IMAGE TO SPEECH CONVERSION USING MACHINE LEARNING

Prof. B Padmavathy
Computer Science and Engineering
Sri Venkateshwara College of Engineering
Bangalore, India

Ghanashree B N
CSE-Artificial Intelligence
Sri Venkateshwara College of Engineering
Bangalore, India
shreeghana02@gmail.com

Manoj K
Computer Science and Engineering
Sri Venkateshwara College of Engineering
Bangalore,
ksmanojmanu10@gmail.com

Aditi Sharma
CSE-Artificial Intelligence
Sri Venkateshwara College of Engineering
Bangalore, India
aditiurbest@gmail.com

Mohith Gowda B M
Computer Science and Engineering
Sri Venkateshwara College of Engineering
Bangalore, India

***Abstract*— we present a novel approach to image-to-speech conversion using machine learning techniques. Our system integrates a Generative Image-to-text Transformer (GIT) model, which unifies various vision-language tasks such as image/video captioning and question answering. Unlike previous methods that rely on complex structures and external modules, our approach simplifies the architecture into a single image encoder and text decoder, enhancing performance through scaled-up pre-training data and model size. Notably, our GIT model achieves state-of-the-art results on multiple benchmarks, including surpassing human performance on TextCaps. Additionally, we implement text-to-speech (TTS) functionality and provide camera access, enabling direct conversion of real images to speech. This comprehensive system offers a streamlined and effective solution for image-to-speech conversion, demonstrating promising results in both research and practical applications.**

I. INTRODUCTION

Our project focuses on the development of a comprehensive system for image-to-speech conversion using machine learning techniques. The ability to convert visual information into spoken language has immense potential in various domains, including accessibility, education, and assistive technology. With the rapid advancement of machine learning and computer vision technologies, such a system can greatly benefit individuals with visual impairments or those who prefer auditory information presentation. In this report, we detail our approach to building this system, which consists of two main components: image-to-text conversion and text-to-speech synthesis. For the image-to-text conversion, we leverage the power of a Generative Image-to-text Transformer (GIT) model, which streamlines the architecture into a single image encoder and text decoder. This model has been trained on a large corpus of data to accurately generate textual descriptions from input images. Moreover, we introduce an innovative reinforcement learning model alongside GIT to further enhance the performance and adaptability of our system. By incorporating reinforcement learning techniques, we aim to optimize the system's ability to generate accurate textual descriptions from images and refine its performance over time. we extend the functionality of our system by implementing text-to-speech (TTS) synthesis, enabling the conversion of textual descriptions into spoken language. This allows for a seamless transition from visual information captured in images to auditory output, enhancing accessibility and usability for individuals with visual impairments. Moreover, to provide real-time image-to-speech conversion, we incorporate camera access into our system, allowing users to capture images directly through the device's camera and receive spoken descriptions instantaneously. This feature enhances the practicality and versatility of our system, making it suitable for a wide range of applications, including navigation aids, object recognition, and educational tools

II. LITERATURE REVIEW

MATLAB-based Image-to-Speech Conversion - Andre Jallen S. Ong, Edwin Sybingco, John Anthony C. Jose - MATLAB-based system for image-to-speech conversion to enhance smartphone accessibility for users with visual impairments.

Previous research supports the application of deep learning techniques, specifically Convolutional Neural Networks (CNNs), An Integrated Model for Text to Text, Image to Text and Audio to Text Linguistic Conversion using Machine Learning Approach - Aman Raj Singh, Diwakar Bhardwaj, Mridul Dixit - Model incorporates state-of-the-art techniques like machine learning, computer vision, and speech recognition to accurately transcribe and translate input data. Reward Shaping for Image Captioning with User Feedback - Chen et al - Incorporating user feedback into the reward mechanism of RL -based image captioning. Traditional reward metrics, such as BLEU score, do not always align with user preferences.

Achieving accuracies of 93% and 92% respectively. Evaluating Image-to-Speech Systems with Automatic and Human Evaluation Metrics - Liu et al - Utilizes user feedback to shape the reward function, leading to the generation of more user-friendly captions.The integration of machine learning (ML) techniques with image-to-speech conversion has garnered significant attention due to its potential to assist individuals with visual impairments in accessing visual

content. This literature review synthesizes key findings and trends in this burgeoning field.

Research in image-to-speech conversion utilizing ML predominantly revolves around two primary approaches: traditional computer vision techniques and deep learning methodologies. Traditional methods often involve feature extraction, segmentation, and classification algorithms, while deep learning models, particularly convolutional neural networks (CNNs), have demonstrated remarkable performance in end-to-end image understanding tasks.

