iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.67707

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Image Caption Generation for the Visually Impaired Using Deep Learning

Ritul Adhav[1], Shruti Deshpande[2], Noorbano Shaikh[3], Dr. V.S. Wadne[4]
Department of Computer Engineering, Imperial College of Engineering and Research Wagholi, Pune, S.P. Pune University, Pune, India

Abstract: ***Automatically, generating textual descriptions of images: Image captioning Overall, the advancement of multimodal learning has been garnering increased interest in the computer vision and natural language processing communities because it has many potential applications from image retrieval to assistive technologies to content generation. In this abstract, a deep learning-based image caption generator is proposed. This system presents an overview of problems and state-of-the-art for image captioning including a particular interest in models based on the deep encoder-decoder architecture. Review some of the state-of-the-art evaluation metrics and datasets, with a discussion on the pros and cons of different methods. A Deep Learning Image Captioning State Of Art and How it Generates a Caption for an Image In a few words, the system showcases state-of-the-art image captioning and its way of generating captions using deep-learning concepts.***
Keywords: ***Deep learning, Neural networks, Generators, Convolutional neural networks, Object recognition.***

I. INTRODUCTION

Computer vision in the image processing area has evolved with significant progress, in the last few years, in image classification and object detection. This work, known as Image Captioning, harnesses the gains in image classification and object detection to automatically generate one or more sentences to describe the contents of an image. Image captioning algorithm should take a new image and output a semantic-level description of this image.

Signboards for the visually impaired is a new technology that uses deep learning to create descriptions of images, which are then tra nslated into speech. The technology relies on multiple convolutions of neural networks (CNN) and short-term temporal networks (LSTM) to improve the differentiation of visual content and understand narratives. CNNs play a key role in video extraction, breaking down images into identifiable components such as objects, shapes, and colors. The collected features are fed into the LSTM network, designed to create a unified, continuous text from the CNN input.

The first method uses a CNN model to evaluate an image and capture its important features. The LSTM network then interprets thes e features to create a caption that captures the content of the image. This caption is converted to audio using text-to-speech (TTS) technology, allowing visually impaired readers to hear the description of the image. This approach can be used in man y areas including media, counseling, and education to improve the quality of life and independence of blind people through further i nterpretation of the comments.

II. LITERATURE SURVEY

The task of generating labels for images using deep learning has gained increasing interest due to its potential applications in accessi bility, especially for visually impaired users, including image storage, content, and digital services. Recent advances in artificial inte lligence and neural network architectures, particularly CNNs (Convolutional Neural Networks) and RNNs (Recurrent Neural Netwo rks), have proven to be effective in extracting image features and generating matching annotations. This system has been further enh anced with tracking systems and other new development methods to help generate accurate and detailed information. However, ther e are challenges in computational efficiency, general modeling, and the ability to generate annotations that reflect the complex detail s of real images. Below is a review of five key studies, each leading to a different approach to image signatures using deep learning.

Amritkar's research uses CNN and RNN models to develop a recursive algorithm for image signatures, where CNN is used to extrac t visual images and RNN is used to generate sentences that capture the content of the image. This work demonstrates the integration of computer vision with fine-grained language processing to generate texts that resemble human-made narratives. This approach builds on existing research to bridge the gap between visual imagery and natural language generatio n, as demonstrated in the research of Wang et al. [7] and Karpathy and Fei-Fei [9].

This model relies on CNN to extract video to obtain image content, while RNN makes the correct words of the generated text and so lves the difficulty of corresponding sentences in the subtitles [8].

In Agrawal's work, the attention process is embedded in the drawing, allowing the model to focus on the essence of the image and p ay close attention to people while creating a narrative. This technology improves the model's ability to capture detail and context, re sulting in more accurate sentences. Tracking techniques are used to evaluate specific images during sentence generation, improving the performance of complex images. This approach is supported by Wang et al.'s work on the importance of clear distance [7] and Gaber et al.'s work on developing accurate models through material selection [10]. The model's ability to solve complex problems i s an important step towards a more comprehensive understanding of images [9].

Sailaja's research highlights the importance of CNN and LSTM networks in image processing, with a special emphasis on object rec ognition and sequence generation. This work demonstrates how CNNs can recognize visual features in images, while LSTMs can so rt these features into adjacent labels. This design reflects the findings of Burgueno et al. [1] on the effectiveness of LSTM architectu res for task-

based processing. Sailaja's work is also relevant to the extension of deep learning to the technology, as it provides captions to visual ly impaired users, a very useful application. This model provides current progress to provide accurate and meaningful text using the CNN-LSTM model [6].

