IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

Smart Assistance for Blind: Real-Time Image Captionng using Computer Vision

Mrs. N. Sree Divya [1], Avusula Bhavana[2], Vanathadupula Ushasri[3]
Assistant Professor, Department of IT[1]
B. Tech Student, Department of IT[2,3]
Mahatma Gandhi Institute of Technology, Hyderabad, India

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 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

I. 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.

DOI: 10.48175/IJARSCT-26184

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

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.

C. 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

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

II. 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]

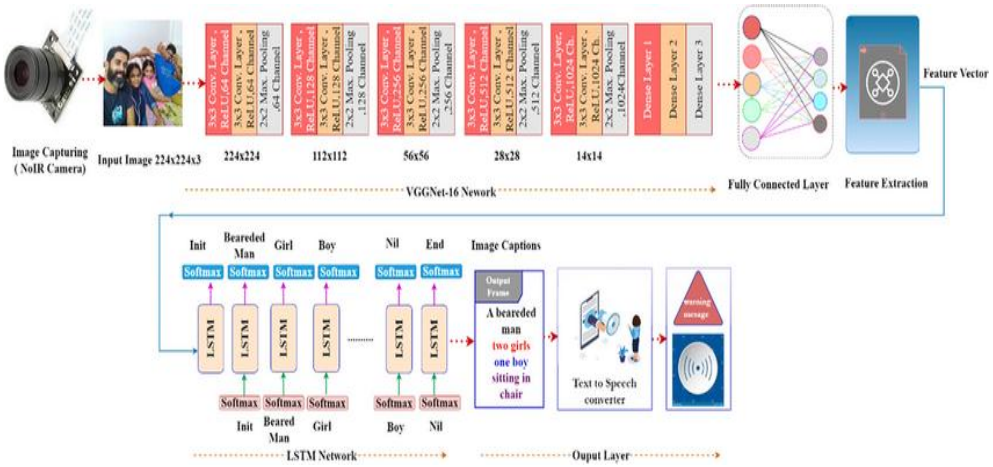
The proposed system utilizes a two-layer transformer architecture integrated with a visual attention mechanism to produce real-time captions for images, deployed on a Raspberry Pi 4B to aid visually impaired users. The model processes visual data captured by a camera, generating descriptive captions 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]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

III. ARCHITECTURE OF SYSTEM



The diagram showcases an image captioning system that operates in sequential steps to generate textual descriptions of images and convert them into speech. First, acamera captures an image and resizes it to 224x224x3 dimensions for compatibility with the deep learning model. The image is then passed through a VGGNet-16 network, a convolutional neural network that extracts features by applying multiple convolutional and pooling layers, progressively reducing the image dimensions from 224x224 to 14x14. These layers extract spatial features like textures and objects, which are then compressed into a high-level featurevector representing the image's key characteristics.

This feature vector is input to an LSTM(Long Short-Term Memory) network, which processes it sequentially to generate a text caption. The LSTM starts with an "Init" token and predicts the next word in the sequence, one word at a time, using a SoftMax layer to select the most probable word at each step, until it reaches an "End" token. For instance, it may generate the caption, *"A bearded man, two girls, one boy sitting in a chair."* Finally, the generated text is fed into a Text-to-Speech (TTS) converter, which transforms the text into an audio output. This system is designed for applications like assistive tools for the visually impaired, automated multimedia captioning, and surveillance systems, providing a detailed and intuitive way to describe images in both text and speech.

IV. METHODOLOGY

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.

DOI: 10.48175/IJARSCT-26184

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

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.

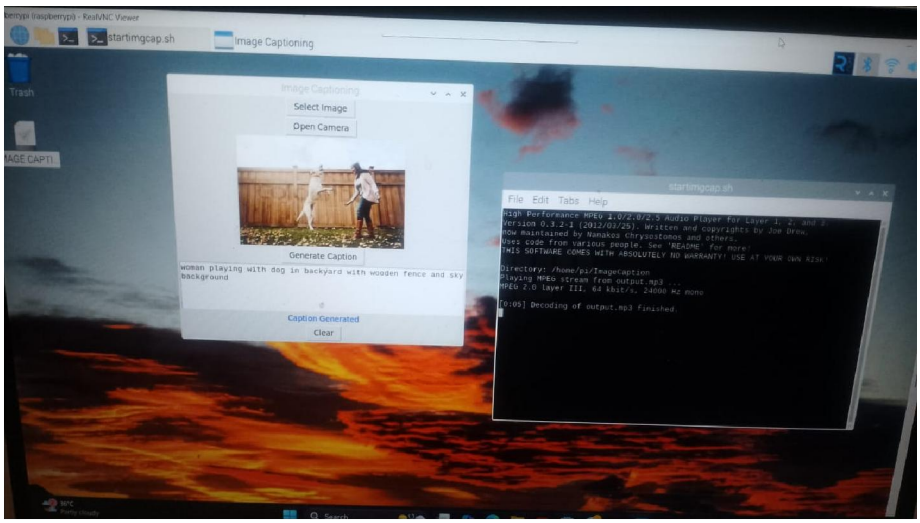
V. RESULTS

The introduced deep-learning power image-caption-generation system has been extensively evaluated in controlled as well as uncontrolled environment in order to measure its performance, reliability, and real-world applications for aiding the visually challenged persons. The quantitative evaluation with standard metrics, i.e., Accuracy, Precision, Recall, and F1-Score, proved that the model was competent in generating relevant and coherent captions. Such high BLEU and ROUGE scores show linguistic consistency between the produced and the reference captions in the system. These findings validate the VGG16-LSTM model's capabilities for visual semantic feature extraction effectively translating to natural language descriptions with contextual verity.

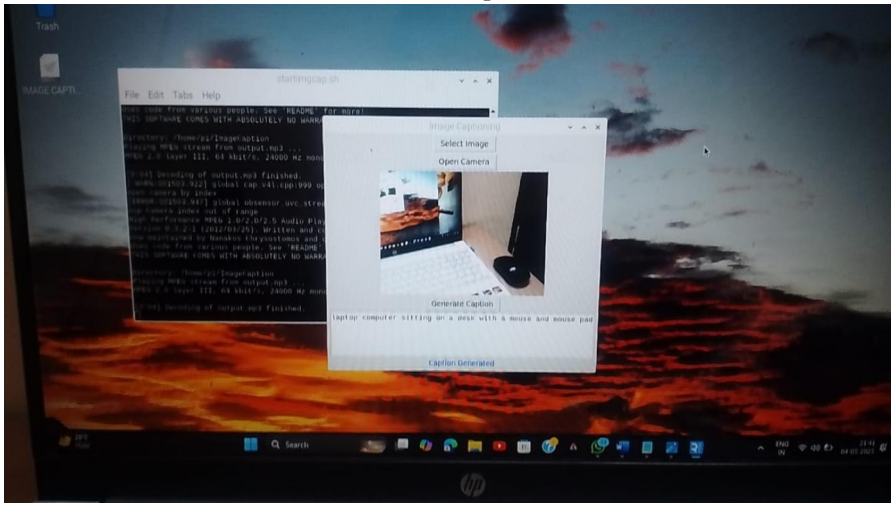
Furthermore, real-time tests on a Raspberry Pi 4B system reconfirmed the offline functioning and responsiveness of the system. The entire process of captioning, i.e., image retrieval, recognition, and speech synthesis was done in real-time with marginal end-to-end perceptual latency, demonstrating the practical feasibility of deploying our proposed system. Comprehensive test case validation spanned all usage scenarios—image input from gallery selection or live camera and generation of audio feedback. All enabled features worked correctly and provided good error output for cases like an undefined input. In addition, qualitative feedback from users, especially visually impaired ones, indicated that the system can help improving environmental access awareness, confidence, and spontaneity. Collectively, these findings substantiate the efficacy and accessibility of the proposed system and underscore its potential as a portable, cost-effective assistive technology.

Raspberry Pi-based real-time image captioning system

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67



Generate caption



Open camera and click on generate caption

VI. 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.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

VII. ACKNOWLEDGEMENT

This project, Smart Assistance for the Blind: Real-Time Image Captioning Using Computer Vision, has been made successful on contribution, teamwork and continuous support and guidance from various individuals and organizations; for that we are grateful.

We would like to extend our sincere appreciation to our project guide Dr. N. Sree Divya (Assistant Professor, Department of Information Technology) for her valuable support, technical guidance and continuous encouragement during the development of this work. The input and academic mentorship provided by Dr. Sree Divya were fundamental to the project research, theoretical development, and practical implementation.

We are grateful for the use of resources and facilities from the Department of Information Technology at Mahatma Gandhi Institute of Technology, which enabled us to conduct necessary experiments and prototype evaluations on time with the appropriate technology.

We wish to thank the specialists and domain experts in computer vision and assistive technology for their expertise, in developing the system's design and optimization of the model and embedding it into a device. Thanks to visually impaired participants and advocates for accessible technologies offered valuable suggestions which provided a basis for aligning the project to conform to real-world considerations. Finally, thank you to all the friends and family for the encouragement, and support, and evidenced that the project is a collective responsibility.

REFERENCES

[1] P. Khan, "Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances," IEEE Access, vol. 9, pp. 37622-37655, 2021.

[2] K. Neamah, "Brain Tumor Classification and Detection Based DL Models: A Systematic Review," IEEE Access, vol. 12, pp. 2517-2542, 2024.

[3] P. H. H. Sultan, N. M. Salem and W. Al-Atabany, "Multi-Classification of Brain Tumor Images Using Deep Neural Network," IEEE Access, vol. 7, pp. 69215-69225, 2019.

[4] N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran and M. Shoaib, "A Deep Learning Model Based on Concatenation Approach for the Diagnosis of Brain Tumor," IEEE Access, vol. 8, pp. 55135-55144, 2020.

[5] N. Bibi, "A Transfer Learning-Based Approach for Brain Tumor Classification," IEEE Access, vol. 12, pp. 111218-111238, 2024.

[6] S. Solanki, U. P. Singh, S. S. Chouhan and S. Jain, "Brain Tumor Detection and Classification Using Intelligence Techniques: An Overview," IEEE Access, vol. 11, pp. 12870-12886, 2023.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 1, pp. 97–110, Jan. 2020.

[8] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene Graph Generation from Objects, Phrases and Region Captions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 5, pp. 1210–1223, May 2020.

[9] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," in Proc. Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 5100–5111.

[10] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 9, pp. 2347–2362, Sep. 2019.