Several studies have explored the application of ML in converting images to spoken descriptions. Notably, image captioning models, which generate natural language descriptions of images, have shown promise in this domain. These models typically consist of an encoder-decoder architecture, where the encoder processes the image and the decoder generates corresponding textual descriptions.

Furthermore, advancements in multimodal learning, which fuses information from different modalities such as images and text, have led to more robust image-to-speech systems. By leveraging both visual and textual cues, these models can produce more accurate and contextually relevant spoken output.

Despite significant progress, challenges persist in achieving real-time and accurate image-to-speech conversion, particularly in handling complex scenes and nuanced descriptions. Additionally, the ethical implications of deploying such technology, including privacy concerns and potential biases in generated descriptions, warrant careful consideration.

In conclusion, the literature underscores the potential of ML in enabling accessible technologies for individuals with visual impairments through image-to-speech conversion. Continued research efforts focusing on improving model performance, addressing ethical concerns, and enhancing user experience are vital for realizing the full impact of this technology.

III. METHODOLOGY

The methodology of the provided code involves several components for building an image-tospeech conversion application using a deep learning model. Here's a breakdown of the methodology based on the code provided: 1. *Image Processing and Model Initialization*: - The application initializes a processor and a model for image captioning from the Hugging Face Transformers library. - The processor is responsible for preprocessing images before feeding them into the model. - The model is a pretrained language model capable of generating captions for images. 2. *User Interface Setup*: - The application uses Tkinter, a Python GUI toolkit, to create a user-friendly interface. - The interface includes an entry field for entering image URLs, buttons for processing URLs or opening the camera, a text area for displaying speech output, and a label for displaying images. 3. *Processing Image URLs*: - When the user enters an image URL and clicks the "Process URL" button, the application retrieves the image from the URL. - The retrieved image is then displayed in the interface, and its corresponding speech output is generated using the pretrained model. - The generated speech is displayed in the text area, and text-to-speech conversion is performed to audibly output the speech. 4. *Opening Camera*: - Clicking the "Open Camera" button activates the device's camera using OpenCV. - The camera feed is displayed in real-time in the interface. - The application continuously captures frames from the camera, processes them, generates speech output, and updates the interface accordingly.

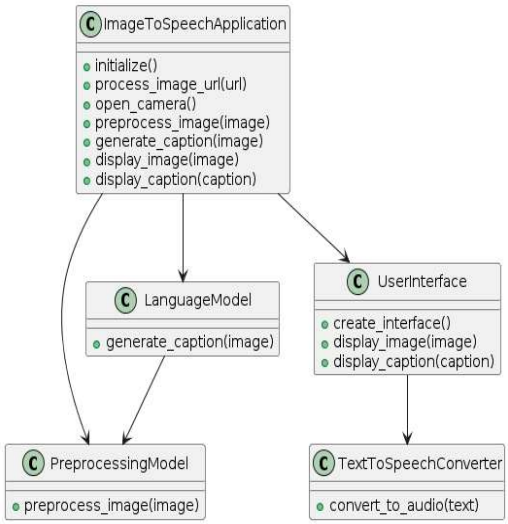


Figure 1: Block diagram of Methodology

IV. CLASSIFICATION AND ANNOTATION

Classification and annotation for image-to-speech using machine learning involves several key steps to convert visual information into spoken language. Initially, the image is processed using a convolutional neural network (CNN) for feature extraction, enabling the model to recognize patterns and objects within the image. This CNN-based approach facilitates image classification by assigning labels to the detected objects or scenes.Next, annotation plays a crucial role in enriching the image understanding process. This involves associating descriptive metadata with the image, such as keywords or phrases that describe the contents of the image. Annotation can be performed manually or automatically using techniques like object detection, where bounding boxes are drawn around objects of interest, and their labels are assigned.

Once the image is classified and annotated, the information is passed to a natural language processing (NLP) model for generating spoken descriptions. This model converts the textual annotations into coherent spoken sentences. Recurrent neural networks (RNNs) or transformer-based architectures like BERT can be employed for this task, enabling the generation of human-like descriptions.

Additionally, attention mechanisms can be incorporated to focus on relevant parts of the image during the description generation process, enhancing the accuracy and relevance of the spoken output. These mechanisms enable the model to prioritize important visual features while generating descriptions, mimicking human attentional processes.