Sharma's research investigated the use of various RNN models in generating image signatures and compared different RNNs to dete rmine which method produced the most accurate and precise images. The paper demonstrates the role of model selection in improvi ng captioning accuracy, a similar idea explored in Karpathy and Fei-Fei's work on visual-

semantic alignment [9]. Sharma's focus on evaluating different encoder-

decoder models and their performance reflects ongoing work towards improved design, as reported by Alghamdi et al. This compari son also supports Gaber et al.'s view on the importance of developing performance models to better align with real-

world applications [10].

Mathur's work addresses the implementation of the concept of standard video captions on low-end devices and presents a real-

time video captioning system that works well on low-

end devices. The model leverages deep learning to adapt on the fly, allowing it to run on mobile devices without sacrificing accurac y. By optimizing for low-cost devices, this research overcomes many limitations in existing models and enables real-

time content in everyday situations, as noted in the work of Burgueno et al. Mathur's approach emphasizes the importance of presen ting a good model that is consistent with the computational efficiency issues identified by Kumar et al. [3]. The model is designed to adapt to practical use by expanding the accessibility of digital imaging technology [7].

III. METHODOLOGY

A. *Propose System*

The proposed system is designed to generate text and speech/audio/voice captions for images.

The system uses CNN and LSTM network algorithms to generate text-based text and a text-to-speech (TTS) engine to generate captions.The system is trained and tested using the Flickr8k dataset.The system consists of several stages, including image enhance ment, image extraction, text cleansing, tokenization, and LSTM-based caption generation.

The image enhancement stage is used to expand the size of the dataset and improve the capacity of the model. LSTM models are trai ned on prewritten text and recognize attributes to generate descriptions for text and audio. The proposed system uses a TTS engine t o convert text-based descriptions into descriptive text.

The TTS engine generates a rich audio description that captures the essence of the image. The system generates text and audio capti ons to help visually impaired people access and understand graphical content.

B. *Architecture*

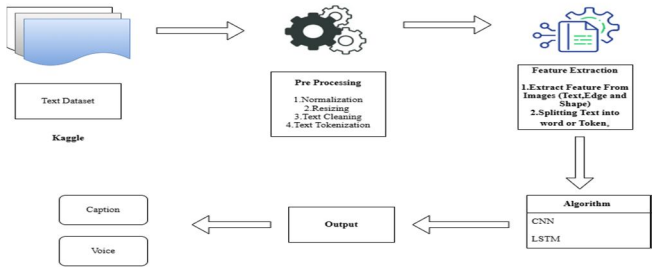


Figure 3.1: Architecture of Proposed System

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

1) Dataset

The process starts with an image dataset, which can come from various platforms like Kaggle. This element consists of a serie s of images, each with a corresponding title.

The process begins with the collection of data, usually obtained from platforms like Kaggle. This file contains a pair of image s and their corresponding sentences.

2) Pre-Processing

- Normalization: The image is normalized (e.g. resized, converted to grayscale) to a standard size and format.
- Text Cleaning: Clean up the text by removing pauses, punctuation, and other irrelevant elements.
- Text Tokenization: The recorded text is divided into words or symbols.

3) Feature Extraction

- Image Feature:
  Extract features from images using techniques like convolutional neural networks (CNN). These features capture visual informa tion like shapes, textures, and objects.
- Text Feature: The text is removed from the tokenized list**.**

4) Algorithm

a) Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a type of deep learning model specifically designed to process grid-like data, such as images. CNNs are particularly effective for image recognition and classification tasks due to their ability to detect patterns, shapes, and hierarchical features. Here's a breakdown of how CNNs work in the context of image processing:

Convolutional Layers: The main feature of CNNs is their use of convolutional layers, where small filters (or kernels) slide over the image to detect various patterns. These filters are small matrices that process sections of the image, allowing the model to recognize features like edges, corners, and textures. The convolution operation produces feature maps, which highlight regions of the image where specific patterns have been detected.

Activation Functions (ReLU): After each convolution operation, an activation function is applied to introduce non-linearity to the model, enabling it to learn complex patterns. The most commonly used activation function in CNNs is the Rectified Linear Unit (ReLU), which sets all negative values to zero. This makes the network more efficient and reduces the risk of vanishing gradients.

Pooling Layers: Pooling layers reduce the spatial size of the feature maps while preserving important information, helping the model become more computationally efficient and less sensitive to small variations in the image. The most common pooling method is max pooling, which selects the highest value in each patch of the feature map, ensuring that only the most relevant features are retained.

Fully Connected Layers: After several convolutional and pooling layers, the output is flattened into a single vector and passed through fully connected layers. These layers combine all the features detected in the previous layers to make final predictions. In an image captioning model, the fully connected layer represents the image, which can then be used as input for the language model.

