iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54470

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 11 Issue VI Jun 2023- Available at www.ijraset.com

Image Orator - Image to Speech Using CNN, LSTM and GTTS

Ayesha Jahan[1], Sanobar Shadan[2], Yasmeen Fatima[3], Naheed Sultana[4]
[1, 2, 3]*UG Student,* [4]*Associate Professor, Department of Information Technology, Stanley College of Engineering and Technology for Women*

Abstract: ***This report presents an image to audio system that utilizes a combination of Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) for image captioning and Google Text-to-Speech (GTTS) for generating audio output. The aim of the project is to create an accessible system that converts images into descriptive audio signals for visually impaired individuals. The proposed system has the potential to provide meaningful context and information about the image through descriptive audio output, making it easier for visually impaired individuals to engage with visual content. In conclusion, the proposed image to audio system, which combines LSTM and CNN for image captioning and GTTS for audio generation, is a promising approach to making visual content more accessible to individuals with visual impairments. Future work may involve exploring different neural network architectures, optimising the system for real-time performance, and incorporating additional audio features to enhance the overall user experience.***

Keywords: Image Captioning, Computer Vision, Natural Language Processing, Convolutional Neural Networks, Long Short-Term Memory, Google Text-to-Speech.

I. INTRODUCTION

An image to audio system is a type of technology that converts visual information, such as pictures or videos, into caption and then audio output. This technology can be useful for individuals who are visually impaired, as it allows them to "hear" images, making it easier for them to understand the content. There are different approaches to creating an image to an audio system, but one common method involves using deep learning algorithms to analyze the content of the image and generate corresponding audio descriptions. These descriptions may include details such as the colors, shapes, and textures of the objects in the image, as well as any text that may be present. Image to audio systems can also be used in applications such as automated image captioning, where images are analyzed and captioned with corresponding audio output. This technology has the potential to make visual information more accessible to a wider audience, and can be particularly helpful in educational and entertainment contexts.

A. Problem Statement

The problem of generating natural language descriptions of an image to describe the visual content has received much interest in the fields of computer vision and natural language processing and support of the visually impaired people.Although the visually impaired people use other senses such as hearing and touch to recognize the events and objects around them, the life quality of those people can be dramatically lower than standard level. For this reason, studies such as “guide dog” [1], “smart glasses” [2] and “image captioning”[3] are reported in order to improve the life quality of the visually impaired. The system should be accurate, reliable, and able to handle various types of images. Additionally, the system should be user-friendly and accessible to individuals with different levels of technical expertise.

B. Motivation To Project

There are several potential motivations for building an image to audio system, including:

1) *Accessibility:* An image to audio system could provide an alternative way for people with visual impairments to experience visual content, such as images, photographs, and graphics.
2) *Creative expression:* An image to audio system could be used as a tool for artists and musicians to explore new forms of expression by converting visual images into soundscapes and musical compositions.
3) *Data sonification:* Image to audio systems could be used to convert complex data sets, such as weather patterns or stock market trends, into audible patterns that can be more easily understood and analyzed.

4) *Assistive technology:* An image to audio system could be used as an assistive technology tool to help people with learning disabilities or cognitive impairments better understand visual information by converting it into an auditory format.
5) *Entertainment:* An image to audio system could be used to create interactive multimedia experiences, such as virtual reality or augmented reality games, that incorporate both visual and auditory elements.

Overall, an image to audio system has the potential to open up new avenues of exploration and creativity across a range of fields and applications. The main motivation for building this system is to help visually impaired people to be able to hear pictures and improve their quality of life.

C. Theory Background

Neural networks are a powerful class of machine learning models that have found numerous applications in image processing. In image processing, neural networks are used to perform tasks such as image classification, object detection, image segmentation, and image restoration, among others. The layers of Neural networks in general are depicted below. Neural networks used in image processing are typically deep neural networks, which consist of multiple layers of interconnected neurons. The input to these networks is an image, which is processed through the layers to produce a desired output.