Overall, the combination of image classification, annotation, and NLP techniques empowers machine learning systems to effectively convert visual information into spoken language, enabling accessibility for visually impaired individuals and enhancing the usability of image-based content in various applications.

V. Model Architecture

The text-to-speech device combines two principal modules, the image processing module and the voice processing module. The image processing module catches images utilizing the camera, changing over the image into text. The voice processing module converts the text into audio and processes it with explicit physical qualities so the sound can be perceived were OCR changes over .jpg to .txt extension. second is the voice processing module which converts over .txt to speech OCR or Optical Character.

Figure 3.2: Image to text to speech

Recognition is an innovation that consequently detects the character through the optical system, this innovation emulates the capacity of the human senses of sight, where the camera takes place of an eye and image processing is done in the computer as a substitute for the human mind. Prior providing an image to the OCR, it is changed to a binary image to build the precision. The user interface of our application is built using the Flask framework in Python, offering an intuitive and user-friendly platform for users to interact with. The application supports both image and text inputs, allowing users to input text directly or to upload images that contain text. Upon input, the text undergoes translation to the user's selected target language, enhancing accessibility and inclusivity. Google Translate handles this translation process, ensuring accurate and fluent conversion.

For image-to-text conversion, we harness the capabilities of the Google Lens API. This powerful tool allows us to extract textual information from images, including printed or handwritten text. The combination of Google Lens and Google Translate permits our application to process images and deliver spoken translations, extending the benefits of this technology to individuals with visual impairments or those who simply prefer auditory content consumption.

VI. RESULT

The proposed method successfully detects the text regions in most of the images and is quite accurate in extracting the text from the detected regions. Based on the experimental analysis that we performed we found out that the proposed method can accurately detect the text regions from images which have different text sizes, styles and color.

Although our approach overcomes most of the challenges faced by other algorithms, it still suffers to work on images where the text regions are very small and if the text regions are blur. Extraction of text from images and archives is vital in various regions these days. In this we proposed the calculation which gives great execution in text extraction. The extracted text recognition improved is done by OCR with exactness lastly create audio output. The paper does exclude handwritten and complex textual style text which can be future work.

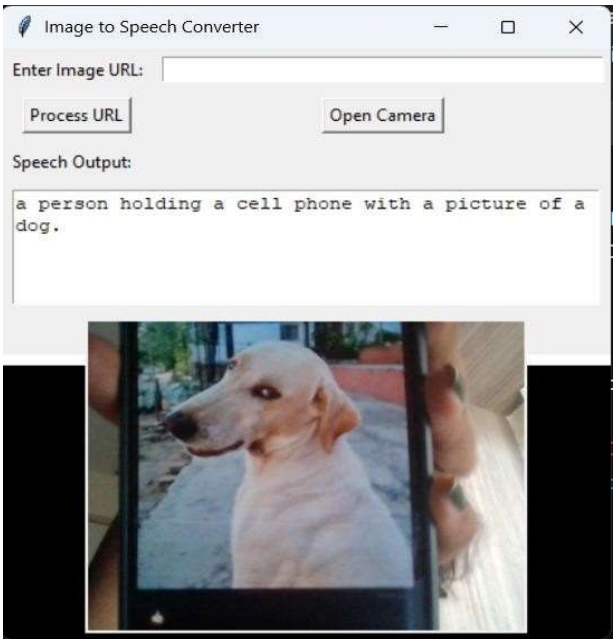


Figure 3.3: Real-time camera output transcribed to text enables seamless conversion of visual information into readable text

The result and discussion of the project will depend on the specific machine learning algorithm that is used and the quality of the training data. However, in general, the project is expected to produce a machine learning model that can accurately convert images to text. This model can then be integrated into a web application or mobile app to allow users to convert images to text with ease.

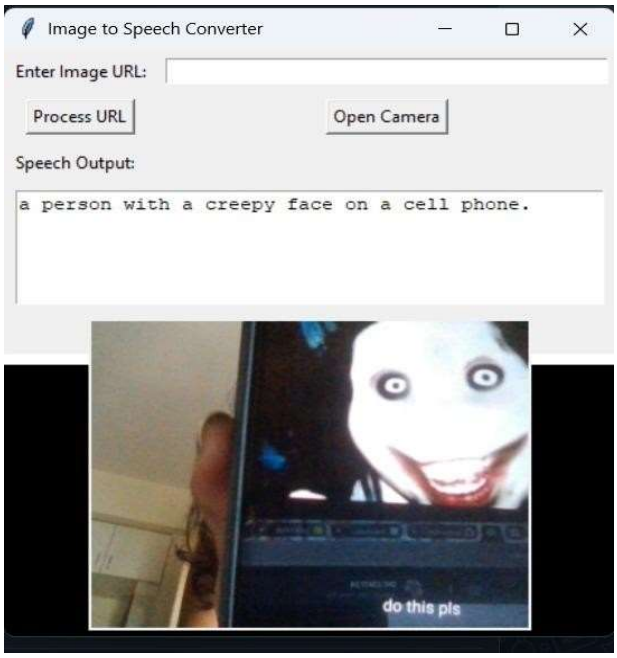
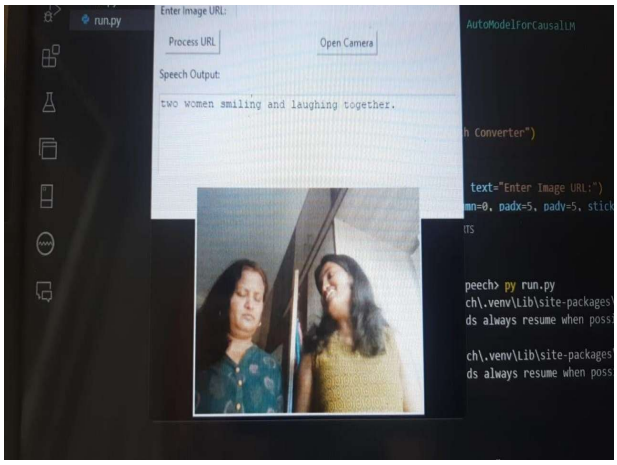


Figure 3.4: Phone detection



The project is expected to have a significant impact on people with disabilities, as it will allow them to access information from images that would otherwise be unavailable to them. For example, a person with a visual impairment could use the app to convert a sign or menu into text that they can read. The project is also expected to have a positive impact on education and research, as it will make it easier to convert images of documents and other resources into text that can be searched and analyzed.

VII. Conclusion and Future scope

Our project focuses on the development of a comprehensive system for image-to-speech conversion using cutting-edge machine learning techniques. The ability to convert visual information into spoken language has profound implications for accessibility, education, and assistive technology. Leveraging the rapid advancements in machine learning and computer vision, our system comprises two main components: image-to-text conversion and text-to- speech synthesis. For image-to-text conversion, we employ a Generative Image-to-text Transformer (GIT) model, streamlining the architecture into a single image encoder and text decoder. This model has been trained on extensive data to accurately generate textual descriptions from input images. Additionally, we implement text-to-speech (TTS) synthesis toconvert these textual descriptions into spoken language, facilitating seamless accessibility for visually impaired individuals or those preferring auditory information presentation. To further enhance the system's capabilities, we integrate camera access, enabling real-time image-to- speech conversion. Users can capture images directly through their device's camera and receive instant spoken descriptions. Moreover, we introduce a reinforcement learning (RL) model to continually refine the accuracy and relevance of generated text, improving the overall performance of the system.

VIII. References

[1] 'Dr.Sujatha. K', 'Shruti P. Pati', "Text and Speech Recognition for Visually Impaired People using Optical Character Recognition (OCR), text-to-speech synthesizer (TTS)", 'IJCRT | Volume 9', (2021).

[2] 'Liu ', "Evaluating Image-to-Speech Systems with Automatic and Human Evaluation Metrics" , Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI-23)(2023).

[3] "Improving Preprocessing for Image Captioning with Text-Guided Image Generation" (Li et al., 2023), Proceedings of the International Conference on Computer Vision (ICCV23).

[4]'Aman Raj Singh', 'Prof. Diwakar Bhardwaj', 'Mridal Dixit', 'Lalit kumar' ,"An Integrated Model for Text to Text, Image to Text and Audio to Text Linguistic Conversion using Machine Learning Approach" (Li et al., 2023), 2023 6th International Conference on Information Systems and Computer Networks (ISCON) GLA University.

[5]'M.Pandu Babu1', 'Anitha G2 ,"OCR Based Image Text to Speech Conversion using KNearest Neighbors and Comparing with Fuzzy K-Means Clustering Algorithm", 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI).

[6]'Mr. Padmavathi P', 'Bunny Bharadwaj Mahadas' ,'Shyam Sundhar Kalluri','Praveen Devarapu','Sowndarya Lakshmi Bandi',"Optical Character Recognition and Text to Speech Generation System using Machine Learning ", 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)