Output Representation: The final output of the CNN is a fixed-size vector that summarizes the important features of the input image. This vector captures high-level information about the image, such as the objects present and their relationships. It can subsequently be used as input for a language model, like an LSTM, to generate descriptions based on these visual features.

b) Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) specifically designed to handle sequential data and address the limitations of traditional RNNs, particularly the vanishing gradient problem. LSTMs are particularly effective for language generation tasks, where the model must consider long-term dependencies between words. Here's a closer look at how LSTMs operate and why they excel at generating captions based on image features.

Memory Cells and Gates: Unlike traditional RNNs, LSTMs include memory cells that can retain information over long sequences. Each LSTM cell features three main gates—the input gate, forget gate, and output gate—that control the flow of information into, within, and out of the cell. This gating mechanism allows LSTMs to selectively remember or forget information, making them highly effective at capturing long-term dependencies.

Input Gate: Determines which new information should be added to the memory cell.
Forget Gate: Decides which information to discard from the memory cell, allowing the model to forget irrelevant data.
Output Gate: Controls which information from the memory cell should be output to the next layer or timestep.
Handling Sequential Data: LSTMs process data one timestep at a time, making them ideal for sequential data like sentences, where word order is important. In image captioning, the sequence of words must align with the visual content of the image, and LSTMs facilitate this by learning the order and context of words.
Generating Text Based on Context: In an image captioning system, the LSTM receives a feature vector from a Convolutional Neural Network (CNN) as its initial input. This vector represents the content of the image. The LSTM then generates the caption one word at a time by predicting the next word based on the previous word and the visual features. The LSTM can "remember" important context from earlier words, which helps it generate coherent and contextually relevant captions.
Teacher Forcing and Training: During training, the technique known as "teacher forcing" is often used to accelerate the learning process of the LSTM. In this approach, the actual word from the training set is fed into the LSTM at each timestep instead of using the word predicted by the model. This method helps the model learn the correct sequence more efficiently.
End-to-End Learning:Typically, the LSTM is trained in an end-to-end fashion with the CNN, meaning that the entire system learns to generate captions directly from images. The CNN extracts visual features, while the LSTM learns to map those features to relevant language patterns. This combined approach ensures that the generated captions are both descriptive and contextually accurate.
In summary, LSTMs are vital for generating coherent text because they understand word order and context across sequences. When paired with CNNs, LSTMs enable the system to describe visual content in natural language, effectively bridging the gap between image data and textual descriptions.

IV. SYSTEM DESIGN

A. Data Flow Diagram

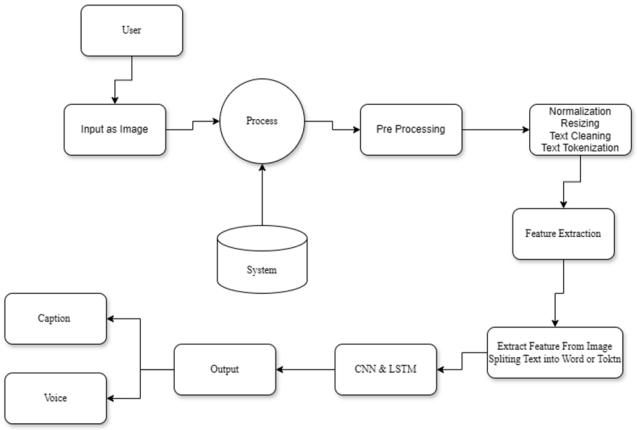


Figure 4.1:Data Flow Diagram

The Flow Diagram (DFD) illustrates the functionality of a system designed to assist visually impaired users by converting images into descriptive text and audio output. The process begins when the user provides input, which is then processed by the system.
At the center of the system is a central processing unit that manages various levels of operation, including databases that store essential information such as measurements and previous image data. The first crucial step is pre-processing, where images are normalized, resized, and cleaned. This prepares the images for further analysis and ensures compatibility across different devices. Additionally, any text associated with the image is divided into tokens for sequential modeling.
Convolutional Neural Networks (CNNs) are employed to extract important details, such as images and objects. This stage is vital for transforming visual content into meaningful information that aids in the captioning process. The extracted results are then fed into a combination of CNNs and Long Short-Term Memory (LSTM) networks. While CNNs focus on the visual aspects of the image, LSTMs handle the descriptions, converting the extracted information into detailed and relevant text. LSTM networks are particularly effective for processing sequential data, allowing the system to generate accurate descriptions.