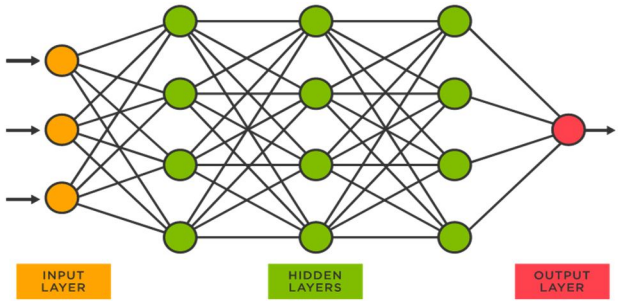


Fig.1: Layers of Neural Networks

One of the most popular types of neural networks used in image processing is the Convolutional Neural Network (CNN). CNNs are specifically designed to process images, and they are very effective at learning features and patterns from images. In a CNN, the input image is processed through a series of convolutional layers, which learn to detect low-level features such as edges and corners. These features are then combined in higher-level layers to detect more complex patterns such as shapes and objects.

Another type of neural network used in image processing is the Recurrent Neural Network (RNN). RNNs are used for tasks that involve sequences, such as image captioning or video classification. RNNs are designed to process sequences of data, and they are capable of learning long-term dependencies between the elements of the sequence.LSTM (Long Short-Term Memory) is a type of Recurrent Neural Network (RNN) that has been used in various applications in image processing. LSTMs are particularly effective for processing sequences of data, including time-series data and spatial-temporal data, which are often encountered in image processing tasks. LSTMs are designed to overcome the vanishing gradient problem that can occur in traditional RNNs, which makes them well-suited for tasks that involve processing long sequences. In image processing, LSTMs have been used for tasks such as image captioning, where the network is trained to generate a textual description of an image. The LSTM is used to encode the image features into a sequence of vectors, which are then decoded into a textual description. Overall, LSTMs have proven to be a powerful tool for processing sequences of data in image processing tasks, and they have enabled significant advancements in tasks such as image captioning and video classification.

1) *CNN Model:* A convolution neural network has multiple hidden layers that help in extracting information from an image. The four important layers in CNN are:

a) Convolution layer
b) ReLU layer
c) Pooling layer
d) Fully connected layer

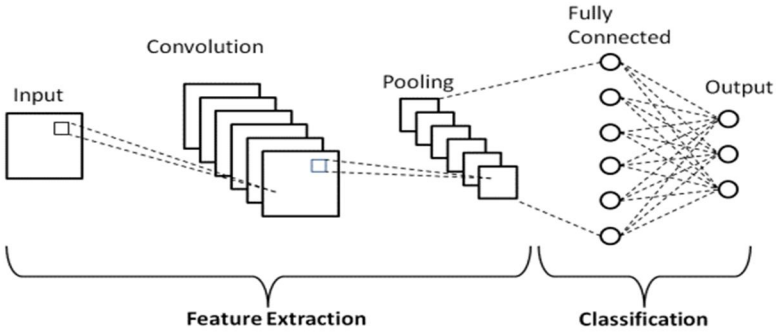


Fig.2: CNN Architecture

2) LSTM Model: Firstly, at a basic level, the output of an LSTM at a particular point in time is dependant on three things:

a) The current long-term memory of the network — known as the cell state
b) The output at the previous point in time — known as the previous hidden state
c) The input data at the current time step

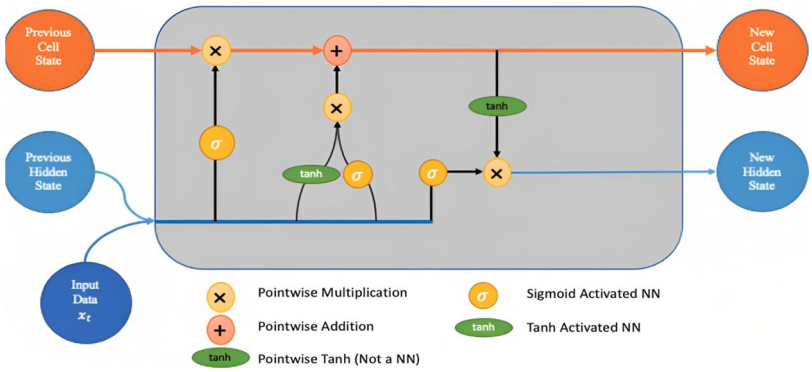


Fig.3: LSTM Architecture

3) Text-to-Speech (TTS): TTS architecture consists of several modules that work together to convert text into spoken audio. The architecture is designed to be modular, with each module performing a specific task in the TTS process. Here are the main modules that make up the Google TTS architecture:

a) Text processing
b) Phonetization
c) Prosody prediction
d) Waveform synthesis
e) Audio output

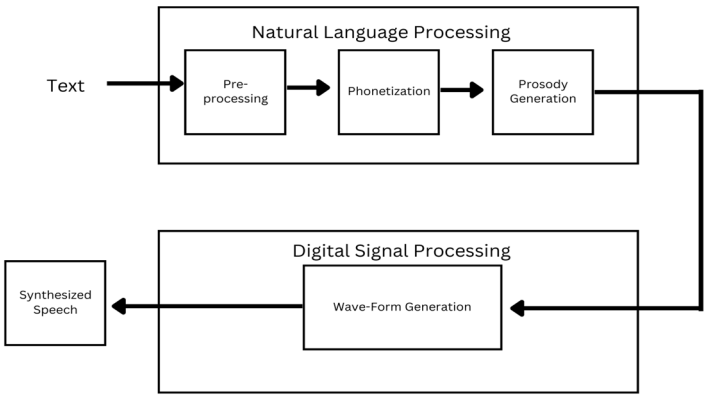


Fig.4: Waveform Synthesis

II. LITERATURE SURVEY

Image to audio systems have shown significant progress in recent years, with various approaches and techniques proposed to convert visual information into sound. The literature review below discusses the different existing systems, such as those based on rule-based methods, image segmentation, and deep learning neural networks. These systems have shown promising results in various applications, including assistive technology, multimedia communication, and entertainment.

However, the literature review also highlights several challenges that need to be addressed to improve the accuracy, efficiency, and user experience of image to audio systems. These challenges include improving the accuracy and efficiency of conversion, developing effective evaluation metrics, and providing a more intuitive user experience. Additionally, there are several future directions for

image to audio systems, including the integration of multiple modalities, exploring new applications, developing more advanced algorithms, and incorporating user feedback.

TABLE I
RESEARCH METHODOLOGY

Authors	Year Of Publication	Methods
Linnea Hammartedt	September 2006	Feasibility Study on text to speech synthesizer
Alex Krizhevsky Ilya Sutskever Geoffrey E Hinton	2012	Imagenet classification with Deep Convolutional Neural Networks
Danish M Rawat R Sharma R	2013	Content based imge retrieval basedon colour, texture, shape and neuro fuzzy
Itunuoluwa Isewon Jelili Oyelade Olufunke Oladipupo	April 2014	Design and Implementation of Text To Speech Conversion for Visually Impaired People
Abdulrehman Mohammmed Cyrus Abanti	March 2016	Image Descriptors in Content based Image Retrieval
Chen He HaiFeng Hu.	2019	Image Captioning with Text based Visual attention
Md. Zakir Hossain	September 2020	Deep Learning Techniques for Image captioning
Ethan Baker	2021	AI Evolution : The Future of text to speech synthesis

III. PROPOSED SYSTEM

Visual information plays a significant role in our daily lives, as it helps us to understand and interpret the world around us. However, for visually impaired individuals, access to visual content can be a challenge. As deep learning models become more sophisticated, the need for effective methods of processing large amounts of data has become increasingly important. One such method is the attention mechanism, which allows a model to focus on the most relevant information when making predictions.

An attention mechanism was introduced to address the bottleneck problem because of fixed-length encoding vector usage, simply defined as an improvement over the encoder-decoder based neural machine translation system in Natural Language Processing (NLP).

Attention Mechanism is an attempt to implement the action of selectively concentrating on fewer relevant things while ignoring the others in deep neural networks.

In many deep learning models, data is processed by passing it through multiple layers of neural networks. These networks are composed of many interconnected nodes organized into layers. Each node processes the data and passes it on to the next layer. This allows the model to extract increasingly complex features from the data as it passes through the network. However, as the data passes through these layers, it can become increasingly difficult for the model to identify the most relevant information. Attention mechanisms were introduced as a way to address this limitation in these models. In attention-based models, the model can selectively focus on certain parts of the input when making a prediction.

A. Dataset Used

The Flickr 8k dataset is a collection of 8,000 images and corresponding captions that was released in 2010 by a group of researchers from the University of Illinois at Urbana-Champaign. The dataset has become a benchmark for research in computer vision and natural language processing, and has been used in a wide range of applications including image captioning, visual question answering, and image retrieval. The images in the Flickr8k dataset are diverse, covering a wide range of scenes and objects. It has 6000 train Images, 1000 Validation Images and 1000 test images. The captions are written in natural language and describe the content of the corresponding image. Each image in the dataset is associated with five captions, providing a rich source of annotations for training and evaluating algorithms.

The Flickr 8k dataset has been used in a variety of research projects. One of the most popular applications of the dataset is image captioning, where the goal is to automatically generate a natural language description of an image. In this task, the image is used as input to a model that generates a caption, which is then evaluated against the ground truth captions in the dataset.

B. System Architecture

An image-to-speech system based on a combination of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Text-to-Speech (TTS) typically follows a multi-stage architecture. Here's an overview of the system architecture:

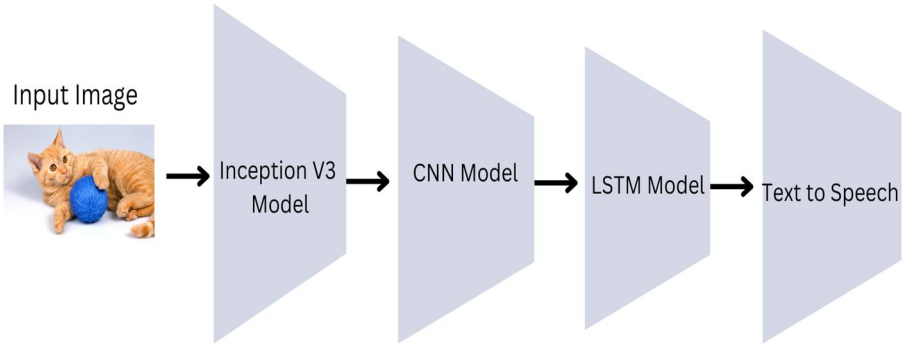


Fig.5: Simple Architectural Design

1) *Image Processing:* The input image is initially processed using a CNN model inception V3. It extracts meaningful visual features from the image, capturing important details and patterns. This step helps in understanding the visual content of the image. Feature extraction is done and using CNN Fully connected layer is formed.
2) *Sequence Generation:* The output from the CNN is fed into an LSTM network. The LSTM processes the sequence of visual features and learns the temporal dependencies among them. It generates a sequence of encoded representations that encode the visual information from the image.
3) *Text Generation:* The encoded representations from the LSTM are then used as input for a text generation module. This module generates a textual description or caption that represents the content of the image. Various natural language processing techniques, such as recurrent neural networks (RNNs) or transformers, can be employed for this task.

4) *Text-to-Speech Synthesis:* The generated textual description is passed to a Text-to-Speech (TTS) synthesis module. The TTS module converts the text into speech by synthesizing human-like speech waveforms. It utilizes techniques like concatenative synthesis, parametric synthesis, or neural network-based synthesis to generate the speech output.
5) *Speech Output:* The synthesized speech output can be delivered through a speaker or an audio playback device to provide the final auditory representation of the image content.

The entire system is trained using a large dataset of images and their corresponding textual descriptions. The training process involves optimizing the parameters of the CNN, LSTM, and TTS components to minimize the discrepancy between the generated speech and the reference speech. It's important to note that the specific architecture and techniques used in an image-to-speech system can vary depending on the implementation and research advancements in the field. The mentioned architecture serves as a general framework that combines CNN, LSTM, and TTS to enable the conversion of visual information from images into synthesized speech.

C. Implementation

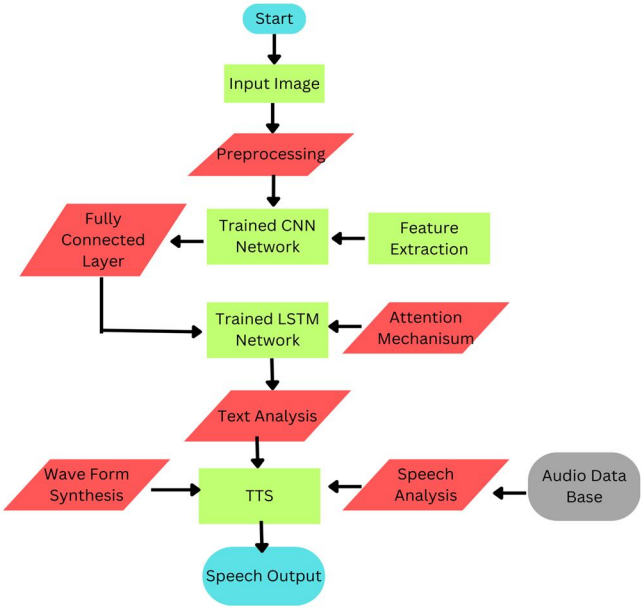


Fig.6: Flow Chart Image to Speech

D. Algorithm

To create an Image to Speech system using CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory),and GTTS (Google Text-to-Speech), you can follow the algorithm outlined below:

1) Dataset Preparation

a) Collect a dataset of images paired with their corresponding textual descriptions or captions.
b) Preprocess the dataset by resizing the images to a fixed size and normalizing the pixel values.
c) Tokenize the textual descriptions and create a vocabulary of words.

2) CNN Feature Extraction:

a) Utilize a pre-trained CNN model (such as VGG16, ResNet, or Inception) to extract image features. We used inception V3
b) Remove the classifier layers from the CNN model.
c) Pass the input images through the CNN model and obtain the output feature maps.
d) Flatten the feature maps and store them as image features.

3) LSTM Text Generation

a) Implement an LSTM-based sequence-to-sequence model to generate textual descriptions given the image features.
b) Initialize the LSTM model with an embedding layer to map words to fixed-dimensional vectors.

c) Combine the image features with the embedded word vectors as input to the LSTM.
d) Train the LSTM model using the image features as inputs and the textual descriptions as targets.
e) The LSTM will learn to generate relevant descriptions based on the image features.

4) Image to Text Conversion
a) Use a pre-trained object detection model Inception V3 to detect objects in the input image.
b) Generate a textual description of the detected objects using the trained LSTM model.
c) Convert the generated textual description into speech using the GTTS library.
d) Save the speech output as an audio file.

5) System Integration
a) Create a user interface or application to accept input images.
b) Utilize the CNN model to extract image features from the input image.
c) Feed the image features into the LSTM model to generate a textual description.
d) Convert the textual description to speech using the GTTS library.
e) Play the speech output to the user or save it as an audio file for later use.

It's important to note that implementing the above algorithm requires a good understanding of deep learning concepts and familiarity with frameworks like TensorFlow or PyTorch. Additionally, acquiring a suitable dataset and training the models may require significant computational resources.

IV. RESULTS & DISCUSSIONS

The accuracy of an image-to-speech system is a crucial performance metric that measures the system's ability to accurately transcribe the content of an image into speech. The accuracy metric evaluates how well the system can recognize and convert the visual information contained in the image into coherent and meaningful spoken output. It encompasses various aspects, including correctly identifying and interpreting objects, scenes, text, and other relevant visual elements within the image.

To evaluate accuracy, we used sparse categorical cross entropy a pre defined function in TensorFlow keras library. Calculates how often predictions match integer labels.

This metric creates two local variables, total and count that are used to compute the frequency with which y_pred matches y_true. This frequency is ultimately returned as sparse categorical accuracy: an idempotent operation that simply divides total by count.

TABLE II
PERFORMANCE OF INCEPTION V3 MODEL

Model	Accuracy	Loss
InceptionV3	0.6212	1.5505

The accuracy of Inception V3 to extract features is 0.6212 which is better that other models indicating a better caption can be generated.

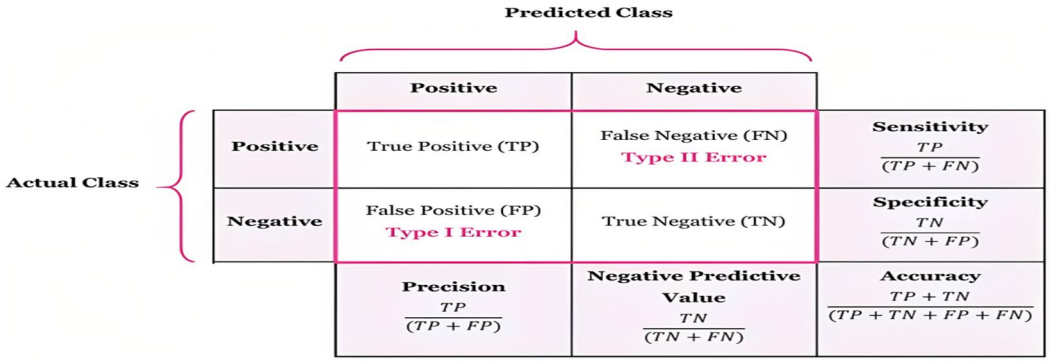


Fig.7: Precision And Accuracy

The formula to find accuracy is given above. We are finding the accuracy using a predefined-function name dsarce categorical entropy which will give accuracy and validation accuracy values.

