages like English, French, German, Tamil, Hindi

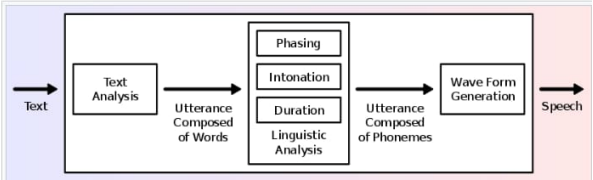

Fig. 7. Speech synthesis

6. Result Analysis

A. Text Generation Analysis Using Bleu Score

BLEU scores are utilized in producing the comparison results of reference sentences to the output result of at least one reference interpretation.

We at that point figure BLEU scores for 1, 2, 3, and 4 combined n-grams. An ideal match brings about a score of 1.0, while an ideal confound brings about a score of 0.0

The sample outcome of image captioning was:

```
BLEU-1: 0.469223
BLEU-2: 0.240794
BLEU-3: 0.157674
BLEU-4: 0.064845
```

Fig. 8. Bleu Scores

These are the evaluations of bleu scores for our image captioning and enhance the good quality description of an image.

B. Image-Caption Analysis

Caption generation is a process of getting a descriptive sentences of a given image, this is a challenging task for both artificial intelligence and computer vision.

Now, let's generate the caption for the image.

Fig. 9.

```
startseq two girls are walking on the sidewalk endseq

C:\Users\sai krishna\Documents\python>
```

Fig. 10.

```
startseq man in red shirt is jumping on the beach endseq

C:\Users\sai krishna\Documents\python>
```

In the above images, we have some limitation on getting in the caption of same color objects at multiple times which can prevent them in future.

C. Speech synthesis

This is connecting the caption to an audio. For the image caption we've added this sound to make this application useful for the visually impaired people.

This speech synthesis has done by importing the libraries GTTS (google text to sound) which may makes any text to play sound and the text generated saved as shown in fig. 11.

```
description = generate_desc(model, tokenizer, photo, max_length)
print(description)
mytext = description
language = 'en'
myobj = gTTS(text=mytext, lang=language, slow=False)
myobj.save("welcome.mp3")
playsound('welcome.mp3')
```

Fig. 11.

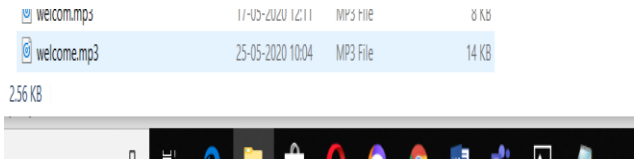

Fig. 12.

As given in the fig. 11, Now the audio file saved in the mp3 format as shown in fig. 12.

7. Conclusion and Future Work

Hereby we conclude that the proposed image recognition method has been done through an attention mechanism. Flick8k dataset has been used for training and testing purposes, which contains 8000 pictures. VGG-16 has been used for extracting the features which are then fed into LSTM for generating the captions. Through these models in the proposed method, we have created an aid for the visually impaired people by converting the captions into speech. However, our method doesn't work best for images with a lot of overlapping objects which need to addressed in future work. Also, the implementation can be enhanced by giving a greater number of images and text datasets with shorter captions for training. It consumes less amount of time to load the sentence and work

with high efficiency.

References

[1] Michal Nijman et al Image captioning using CNN," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CANADA, USA, May 2017.

[2] Rebecca Mason, " Data driven image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, UK, USA, MAY 2017.

[3] Andrej Karapathy at al., "Connecting images and natural language," in Proceedings of the IEEE Conference on International Conference on Computer Vision, pp. 4904–4912, Venice, Italy, April 2016.

[4] Marco Pedersoli, "Areas of Attention in Image captioning," in Proceedings of the IEEE Conference on International Conference on Computer Vision, US, UK June 2018.

[5] Quan Zeng You, "Image captioning using semantic attention," in Proceedings of the IEEE ConfeSemStyle: figuring out how to produce stylised picture inscriptions utilizing unaligned content," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition on International Conference on Computer Vision, Venice, Italy, May 2017.

[6] Der Forschung, Der Lehre, Der Balding: how solving the image captioning-Retrieval problem creates conversations, July 2018.