Vision Assist: AI-Powered Real-Time Image Captioning for the Visually Impaired

Mrs. N. Sree Divya [1], Avusula Bhavana[2], Vanathadupula Ushasri[3]
[1]*AssistantProfessor, Mahatma Gandhi Institute of Technology*
[2,3]*UG Student, Mahatma Gandhi Institute of Technology*

Abstract: Recent advancements in image captioning technology have significantly improved the lives of people with visual impairments, promoting social inclusivity. Using computer vision and natural language processing, images become more accessible and understandable through textual descriptions. Notable progress has been made in developing photo captioning systems specifically for visually impaired users. However, challenges remain, such as ensuring the accuracy of automated captions and managing images with multiple objects or scenes. This study introduces a pioneering architecture for real-time image captioning based on a VGG16-LSTM deep learning model, supported by computer vision. The system has been built and implemented on a Raspberry Pi 4B single-board computer with GPU capabilities. This setup enables the automatic generation of suitable captions for images taken in real-time with a NoIR camera module, making it a convenient and portable solution for visually impaired individuals. The performance of the VGG16-LSTM model is assessed through extensive tests involving both sighted and visually impaired participants in various environments. The results reveal that the proposed system functions effectively, producing accurate and contextually relevant real-time captions. User feedback indicates a notable enhancement in understanding visual content, thereby aiding the mobility and interaction of visually impaired individuals within their surroundings. Multiple datasets were utilized, including Flick8k, Flickr30k, VizWiz captioning, and a custom dataset, for the training, validation, and testing of the model.

Keywords: image captioning technology, visual impairments, social inclusivity, computer vision, natural language processing (NLP), textual descriptions, photo captioning, accuracy, real-time image captioning, VGG16-LSTM deep learning model, portable solutions, automatic generation, extensive testing, contextually relevant captions.

INTRODUCTION

Visually impaired individuals often struggle to understand and engage with visual content like images and videos, hindering their ability to navigate effectively. Image captioning technology offers a vital solution by providing text descriptions of visual content, helping visually impaired users understand and interact with their environment. This enhances their independence and fosters inclusivity. Developing automated image captioning systems that generate accurate, meaningful, and contextually relevant descriptions is essential to bridge this accessibility gap. Recent advancements in deep learning, particularly in computer vision and natural language processing (NLP), have enabled the creation of such systems. Using platforms like Raspberry Pi 4B, these systems can be portable, cost-effective, and operable offline, further helping visually impaired users. This paper aims to present a deep learning-based image captioning system tailored for the visually impaired, focusing on generating precise and meaningful captions to support their daily lives.

A. *Problem Statement.*

Visually impaired individuals face significant barriers when trying to understand and interact with visual content like images and videos due to a lack of accessible interpretation tools. This limitation affects their ability to navigate and access essential information in education, communication, and independent living. While some image captioning technologies are available, many rely on internet

connectivity, have substantial computational demands, or are not specifically designed for visually impaired users. There is an urgent need for a portable, user-friendly automated image captioning system capable of generating real-time, meaningful captions. By leveraging advances in deep learning, computer vision, and NLP, such a system can transform accessibility. Integrating these technologies into a compact, affordable platform like Raspberry Pi 4B can provide visually impaired individuals with a reliable tool for understanding and engaging with their surroundings.

B. Existing System

Current systems aiding visually impaired individuals combine human input and advanced AI for effective solutions. These systems address various aspects of visual accessibility. VizWiz empowers visually impaired users to upload images and seek specific answers, exploiting crowd-sourced inputs for personalized responses. By concentrating on user-specific inquiries, VizWiz enhances utility for tasks such as object identification, text reading, or image comprehension. Be My Eyes connects visually impaired users with sighted volunteers via live video calls, where volunteers provide immediate visual assistance for tasks like reading labels, identifying objects, or navigating new environments. This human-centric method prioritizes interaction, offering detailed and empathetic aid.

Meanwhile, Google Cloud Vision API leverages AI-driven image analysis, using machine learning for tasks like captioning, object detection, text recognition, and scene identification, at high accuracy. Its automation and scalability enable rapid processing of vast image volumes, making it valuable for developers creating assistive applications for the visually impaired. These existing systems illustrate the strengths of human and AI solutions. Human-centered platforms like VizWiz and Be My Eyes deliver contextual assistance, while AI-based tools like Google Cloud Vision API offer efficiency and precision, catering to the diverse needs of the visually impaired community.

LITERATURE SURVEY

The proposed cloud-based system for the visually impaired utilizes AI, NLP, and uniquely trained models to offer real-time help via a gesture-controlled mobile application. Key features encompass image captioning using CNN for object recognition and spatial analysis, Optical Character Recognition (OCR) for reading text, multilingual voice support via Text-to-Speech (TTS), real-time location guidance with Maps API, and a custom Vision API for object and text identification. However, the system's limitations include a heavy dependency on internet connectivity, rendering it ineffective in areas with weak or no network access. It also demands high computational power, making it less suitable for low-power devices, and does not completely utilize edge devices like Raspberry Pi for offline functionality. Additionally, dependence on cloud computing can introduce latency, potentially affecting the system's real-time capabilities.[1]

The proposed wearable device, resembling smart glasses, is specifically created to aid blind individuals by providing real-time descriptions of their surroundings. It combines advanced deep learning models, including Convolutional Neural Networks (CNN) for image feature extraction and Recurrent Neural Networks (RNN) for generating meaningful, context-aware captions. The device captures images through an embedded camera, processes them instantly, and delivers audio descriptions via an integrated speaker or headphones, enabling users to effectively understand their environment. While the device shows promising performance in controlled tests, further assessment in diverse real-world environments is necessary to ensure robustness against varying conditions such as lighting, movement, and complexity.[2]

This paper introduces a smartphone app aimed at assisting visually impaired users through object detection and image captioning technologies to provide real-time descriptions of their environment. Leveraging the smartphone camera to capture images, advanced algorithms analyze the visual content to identify objects and generate descriptive captions, which are subsequently converted into speech, allowing users to hear detailed information about their surroundings. While the app operates effectively on smartphones, its performance is constrained by the device's processing power, which may lead to delays or reduced accuracy in complex or fast-paced environments.[3]

The proposed system utilizes a two-layer transformer architecture integrated with a visual attention mechanism to produce real-time captions for images, deployed on a Raspberry Pi 4B to aid visually impaired users. The model processes visual data captured by a camera, generating descriptive captions that are then transformed into speech, providing immediate auditory feedback about the surroundings. Although the system demonstrates a promising approach to real-time image captioning, its performance heavily relies on the quality and diversity of labeled datasets used for training the transformer model. Limited, biased, or unrepresentative datasets may result in inaccurate or incomplete captions, especially in complex or unfamiliar environments.

This study employs deep learning models, particularly transformers, to create real-time image captions for blind users, focusing on delivering accurate and detailed descriptions of their surroundings. The system prioritizes text reading, such as interpreting signs or labels, over comprehensive scene-based descriptions. Therefore, while it excels at reading text, it may not provide intricate descriptions of complex visual scenes or environments.[4]

This paper introduces a system aimed at helping visually impaired people read text from public areas in real-time. Using computer vision and Optical Character Recognition (OCR), the system captures and processes text from different settings like street signs, menus, and labels. Once the text is identified, it is converted to speech, giving users instant audio feedback to assist them in navigating their environment. While effective for reading text, it does not offer detailed scene descriptions, which limits its effectiveness in dynamic or unfamiliar locations.[5]

PROPOSED SYSTEM

A. Architecture of Proposed System:

The proposed system emphasizes smart, real-time image captioning using advanced deep-learning models to analyze and interpret visual data. It produces clear descriptions of surroundings, objects, or scenes instantly. Unlike many current solutions, it functions offline via a Raspberry Pi 4B, making it viable in remote areas or regions with poor network coverage. The system utilizes efficient models like VGG16 for key image feature extraction and LSTM (Long Short-Term Memory) networks to transform these features into coherent, contextually accurate captions. This combination ensures fast and precise performance, even in complex settings. Once the captions are generated, they are instantly converted into natural-sounding speech through a Text-to-Speech (TTS) engine, enabling users to receive auditory information without needing to read. Crafted with portability and lightweight materials, the system is easy to transport and suitable for various environments, from public venues to indoor spaces. Moreover, the system is trained on diverse datasets, allowing it to adapt effectively to real-world scenarios, including different lighting conditions, intricate objects, and both crowded and open spaces.

B. *Advantages of Proposed System:*

- Offline operation
- Rapid, precise image captioning

- Real-time audio feedback for users with visual impairments
- Portable for on-the-go usage
- Adjustable to varying lighting and environments

SYSTEM DESIGN AND METHODOLOGY

A. *Design:*

This technology attempts to give visually challenged people real-time captions for photos using deep learning and computer vision. It consists of a number of components that cooperate to record, process, and characterize the user's environment. A NoIR camera at the device's center continuously captures images of the user's environment. To make sure these photos are compatible with our deep-learning model, they are scaled to 224x224 pixels and modified to match the RGB color standard.

The photos are processed using the VGG16 model, a kind of deep convolutional neural network (CNN) that, after being scaled, extracts important features from the images. These features are used to produce a feature vector that highlights the key elements of the picture. The LSTM (Long Short-Term Memory) network starts predicting the words that will make up a cohesive sentence using this feature vector. A SoftMax layer helps choose the best word at each step, and the process keeps going until it reaches a "End" token. To enable the visually impaired user to hear what is happening around them, the LSTM transforms the final caption into voice using a Text-to-voice (TTS) module such as gTTS. Because the entire arrangement is powered by a Raspberry Pi 4B, it is portable and operates offline. Because of its architecture, which enables real-time use without an internet connection, it is an easy-to-use option for people with visual impairments.

B. *Modules:*

1. Image Processing Module

This module handles the taking and preprocessing of images. We use a NoIR camera to obtain real-time imagery from our surroundings. The images are resized to 224x224 so they match the input size of the deep learning model, but they still retain the RGB color format to preserve all the details. Standardizing images helps to extract features efficiently.

2. Caption Generation Module

In this step, the processed images are passed through the VGG16 model (a convolutional neural network (CNN) meant for image recognition that extracts relevant features of the input image). Which in turn these features then converted to a feature vector which served as the input to the LSTM (Long Short-Term Memory) network. These features are passed into an LSTM to generate a semantic text description of the image. It generates the words one by one, by employing a SoftMax layer each time to find the best word, and all these words together make a complete and meaningful sentence.

3. Audio Output Module

A TTS module like gTTS converts this generated text caption to speech. That assists the system provide real-time audio feedback so that users can perceive their environment through words. Finished outputs with audio can be heard through speakers and headphones for easy listening.

Everything runs on a Raspberry Pi 4B, so it's portable and can work offline. This setup with three modules allows it to run smoothly with real-time image descriptions, and without needing an internet connection.

D. Algorithms:

1. VGG16:

VGG16 is a CNN architecture used to recognize images and extract important features from them. Its architecture contains 16 layers (including convolutional, pooling, and fully connected layers). Single image is taken by us, then it gets resized to 224x224px and goes through several convolution layers.

Small filters here detect patterns such as edges, textures and objects. It works to reduce the size of

the images while preserving important features. It down samples the image to ensure that the key elements are encoded in the image and these features are then processed in fully connected layers to produce a feature vector of the image that describes its salient characteristics. This feature vector is then passed through into the LSTM model to enable the caption generation.

VGG16 is a popular deep learning model due to the accuracy it affords, the extraction of deeper features than its predecessors, and the availability of pre-trained models on large datasets such as ImageNet that help increase performance while reducing training time.

2. LSTM:

LSTM (Long Short-Term Memory) is a special kind of RNN that is capable of learning long-term dependencies (sequential data generation). Here we use LSTM in this project to translate the extracted features of images into appropriate textual descriptions. Unlike regular RNNs, LSTMs can manage long-term dependencies effectively, which is perfect for creating structured captions. The feature vector from VGG16 goes into the LSTM model, which processes it step-by-step, predicting each word in sequence to make a clear caption. At each point, a SoftMax function picks the most likely word from the vocabulary. The model continues predicting words one at a time, using the context of previously predicted words until it hits an 'End' token, marking the end of the caption.

CONCLUSION

This research provides an in-depth analysis of real-time image captioning methods for aiding visually impaired individuals. It centers on the performance of deep learning models like VGG16 and LSTM. By merging computer vision with natural language processing, the suggested system creates contextually relevant image descriptions and transforms them into speech via Text-to-Speech (TTS) technology. The use of Raspberry Pi 4B allows for offline use, eliminating the need for cloud processing while ensuring real-time effectiveness.

The study emphasizes the importance of combining CNN-based feature extraction with LSTM-based caption generation to improve the precision and usability of assistive technologies. Nonetheless, issues such as recognition limits, caption diversity, and real-world adaptability remain points for future research. Upcoming work should target advancements in object detection, enhancement of caption relevance, and improvement in speech synthesis to elevate user experience. By evolving these techniques, the study aspires to support the advancement of more accessible and efficient assistance systems for visually impaired individuals.

ACKNOWLEDGMENT

This survey of the project would not have been possible without the guidance and help of various people, and their encouragement and contribution was instrumental in this work.

We would like to thank Prof. G. Chandramohan Reddy, Principal of MGIT and Dr. D. Vijaya Lakshmi, Professor and Head of the Department of IT, MGIT, for providing us with excellent infrastructure and environmental support for completing our project.

Most importantly we would like to express our gratitude to our project guide Dr. N. Sree Divya, Assistant Professor, Department of IT, MGIT, we are very thankful for her continuous support, valuable suggestions and great patience throughout the project. She has been pivotal to the success of our work, both with her expertise and her invaluable feedback.

Last but not least We would like to thank our project coordinator Dr. N. Sree Divya Assistant Professor, Department of IT, MGIT who has been a constant source of guidance and encouragement without which this project would not have been possible.

REFERENCES

[1] P. Khan Image Captioning for the Visually Impaired and Blind: A Recipe for Low Resource Languages Batyr Arystanbekovi Aur Kuzdenovi Shakhizat Nurgaliyev), and Huseyin Ataka

[2] A transformer based real-time photo captioning framework for visually impaired people with visual attention Abubeker Kiliyanal Muhammed KunjuS. Baskar Sherin Zafar 3Bushar AR4 Rinesh S5 Shafeena Karim Al

[3] A Realtime portable and accessible aiding system for the blind- a cloud based approachS Venkat Ragavani & A. H. Tannl & S. Yogeeshward B. S. Vishwath Kumar &S. Sedana Reka

[4] Audio Assistance for Visually Impaired Using Image Captioning Krmal Tule, Krishna Paril, Manas Yeole, Shrenik Shingi, Dr. Rashmi Phalnikar

[5] Generalized Image Captioning Fix Multilingual SupportSuhyun Cho land Hayoung Ois "IEEE Access, vol. 9, pp. 37622-37655, 2021.