TABLE III
IMAGE ORATOR ACCURACY AND LOSS

Epoch	Time taken	Accuracy	val_accuracy
1/30	125s 974ms/step	0.3753	0.5629
2/30	71s 744ms/step	0.5748	0.6101
4/30	71s 736ms/step	0.6326	0.6396
6/30	70s 726ms/step	0.6597	0.6559
8/30	69s 721ms/step	0.6807	0.6652
10/30	68s 707ms/step	0.6987	0.6707
12/30	69s 713ms/step	0.7145	0.6727
14/30	69s 717ms/step	0.7288	0.6739
16/30	69s 718ms/step	0.7451	0.6738

The accuracy value in the final epoch is 0.74 indicating better precision. 74% accuracy is better than the accuracy values for existing systems which were lesser than 70%.

V. CONCLUSION

A. Summary

In conclusion, image to text captioning using deep learning is a powerful technology that enables machines to describe images accurately in natural language. This technology combines various deep learning techniques such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and attention mechanisms to analyze and understand the content of an image. With the use of GTTS (Google Text-to-Speech) technology, this method allows the machine to generate voice output of the image captions, making it even more accessible to people with visual impairments.

The advancements made in image to text captioning using deep learning have enormous potential in various fields, including education, healthcare, and entertainment. It enables machines to recognize and describe images accurately, which can be useful in creating applications for the visually impaired or for automated image captioning in social media platforms.

Overall, the ability of deep learning models to understand images and generate accurate captions using natural language is a significant milestone in the field of computer vision. The combination of CNN, LSTM, and attention mechanisms, along with GTTS, represents a powerful technology as it is a highly valuable and versatile system that has a wide range of applications, from assistive technology to entertainment, and is well-positioned to become an integral part of future multimedia systems such that it can greatly enhance the accessibility of images for individuals who are visually impaired.

B. Future Enhancements

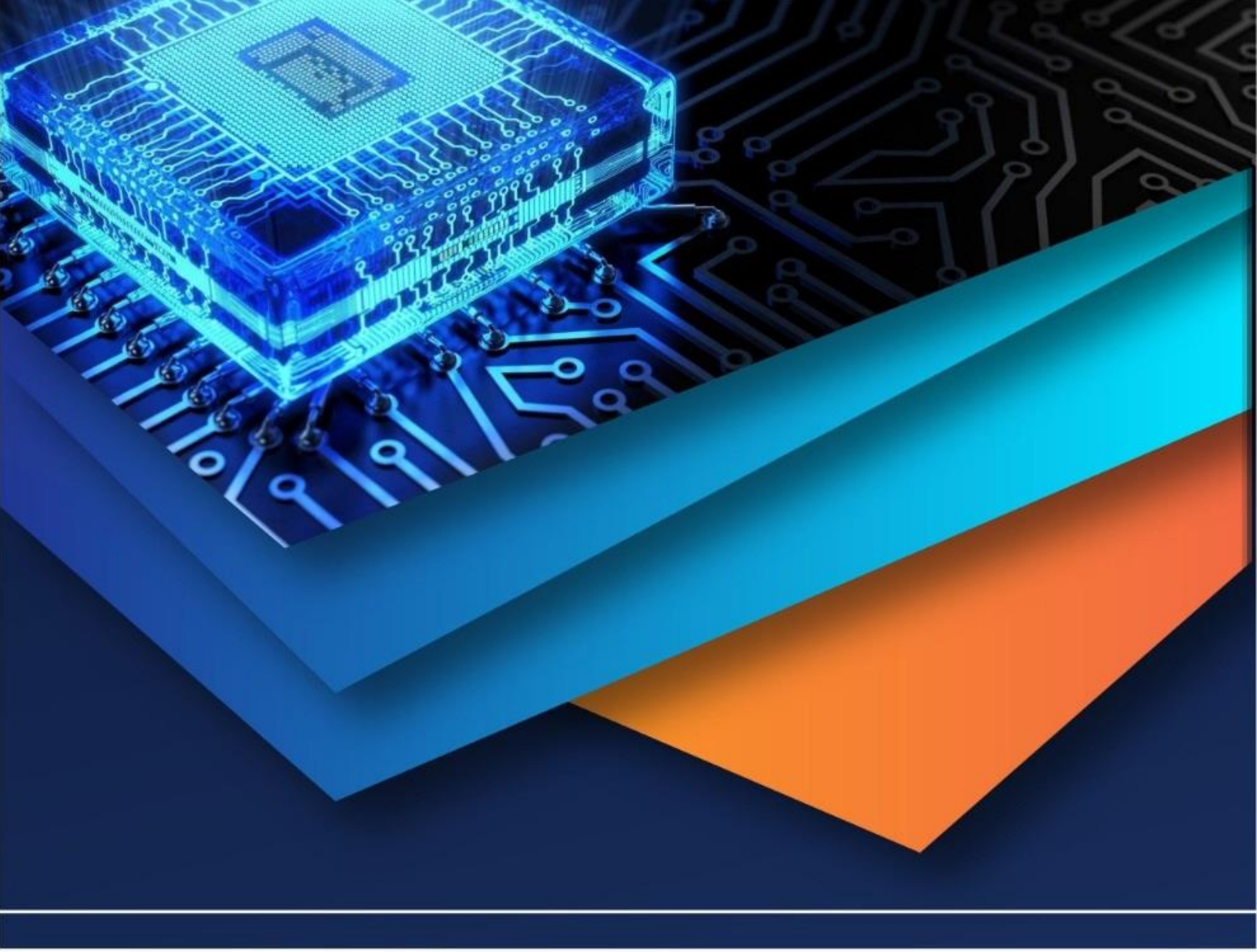
Image to text captioning using deep learning, CNN, LSTM, attention mechanism, and GTTS has made significant progress in recent years. However, there are still several areas where further research and development can improve the technology. Here are some potential future enhancements:

1) *Multimodal Image Understanding:* Combining multiple modalities such as audio and video with image captioning can improve the accuracy and richness of the generated captions.
2) *Real-time Image Captioning:* Developing real-time image captioning technology that can generate captions for streaming video in real-time can open up new applications such as live event captioning and video conferencing.
3) *Multi-task Learning:* Multi-task learning can be used to develop models that can perform multiple tasks such as object detection and image captioning simultaneously, leading to more comprehensive and accurate image understanding.
4) *Improved Attention Mechanisms:* Developing improved attention mechanisms that can better focus on important parts of the image and learn to attend to different aspects of the image for different tasks.
5) *Better Language Modeling:* Developing language models that can better understand the semantics of language and improve the generation of natural language descriptions.
6) *Multilingual Audio Generation:* The system being able to generate audio in various other languages rather than English.

In conclusion, image to text captioning using deep learning, CNN, LSTM, attention mechanism, and GTTS has come a long way, but there are still several opportunities for future enhancements. These enhancements can lead to more accurate and comprehensive image understanding, real-time captioning, multi-task learning, improved attention mechanisms, better language modeling, and improved accessibility for people with disabilities.

REFERENCES

[1] Abdulrehman Mohamed,Cyrus Abanti - March 2016 - “Image Descriptors in Content Based Image Retrieval”.
[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. -2012- “Imagenet classification with deep convolutional neural networks”.
[3] Andrej Karpathy and Li Fei-Fei.-2015- “Deep visual-semantic alignments for generating image descriptions”.
[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. -2017- “Attention is all you need”.
[5] Changzhi Bai, Hangil Park-Classification of gas dispersion states via deep learning based on images obtained from a bubble sampler-arch 2021.
[6] Chen He and Haifeng Hu. -2019- “Image captioning with text-based visual attention”. In: Neural Processing Letters 49.1 (2019), pp. 177–185.
[7] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. -2016- “Image captioning with deep bidirectional LSTMs”.
[8] Danish M., Rawat R., & Sharma R - 2013- “A Survey: Content Based Image Retrieval Based On Color, Texture, Shape & Neuro Fuzzy”. Int. Journal Of Engineering Research And Application.
[9] Ethan Baker-2023-AI Evolution: The Future of Text-to-Speech Synthesis.
[10] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. -2011- “Baby talk: Understanding and generating image descriptions”.
[11] Itunuoluwa Isewon, Jelili Oyelade, Olufunke Oladipupo-April 2014-Design and Implementation of Text To Speech Conversion for Visually Impaired People.
[12] Linnea Hammarstedt-Sept 2006-Feasibility study on text to speech synthesizer.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY