International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Conversational Image Recognition Chatbot: A Multimodal AI System for Interactive Image Understanding

Nikhil C[1], Sai Saran Reddy S[2], Jagadeeshwar Raju S[3], Amon Limbu[4], Kiran J[5]

[1,2,3,4]Student, School of CSE, Reva University, Bengaluru, India
[5]Assistant Professor, School of CSE, Reva University, Bengaluru, India

Abstract
A Conversational Image Recognition Chatbot is an intelligent multimodal system that enables users to interact with images using natural language conversations. By integrating computer vision and natural language processing, the system can analyze uploaded images, identify objects and scenes, and respond to user queries in real time. Traditional image recognition systems provide static outputs such as labels or captions, whereas conversational systems enable context-aware, interactive understanding.

Keywords: Conversational AI, Image Recognition, Multi- modal Systems, Computer Vision, Natural Language Processing

I. INTRODUCTION

The rapid advancement of artificial intelligence has enabled machines to interpret and communicate about visual information with increasing sophistication. Traditional image recognition systems focus primarily on object detection or classification and provide limited interaction with users. These systems lack the ability to answer follow-up questions or explain visual content in a human-like manner. Conversational Image Recognition Chatbots combine computer vision and natural language processing to allow users to upload images and ask questions about them using natural language. Such systems are increasingly important in applications such as accessibility support for visually impaired users, educational tools, smart assistants, and customer service platforms. models with natural language processing to generate image captions and answer visual questions. Attention mechanisms and vision–language models further enhanced contextual understanding. Recent multimodal models enable conversational-level interaction with images, supporting multi-turn dialogue and semantic reasoning.

R Furthermore, these advanced systems can understand complex visual scenes, relate objects with contextual information, and assist in real-world applications such as autonomous driving, medical imaging, security surveillance, and human–computer interaction, making image recognition an essential component of modern artificial intelligence systems.

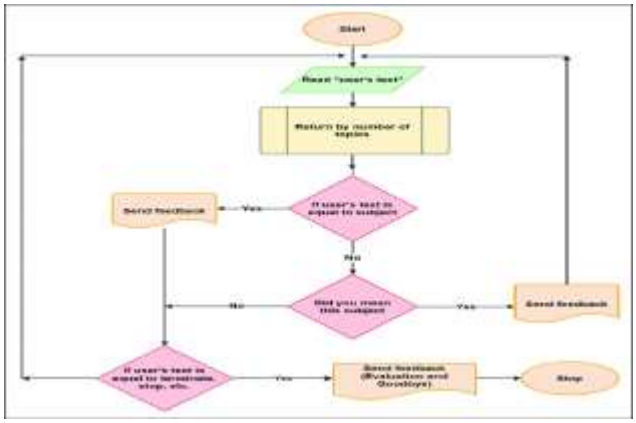

Fig. 1 Flowchart of User Text Processing and Feedback Generation

II. RELATED WORK

Early image recognition systems relied on handcrafted features and traditional machine learning techniques, offering limited accuracy and no interaction capability. The introduction of deep convolutional neural networks significantly improved image classification and object detection performance. Subsequent research integrated vision

III. SYSTEM OVERVIEW

The Conversational Image Recognition Chatbot is a web based multimodal system designed to analyze images and respond to user queries conversationally. Users can upload images and interact with the system through a chat interface. The system consists of image preprocessing, image recognition, conversational AI, and context management modules.

The chatbot maintains conversation history to support follow up questions and provide coherent responses. The image preprocessing module prepares the uploaded image by resizing, normalization, and noise reduction to ensure accurate recognition. The processed image is then passed to the image recognition module, which uses deep learning models such as convolutional neural networks to detect objects, scenes, and visual features. The detected visual information is converted into structured data that can be interpreted by the conversational AI module.

The conversational AI component processes user questions using natural language processing techniques and combines them with the extracted visual information to generate meaningful responses. Context management maintains the conversation history, allowing the system to support multi-turn interactions and follow-up questions. The chatbot can explain objects, describe scenes, or answer queries related to the uploaded image. This integration of computer vision and conversational intelligence enhances user experience by enabling intuitive and interactive image understanding.

IV. METHODOLOGY

The project follows a structured development methodology starting with requirement analysis and literature survey. System architecture is designed to integrate frontend, backend, image processing, and conversational modules. Deep learning models are used for image recognition tasks such as object detection and scene understanding. A conversational AI model processes user queries and generates natural language responses based on visual context. The system is tested for accuracy, performance, and usability before deployment.

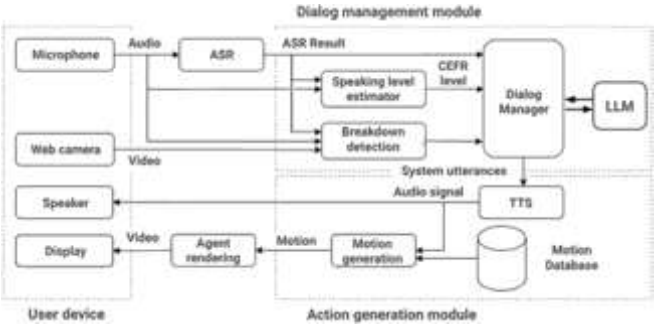

Fig. 2 Architecture of the Multimodal Conversational Dialog Management System

Additionally, data preprocessing and model training are performed using large annotated datasets to enhance prediction accuracy and model reliability. Various evaluation metrics such as precision, recall,

accuracy, and F1-score are used to measure system performance. Optimization techniques and hyperparameter tuning are applied to improve model efficiency. The system is also evaluated under different scenarios to ensure robustness, scalability, and smooth user interaction in practical real-world environments.

V. RESULTS AND DISCUSSION

Future enhancements include multilingual support, voice-based interaction, real-time video analysis, and integration with wearable or assistive devices. Ethical AI considerations such as bias mitigation and uncertainty-aware responses will also be addressed to improve reliability and trustworthiness. Additionally, adaptive learning mechanisms can be introduced to personalize responses based on user preferences and usage patterns. Strengthening data privacy, security, and transparent explanations will further enhance user confidence and responsible deployment.

Moreover, adding edge computing support can help the system work faster and even function in low-connectivity environments. Regular model updates and domain-specific tuning will gradually improve accuracy. Conducting user studies and gathering real feedback will also help refine the design and make the system more practical and user-friendly.

VI. FUTURE WORKS

The implemented system successfully demonstrates interactive image understanding through conversational dialogue. Users can ask descriptive and contextual questions about uploaded images and receive meaningful responses. The chatbot improves accessibility and usability compared to traditional image recognition systems. Its modular architecture allows easy extension and integration of advanced features. Performance evaluation shows effective image recognition and conversational response generation in real time.

In addition, user feedback indicates that the system offers a more engaging and intuitive interaction experience, especially for non-technical users. The response accuracy remains consistent across varied image types, including everyday scenes and complex visual contexts. Efficient processing pipelines ensure low latency during inference, supporting smooth, continuous conversations. The design also allows future enhancements such as multilingual support, personalized context retention, and deployment across web and mobile platforms, thereby broadening its practical applicability.

VII. CONCLUSIONS

The Conversational Image Recognition Chatbot demonstrates the effective integration of computer vision and conversational AI to enable interactive and accessible image understanding. By moving beyond static recognition outputs, the system provides context-aware dialogue and improved user interaction. The project validates the feasibility of multimodal conversational systems and highlights their potential impact in accessibility, education, and assistive technologies.

Through practical testing, the system shows that users can naturally explore image content by asking follow-up questions and receiving clear explanations. This makes the interaction feel more intuitive and engaging compared to conventional tools. The overall results suggest that such chatbots can serve as helpful companions in everyday scenarios, supporting users in learning, exploring visual data, and gaining better insights from images in a simple and conversational way.

It also highlights how combining vision and language can make technology feel more approachable. With further improvements, the system can become a reliable assistant that understands visual content and responds in a natural, user-friendly manner.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the faculty of the **School of Computing and Artificial Intelligence and Machine Learning, REVA University, Bangalore**, for their continuous support and guidance throughout the development of this project. Their valuable insights and encouragement played an important role in completing this work successfully.

The authors also thank their mentors and colleagues for their constructive suggestions and feedback during the design and implementation of the Conversational Image Recognition Chatbot system. Their discussions and technical assistance greatly contributed to improving the quality of this research.

Finally, the authors acknowledge the support of **REVA University** for providing the academic environment, resources, and infrastructure required to carry out this project, , enabling effective research, experimentation, collaboration, and successful completion of this work.

REFERENCES

1. J. Deng et al., “ImageNet: A Large-Scale Hierarchical Image Database,” IEEE, 2009.
2. A. Krizhevsky et al., “ImageNet Classification with Deep Convolutional Neural Networks,” NeurIPS, 2012.
3. O. Vinyals et al., “Show and Tell: A Neural Image Caption Generator,” IEEE, 2015.
4. K. Xu et al., “Show, Attend and Tell,” ICML, 2015.
5. A. Radford et al., “Learning Transferable Visual Models from Natural Language Supervision,” OpenAI, 2021.
6. J. Li et al., “BLIP: Bootstrapped Language-Image Pretraining,” ICML, 2022.
7. S. Antol et al., “VQA: Visual Question Answering,” ICCV, 2015.
8. Y. Goyal et al., “Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering,” CVPR, 2017.
9. L. H. Li et al., “VisualBERT: A Simple and Performant Baseline for Vision and Language,” arXiv preprint arXiv:1908.03557, 2019.
10. W. Kim et al., “ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision,” ICML, 2021.

11. C. Jia et al., “ALIGN: Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision,” ICML, 2021.