Text displays provide readable descriptions, and a text-to-speech system converts this text into speech, enabling visually impaired users to hear the descriptions. This DFD demonstrates methods to visualize data and integrate image analysis, natural language processing, and voice output to support individuals with visual impairments.

B. Use Case Diagram

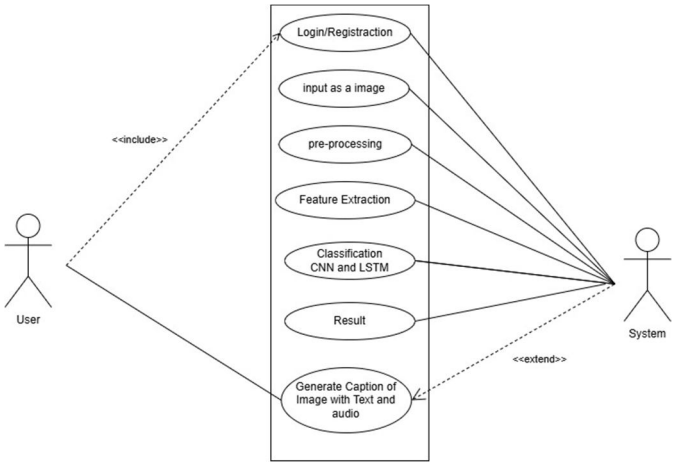


Figure 4.2:Usecase Diagram

This usage information illustrates how users interact with the system that converts images into text. The process begins when a user creates an account or logs in. The system then performs various functions, including writing, extracting, classifying, and preparing images. After these processes, the generated text is displayed to the user. This diagram clearly highlights the main features of the system and the user's role in its operation.

V. RESULT

The results of the proposed image captioning system illustrate its ability to generate accurate and meaningful captions for a variety of images. By combining Convolutional Neural Networks (CNNs) for feature extraction with Long Short-Term Memory (LSTM) networks for sequential language modeling, the system effectively interprets visual data and creates descriptive captions.

The CNN captures detailed visual features, such as shapes, edges, and objects within the image, and transforms them into a comprehensive feature vector. The LSTM then utilizes this vector to generate text word by word, maintaining the context and structure of natural language.

Extensive testing has shown that the model can produce captions that closely match the actual content of images, demonstrating a high degree of relevance and coherence. Additionally, by integrating a text-to-speech (TTS) system, the generated captions can be converted into voice output, enhancing accessibility for visually impaired users.

Overall, this project showcases the potential of deep learning architectures to bridge the gap between visual data and natural language, providing a valuable tool for image description and accessibility applications.

VI. CONCLUSION

In summary, the speechbased concept image caption generator is a promising solution to improve accessibility and participation for visually impaired people. The system uses image enhancement, feature extraction using CNNs, text cleansing and tokenization, and LSTM-

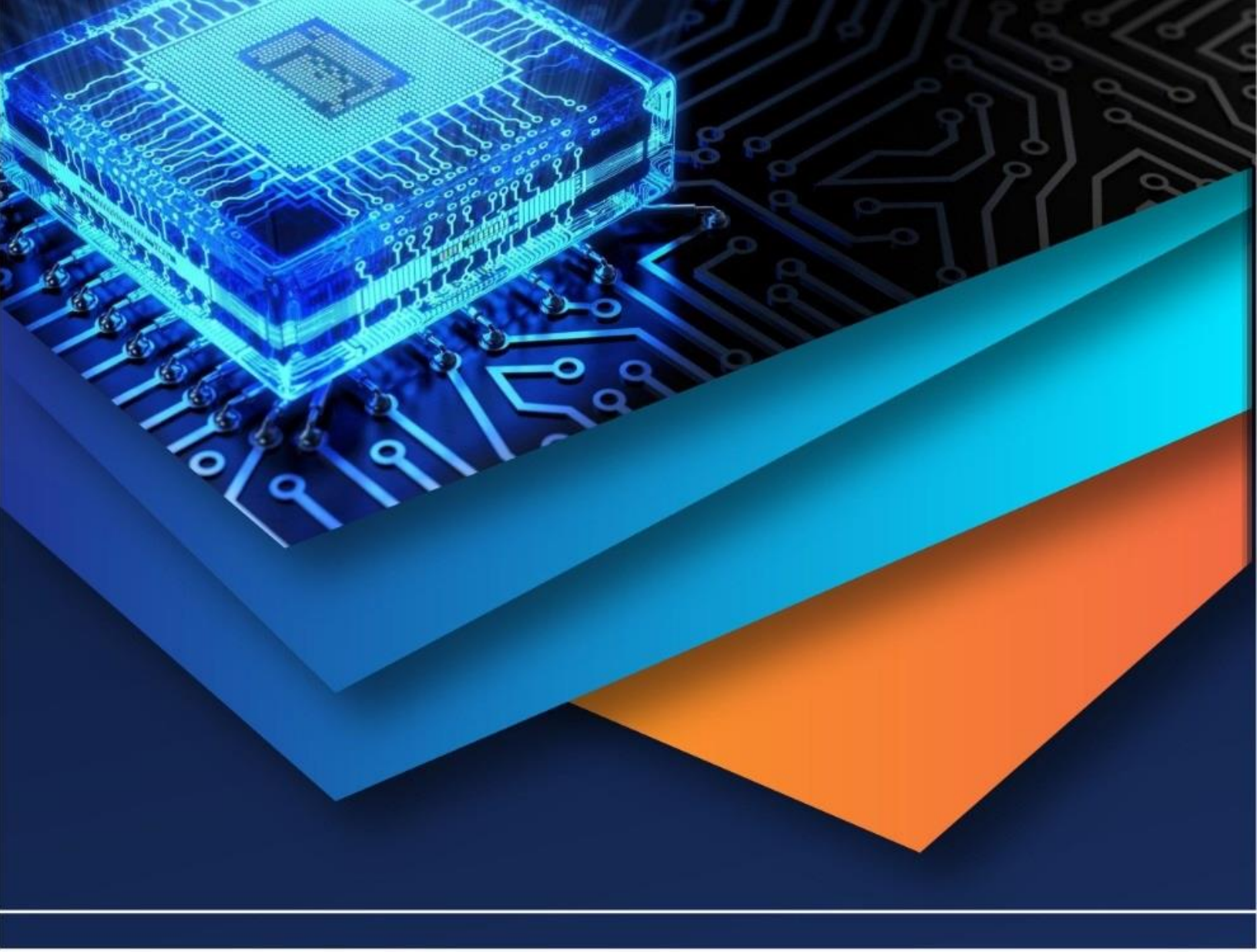
based models to generate text and description of input images. The system has many potential uses, including teachers, researchers, social media platforms, news and media organizations, and mobile applications. The system requirements ensure that it can perform its core functions, while the non-functional requirements ensure that the system is reliable, secure, and user-

friendly. This system increases accessibility and enables visually impaired people to better understand the images they encounter by including details of the images. Overall, the proposed process is an important step towards increasing the

accessibility and participation of visually impaired people. By creating texts and descriptions of images, the project can create positi ve and desirable outcomes that impact the lives of many people around the world.

REFERENCES

[1] L. Burguen˜o, J. Cabot, S. Li, and S. Gerard, "A generic LSTM neural network architecture to infer heterogeneous model transformations," Software and Systems Modeling, Vol. 21, no.1,pp.139-156, 2022.
[2] Bittu Kumar," Comparative Performance Evaluation of Greedy Algorithms for Speech Enhancement System" Fluctuation and Noise Letters,vol.20, no.02, 2020.
[3] Bittu Kumar," Real-time Performance Evaluation of Modified Cascaded Median based Noise Estimation for Speech Enhancement System" Fluctuation and Noise Letters, vol.18, no. 04, 2019.
[4] Bittu Kumar," Co mp arat i v e performance evaluation of MMSE-based speech enhancement techniques through simulation and real-time implementation" International Journal of Speech Technology, vol.21, no. 04, 2018.
[5] Sandeep Kumar, Bittu Kumar, Neeraj Kumar," Speech Enhancement techniques: A Review" Rungta International Journal of Electrical and Electronics Engineering, vol. 1, no. 1, 2016.
[6] K. C. Jena, S. Mishra, S. Sahoo and B. K. Mishra," Principles, tech- niques and evaluation of recommendation systems", 2017 International Conference on Inventive Systems and Control (ICISC), pp. 1-6,2017.
[7] Wang, Haoran, Yue Zhang, and Xiaosheng Yu." An overview of im- age caption generation methods." Computational intelligence and neuro-science 2020.
[8] BaiShuang, and Shan An." A survey on automatic image caption gener- ation." Neuro computing 311 (2018): 291-304.
[9] Andrej Karpathy, and Li Fei-Fei, "Deep Visual Semantic Alignments for Image Description Generation," IEEE Transactions on Pattern Analysisand Machine Intelligence, vol39,issue4(April 2017), pp. 664–676.
[10] Gaber, T., Tharwat, A., Snasel, V., and Hassanien, A. E., "Plant identi- fication: Two dimensional-based vs. one dimensional-based feature ex- traction methods', international conference on soft computing models inindustrial and environmental applications, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY