IMAGE PROCESSING ALGORITHM FOR TEXT RECOGNITION AND FEATURE EXTRACTION FOR VISUALLY UNSIGHTED PEOPLE

BalajiChaugule
Zeal College of Engineering and Research, Pune.
Asst. Prof. in Computer Department
SavitribaiPhule Pune University,Maharashtra.

SahilSharafatSayyed
Zeal College of Engineering and Research, Pune.
Student inComputer Department,
SavitribaiPhule Pune University,Maharashtra.

Omkar Ware
Zeal College of Engineering and Research, Pune.
Student inComputer Department,
SavitribaiPhule Pune University,Maharashtra.

Shubham Suresh Kadam
Zeal College of Engineering and Research, Pune.
Student inComputer Department,
SavitribaiPhule Pune University,Maharashtra.

SaurabhSandeep Singh
Zeal College of Engineering and Research, Pune.
Student inComputer Department,
SavitribaiPhule Pune University,Maharashtra.

***Abstract*:**In recent years, with the rapid development of artificial intelligence, image to text to audio has gradually attracted the attention of many researchers in the field of artificial intelligence has become an interesting and onerous task.

Picture generates natural language descriptions automatically according to the context observed in an image, it's is an important part of scene understanding, which combines the knowledge of computer vision and NLP.

Application of a photo is extensive and significant, for example, the realization of human-computer interaction.

Our base paper summarizes the related methodology and focuses on the attention mechanism which plays an important role in computer vision and is recently widely used in Machine learning tasks. Furthermore, the benefits and the shortcomings of these methods are discussed, providing the commonly used datasets and evaluation criteria in this field.

Keywords: Machine Learning, Feature Extraction, Text Conversion, Image Conversion, CNN.

I. INTRODUCTION

The machine to be able to automatically describe objects in a picture along with their relationships or the actions that are being performed using a learnt language model is a very challenging task, but with massive impact in many areas. For example it could help people with visually impairment better understand visual inputs, thereby acting as an assistant or a guide. The model be able to solve the computer vision challenges of identifying the objects in an image, but it must also be intelligent enough to capture and express the object's relationships in natural language. For this reason, image caption generation has long been considered as a difficult problem. Its purpose is to mimic the human ability to comprehend and process huge amounts of visual information into a descriptive language, making it an attractive problem in the field of AI.

Image can be used for a variety of use cases such as assisting the blind using text to speech by real time responses about the surrounding environment through a camera video, increasing social medial experience by converting captions for images in social media feed as well as messages to speech. Assisting young children in recognizing objects as well as learning the English language.

The captions for every photo on the internet can lead to faster and descriptively accurate image searches as well as indexing. In robotics, the perception of environment for an agent can be given a context through natural language representation of environment through the captions for the images in the camera feed.

To address these issues, deep neural network architectures were used along with a language model, which coembeds the images and captions in the same vector space. The simple and basic approach is essentially the same – a Convolutional Neural Network (encoder) that generates a sequence of features which are fed into a sequence-to-sequence model (decoder) that learns the language model to generate natural language descriptions.

Our model is inspired by Show and which uses GoogLeNet CNN to extract image features and generate captions using Long Short Term Memory cells. We differ from theimplementation to optimize for real-time scene. Show, Attend makes use of new developments in machine translation and object detection to introduce an attention based model that takes into account several "spots" on the image while generating the captions.

They extract features from lower convolutional layer instead of extracting from the penultimate layers, resulting in a feature vector of length for every image. This number can be interpreted on the image, each having a feature vector of length. These are taken into consideration while generating the captions. Using visual attention, the model was able to learn to fix its sight on the important objects in the image when generating captions. They introduced two attention mechanisms, a "soft" deterministic attention mechanism trained with back-propagation methods; and a "hard" stochastic attention mechanism trained by maximizing an approximate variation lower bound. Adding focus and attention improves the quality of the generated captions, but with a huge cost of additional trainable weights. Also, processing and storing the data into the database encoded features takes a lot of computational time, thereby rendering this technique futile for real-time applications onconsumer cell phone hardware.

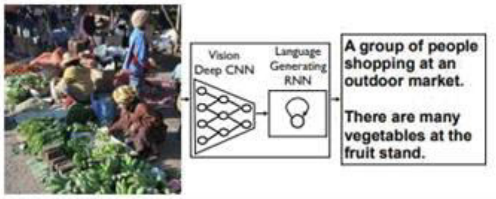
Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Fig 1.A neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image.

II. RELATEDWORK

[1]. Nada Farhani, NaimTerbeh, MounirZrigui, "Image to text conversion: state of the art and extended work," IEEE/ACS 14th International Conference on Computer Systems and Applications, 2017.
Limitations :Authors have implemented Image to text conversion in this base paper.
Solution : We are implementing image to text and then text to audio conversion. This is an advancement of the implementation work.

[2]. Azmi Can Özgen, MandanaFasounaki, Hazım Kemal Ekenel, "Text detection in natural and computergenerated images," 2017.
Limitations :Authors are using a tesseractocr software to implement their project.
Solution : We are using two main components that are CNN and RNN which are easy and reliable.

[3]. D. B. K. Kamesh, S. Nazma, J. K. R. Sastry , S. Venkateswarlu , "Camera based Text to Speech Conversion, Obstacle and Currency Detection for Blind Persons" Indian Journal ofScience and Technology, Vol 9(30), DOI: 10.17485/ijst/2016/v9i30/98716, August 2016.
Limitations :They are using various hardware devices that can be used to recognize various types of currency only.
Solution : We are implementing it for whole image and extracting all the data and then converting it to the speech format.

[4]. Dhulekar, Pravin A., NiharikaPrajapatr, Tejal A. Tribhuvan, and Karishma S. Godse. "Automatic voice generation system after street board identification for visually impaired." In 2016 International Conference on Global Trends in Signal

Processing, Information Computing and Communication (ICGTSPICC), pp. 91-96.IEEE, 2016.

Limitations :This base paper shows the implementation of automatic voice generation system after street board identification it is limited only to the obstacle.

Solution :We are making it to recognize anything that is captured in an image it may be anything that can be recognized.

III.PROBLEMSTATEMENTANDOBJECTIVES

A.Problem Statement

We can create an application for which we will guide blind people travelling on the roads without the support of anyone else; we can do this by first converting the scene into text and then the text to voice. Both are now famous applications of Deep Learning.

B.Objectives

- Recognize objects in the image
- Generate a fluent description in natural language
- CNN is architecture specialized in finding topological invariants the input.
- Finds relationships between atoms and infers higher abstractions.
- Highly resistant to spatial transformations.
- It learns automatically what the relevant features to extract from an input are not limitedto images: CNNs can be applied to text, audio, etc.

IV. PROPOSEDWORK

The following diagram shows the high level working of the system. The diagram shows that the system consists of two main modules which are :

- Image to text conversion or text extraction from image.
- Text to speech conversion.

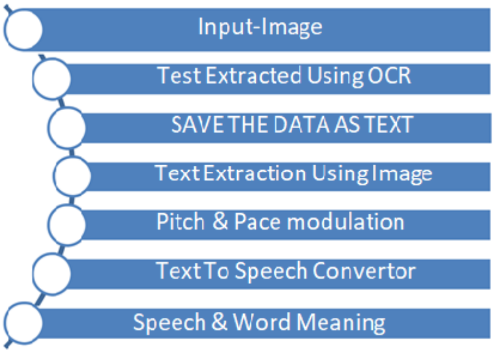

Fig 2.Image to text conversion or text extraction from image.

The process of extraction of text from different inputs is different. As shown in the diagram, extraction of text from a captured image requires natural image processing.

Text to Speech Conversion Text which may be extracted from any kind of input is stored as a text file in the working directory. Microsoft provides a Speech API which is used to provide various voices and algorithm is written to convert this text into speech.

Text extracted is then written in a file in the working directory for further processing. Extraction of text from PDF is achieved using a .jar file. It is integrated with the application for a seamless extraction from PDF files. It must be kept in mind that algorithms to process natural image are such that the quality of image does not hinder the accuracy with which the text is extracted. The OCR works sufficiently accurately for images in which text orientations are not necessarily straight and resolution is low. Text may be present in any font style and size in image.

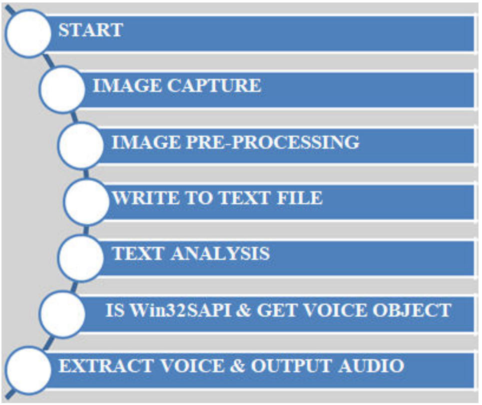

Fig 3.Text to speech conversion.

A captioning model relies on two main components. They are :

1. CNN
2. RNN

- Captioning is all about merging the two to combine their most powerful attributes i.e. CNNs (Convolutional Neural Networks at preserving spatial information and recognize objects in images.

- RNNs (Recurrent Neural Networks) work well with any kind of sequential data, such as generating a sequence of words. So by merging the two entities you can get a model that can find patterns and images, and then use that information to help generate a description of those images.

- Convolutional Neural Networks, or CNNs, were designed to map image data to an output variable. It has proven so effective that they are the go-to method for any type of prediction problem involving image data as an input.

- The benefit of using CNNs is their ability to develop an internal representation of a two-dimensional image. This allows the model to learn the position and scale in variant structures in the data, which is important when working with images.

Use CNNs for

- Image data
- Classification Conversion problems
- Regression prediction problems

More generally, CNNs method work well with the data that has a spatial relationship

Although it is not specifically developed for non-image data, CNNs achieve state-of-the-art results on problems such as document classification used in sentiment analysis and related problems.
RNNs in general & LSTMs in particular have received the most success when working withthese sequences of words and paragraphs, generally called natural language processing.
This includes both the sequences of text and sequences of spoken language represented as a time series. It is also used as generative models that require a sequence output, not only with text, but on applications such as generating handwriting.

Use RNNs for

- Text data
- Speech data
- Classification prediction problems
- Regression prediction problems
- Generative models

Design Steps :

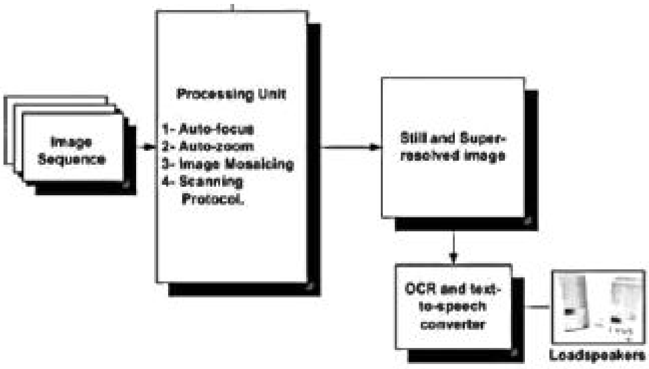

Fig 4. Design step-wise Diagram

OCR pre-processing techniques include image de-skewing, de-speckling, binarization, line removal, layout analysis, line and word detection script recognition, character segmentation or isolation, scaling and aspect ratio normalization. Despite various multiple techniques are available, OCR operation is performed on one block of text or one word at a time. The feature recognition andpattern detection algorithms are used for identification of characters. The feature detection is informed of the features of numbers or letters individually for the recognition of these characters in the document. The number of crossed lines, curves or angled lines and such features can be used for comparison of characters. The method for the pattern recognition, sample formats and fonts of textual data are fed to the system for the purpose of recognition of characters by comparison with the scanned document.

In order to enhance the accuracy of the system, handling of the complex layouts, proofreading and correction of basic errors is to be done and the document must be saved for further use. The speech synthesizer changes the final content into audio output to the visually impaired user. This system produces human speech artificially. It can be executed in either hardware or software form. The phonetic transcription and linguistic symbols representation can be converted into sound signals. The entirely synthetic model can be used for creation of voice output by replicating human voice characteristics and the vocal tract model.

CNN Architecture

Convolutional Neural Networks (CNN) is a deep model that performs well with a variety of tasks such as image classification, natural language processing, and signal processing. CNNs are explicitly designed to deal with multi-dimensional input and overcome the high number of parameters that are requested by standard FNN. For example, a single RGB image of size 64x64, in an FNN would require: 64 • 64 • 3 = 12288 neurons as input. The issues that arise when the FNN is over parameterized are the following: 1. a huge number of input neurons will require more layers at a high computation cost and time required for training 2. Over parameterization is a symptom of over fitting: in the specific case of an image, the FNN would behave too meticulous since it will take into account each single pixel.

The performance of the proposed DL structure was compared to three previous studies in the literature, due to limited previous research classifying heartbeats with noisy data using ML including SVM, FBNNs, and CNNs. The results found that, with an accuracy level of 99.34%, our proposed DL structure outperformed the SVM, FBNN, and CNN classifiers of ECG noise removal and feature representation algorithms.

Convolutional Neural Networks (CNN) is a deep model that performs well with a variety of tasks such as image classification, natural language processing, and signal processing. CNNs are explicitly designed to deal with multi-dimensional input and overcome the high number of parameters that are requested by standard FNN. For example, a single RGB image of size 64x64, in an FNN would require: 64 • 64 • 3 = 12288 neurons as input. The issues that arise when the FNN is over parameterized are the following:

- A huge number of input neurons will require more layers at a high computation cost and time required for training.
- Over parameterization is a symptom of over fitting: in the specific case of an image, the FNN would behave too meticulous since it will take into account each single pixel.

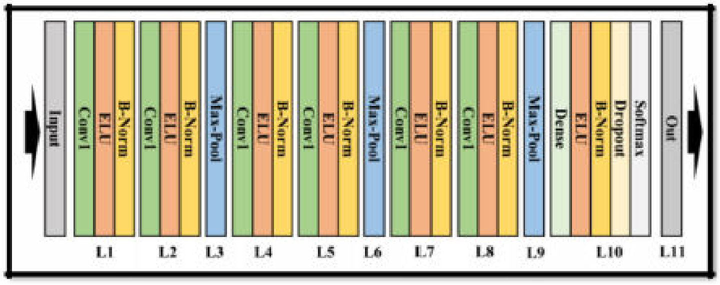

Fig 5. Architecture of CNN Model

In order to take into account, the multidimensional input, CNN"s neurons are organized in three dimensions and to reduce the overall complexity, the neurons in each layer are connected only to a portion of the previous one. This is the opposite of what happens in fully connected neural networks.

The main types of layers used in CNN are:

- Convolutional: consists of groups of neurons apply a scalar product with the connecting portions of the input.
- Pooling: down sampling to reduce the dimensionally.
- Fully connected: produces the final predictions. Generally, it is preceded by a flattening operation since the convolutional outputs are always 3-dimensional.

VI .CONCLUSION

- An image is often rich in content. Our model should be able to generate description sentences corresponding to multiple main objects for images with multiple target objects, instead of just describing a single target object.
- For the corpus description of different languages, a general image description system capable of handling multiple languages should be developed.
- Evaluating the result of natural language generation systems is a difficult problem. 0e best way to evaluate the quality of automatically generated texts is subjective assessment by linguists, which is hard to achieve. In order to

improve system performance, the evaluation indicators should be optimized to make them more in line with human experts" assessments.

- The real problem is the speed of training, testing, and generating sentences for the model should be optimized to improve performance.

VII. RESULT ANALYSIS

Below are the screenshots of the output of our project. There are three figures which indicate the line of codes and an image where the image is recognized and the output is shown on the screen as in the caption.

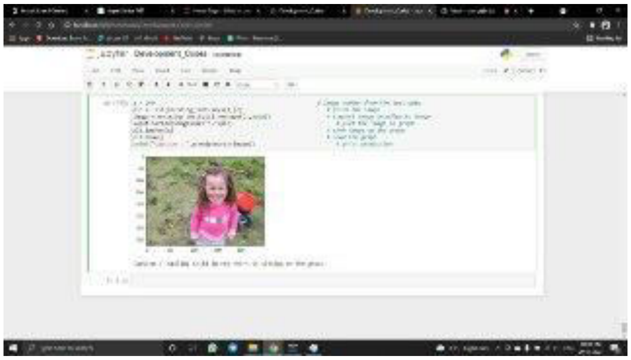

Fig 6. Smiling child in red shirt is sitting on the grass.

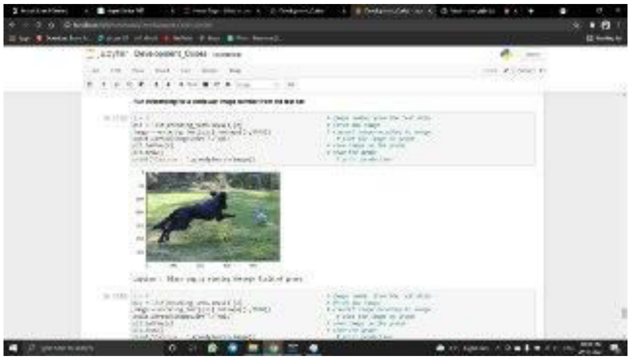

Fig 7. Black dog is running through the field of grass.

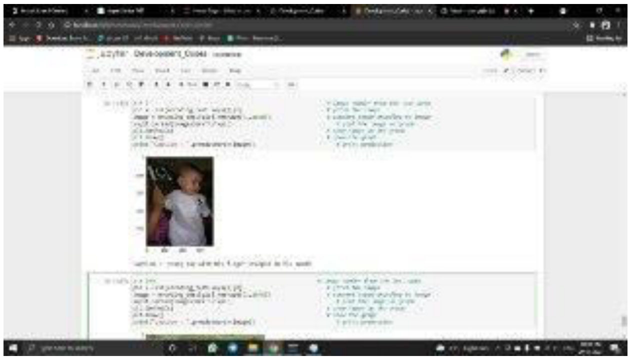

Fig 8. Young boy with his finger wrapped in his mouth.

Graphical representation of dataset analysis :

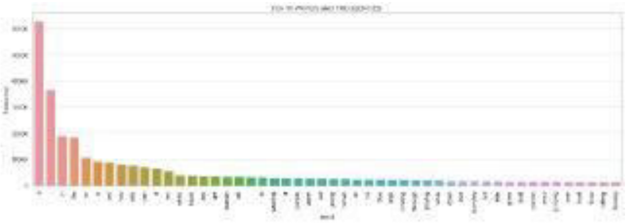

Fig 9. Bar Graph that represents top 50 words and frequencies.

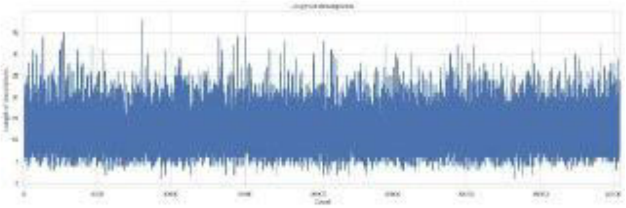

Fig 10. Figure indicates length of descriptions in graphical manner.

VIII .REFERENCES

[1] Mohammad MarufurRahman, Md. Milon Islam, Saeed Anwar Khan ,2020, "Obstacle and Fall Detectionto Guide the Visually Impaired People with Real Time Monitoring", Madhav Institute of Technology,Madhya Pradesh.

[2] ChandanDehanth, 2019, "Development of an Automated Obstacle Detector for Blind People",Department of Computer Science and Engineering, Daffodil Institute of IT, Dhaka, Bangladesh.

[3] KolapoSulaimonAlli, Moses OluwafemiOnibonoje, 2019, “Development Of An Arduino-based ObstacleAvoidance Robotic System For An Unmanned Vehicle", Babalola University, Ado Ekiti, Nigeria.

[4] AshimaArora, 2017, “Automatic number plate detection and unmanned chalan generation forodd/even rules in Delhi", IIT, Ropar.

[5] Priya Lakshmi S, Umamashewari ,2021, “Robo kart for Visually impaired people", SRM Institute, India

[6] M Geetha, R. C. Pooja ,2020, “Implementation of text recognition and text recognition on FormattedBills", International Journal of Test and Automation.