International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026

Peer Reviewed, Referred & Indexed Journal

Transformer-Driven Intelligent Image Understanding and Question Answering System

GOLUKONDA SANDHYA KUMARI

PG Scholar. Department of MCA, DNR College, Bhimavaram, Andhra Pradesh

V.SARALA

(Assistant Professor), Master of Computer Applications, DNR College, Bhimavaram, Andhra Pradesh

ABSTRACT

The rapid advancement of artificial intelligence has significantly transformed the field of image analysis, enabling systems to not only recognize visual content but also interpret and describe it in natural language. This paper presents an intelligent image analysis and Visual Question Answering (VQA) system that integrates deep learning and transformer-based models to provide automated image understanding, caption generation, and user-interactive query responses. The system is designed to handle both medical and general image datasets, making it versatile for multiple real-world applications such as healthcare diagnostics, wildlife recognition, and intelligent assistants.

The proposed system utilizes a convolutional neural network, specifically ResNet50, for image classification tasks. Pretrained models are fine-tuned to categorize images into predefined classes such as medical conditions or animal types. To enhance interpretability, the system incorporates the BLIP (Bootstrapping Language-Image Pretraining) model for generating human-like captions that describe the content of the image. This enables users to gain contextual understanding of visual data without manual inspection.

Furthermore, the system integrates a transformer-based language model (FLAN-T5) to support Visual Question Answering. Users can input queries related to the selected image, and the system generates meaningful answers by leveraging natural language processing capabilities. This combination of vision and language models creates a knowledge-based interactive system that bridges the gap between image recognition and semantic understanding.

A user-friendly graphical interface is developed using Tkinter, allowing seamless interaction with the system. Users can upload images either from local storage or via URL, input questions, and receive results including captions, classifications, and answers in real-time. Additionally, a SQLite database is employed to manage user authentication and store analysis results, ensuring data persistence and usability.

Experimental evaluation demonstrates that the system effectively combines multiple AI techniques to deliver accurate image classification, meaningful captions, and context-aware responses. The integration of deep learning and transformer models enhances the system's performance and adaptability across different domains.

In conclusion, this work highlights the potential of combining computer vision and natural language processing for intelligent image analysis. The proposed system provides a scalable and efficient framework for developing advanced AI applications capable of understanding and interacting with visual data.

Keywords: Image Captioning, Visual Question Answering (VQA), Deep Learning, BLIP Model, ResNet50, Transformer Models, Medical Image Analysis, Natural Language Processing, GUI Application, Knowledge-Based Systems

I. INTRODUCTION

The integration of computer vision and natural language processing has opened new frontiers in artificial intelligence, enabling machines to interpret, analyze, and communicate visual information effectively. Traditional image processing systems were limited to classification and object detection tasks, providing only basic insights into visual data. However, modern applications demand more advanced capabilities, such as generating descriptive captions and answering user queries based on image content. This has led to the development of Visual Question Answering (VQA) systems, which combine vision and language understanding into a unified framework.

Image analysis plays a crucial role in various domains, including healthcare, surveillance, agriculture, and autonomous systems. In the medical field, for instance, accurate image interpretation can assist in disease diagnosis and treatment planning. Similarly, in wildlife monitoring, automated image classification helps in identifying species and tracking biodiversity. Despite these advancements, many existing systems lack the ability to provide contextual explanations or interact with users in a meaningful way.

Recent developments in deep learning have significantly improved the performance of image analysis systems. Convolutional Neural Networks (CNNs), such as ResNet50, have demonstrated remarkable accuracy in image classification tasks by learning hierarchical feature representations. At the same time, transformer-based models have revolutionized natural language processing by enabling machines to understand and generate human-like text. Models like BLIP and FLAN-T5 have further extended these capabilities to multimodal applications, allowing seamless integration of vision and language.

This paper presents an intelligent image analysis system that leverages both CNN and transformer

architectures to provide comprehensive image understanding. The system is capable of classifying images, generating descriptive captions, and answering user queries, making it a complete solution for interactive image analysis. The use of pretrained models ensures high accuracy and reduces the need for extensive training data.

A graphical user interface (GUI) is developed to make the system accessible to non-technical users. The interface allows users to upload images, input queries, and view results in an intuitive manner. Additionally, a database system is integrated to manage user data and store analysis results, enhancing the system's usability and scalability.

The motivation behind this work is to bridge the gap between visual perception and semantic understanding by combining state-of-the-art AI techniques. By integrating image classification, caption generation, and question answering into a single framework, the proposed system aims to provide a more intelligent and interactive approach to image analysis.

II. LITERATURE SURVEY (WITH EXISTING METHODS)

The field of image analysis and Visual Question Answering (VQA) has evolved significantly with advancements in deep learning and multimodal artificial intelligence. Early image analysis systems primarily relied on traditional computer vision techniques such as edge detection, feature extraction, and template matching. While these approaches were effective for simple tasks, they lacked the ability to generalize across complex and diverse datasets.

With the introduction of Convolutional Neural Networks (CNNs), image classification accuracy improved substantially. Architectures such as ResNet, VGG, and Inception became widely adopted due to their ability to learn hierarchical feature representations from large-scale datasets. ResNet, in particular, addressed the vanishing gradient problem through residual connections, enabling deeper networks and improved performance in image recognition tasks.

In parallel, image captioning systems emerged to bridge the gap between vision and language. Early methods combined CNNs for feature extraction with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for sequence generation. These models generated basic textual descriptions of images but often lacked contextual understanding and semantic richness.

The development of transformer-based architectures marked a significant breakthrough in natural language processing. Models such as BERT, GPT, and T5 demonstrated superior performance in language understanding and generation tasks. These models were later extended to multimodal applications, enabling the integration of visual and textual data.

Visual Question Answering systems combine computer vision and NLP to answer questions based on image content. Early VQA models used CNNs for image encoding and LSTMs for

question encoding, followed by simple fusion techniques. However, these models struggled with complex reasoning and long-range dependencies.

Recent approaches utilize transformer-based models such as Vision Transformers (ViT) and multimodal frameworks like BLIP (Bootstrapping Language-Image Pretraining). These models leverage attention mechanisms to align visual and textual features, resulting in improved performance in captioning and VQA tasks.

Despite these advancements, existing methods often focus on isolated functionalities such as classification, captioning, or question answering. Many systems lack integration, real-time interaction, and user-friendly interfaces. Additionally, limited support for domain-specific applications such as medical image analysis remains a challenge.

The proposed system addresses these gaps by integrating CNN-based classification, transformer-based caption generation, and VQA into a unified framework with an interactive GUI, enabling comprehensive and practical image analysis.

III. EXISTING SYSTEM

Existing image analysis systems are typically designed to perform specific tasks such as image classification, object detection, or caption generation. These systems often operate independently and lack integration with other functionalities. For example, CNN-based models like ResNet and VGG are widely used for classification tasks, while separate models are employed for caption generation and natural language processing.

One major limitation of existing systems is their inability to provide interactive user experiences. Most systems do not allow users to query images or obtain contextual explanations. As a result, users are limited to predefined outputs such as class labels or simple captions, which may not provide sufficient insight into the image content.

Another drawback is the lack of multimodal integration. Traditional systems treat visual and textual data separately, resulting in limited semantic understanding. Early VQA systems attempted to combine these modalities, but they relied on basic fusion techniques and lacked the ability to handle complex queries.

Additionally, many existing systems do not support real-time processing or user-friendly interfaces. They often require technical expertise to operate and are not suitable for general users. The absence of database integration further limits their usability, as analysis results cannot be stored or retrieved efficiently.

In domain-specific applications such as medical imaging, existing systems face additional challenges. They often require large annotated datasets and specialized models, making them less

adaptable to different use cases.

Overall, existing systems suffer from limitations in integration, interactivity, scalability, and usability, highlighting the need for a more comprehensive and intelligent solution.

IV. PROPOSED METHOD

The proposed system introduces an integrated framework for intelligent image analysis and Visual Question Answering (VQA) using deep learning and transformer-based models. Unlike existing systems, this approach combines multiple functionalities into a single unified platform, enabling comprehensive understanding and interaction with visual data.

The system utilizes a pretrained ResNet50 model for image classification. This model extracts high-level features and categorizes images into predefined classes such as medical conditions or animal types. To enhance interpretability, the system incorporates the BLIP model for generating descriptive captions that provide contextual information about the image.

For interactive analysis, the system integrates a transformer-based language model, FLAN-T5, to perform question answering. Users can input queries related to the image, and the system generates meaningful responses based on both visual and textual understanding.

A key feature of the proposed system is its user-friendly graphical interface developed using Tkinter. The interface allows users to upload images from local storage or via URLs, input questions, and view results in real-time. This makes the system accessible to both technical and non-technical users.

The system also includes a SQLite database for user authentication and data storage. This ensures secure access and enables users to save and retrieve analysis results for future reference.

Overall, the proposed system provides a scalable, efficient, and interactive solution for image analysis by integrating computer vision and natural language processing techniques into a single framework.

V. IMPLEMENTATION

The implementation of the proposed system is carried out using Python, leveraging its extensive ecosystem of libraries for machine learning, image processing, and GUI development. The system is designed in a modular manner to ensure scalability, maintainability, and efficient integration of different components.

The first module involves database initialization using SQLite. A local database is created to store user credentials and analysis results. Two tables are defined: one for user authentication and another for storing image analysis data, including captions, classifications, and query responses. This ensures persistent storage and secure access control.

The user authentication module provides a login interface built using Tkinter. Users are required to enter valid credentials to access the main application. The system verifies the credentials against the database and grants access upon successful authentication.

The image input module allows users to upload images either from local storage or via a URL. The selected image is displayed on the GUI using the PIL and ImageTk libraries. This module ensures flexibility in data input and enhances user experience.

The preprocessing module prepares the image for model inference. It includes resizing, normalization, and data augmentation techniques such as random rotation and flipping. These transformations ensure compatibility with the ResNet50 model and improve generalization.

The classification module uses a pretrained ResNet50 model for image classification. The model is fine-tuned for specific categories and predicts the class label based on input features. Softmax probabilities are used to determine the most likely class.

The caption generation module employs the BLIP model to generate descriptive captions. The model processes the image and produces natural language descriptions that summarize its content.

The VQA module integrates the FLAN-T5 transformer model. User queries are tokenized and processed by the model to generate context-aware answers. This enables interactive communication between the user and the system.

The GUI module integrates all components into a cohesive interface. Users can perform all operations, including image selection, query input, and result visualization, within a single window.

Finally, the system stores all results in the database, ensuring data persistence and enabling future analysis.

VI. ALGORITHMS

The proposed system employs three primary algorithms for image classification, caption generation, and question answering.

The first algorithm is the ResNet50-based classification algorithm. It processes input images

through multiple convolutional layers to extract hierarchical features. Residual connections are used to prevent vanishing gradients, enabling deeper network architectures. The final fully connected layer produces class probabilities, and the class with the highest probability is selected as the output.

The second algorithm is the BLIP-based caption generation algorithm. This model combines vision and language understanding by encoding image features and generating textual descriptions using a transformer decoder. The model uses attention mechanisms to focus on relevant regions of the image while generating captions.

The third algorithm is the transformer-based VQA algorithm using FLAN-T5. The model takes a textual query as input and generates a response based on learned language representations. It uses encoder-decoder architecture and attention mechanisms to process input sequences and produce meaningful outputs.

These algorithms work together to provide a comprehensive image analysis system. The classification algorithm identifies the category of the image, the captioning algorithm describes its content, and the VQA algorithm enables interactive querying.

VII. SYSTEM DESIGN

The system design follows a modular architecture that integrates multiple components for efficient image analysis and interaction.

The first module is the user authentication module, which manages login functionality. It ensures secure access to the system and prevents unauthorized usage. User credentials are stored in a SQLite database and validated during login.

The second module is the image input module, which allows users to upload images from local storage or via URLs. This module provides flexibility and enhances usability.

The preprocessing module prepares the input image for model inference. It includes resizing, normalization, and augmentation to ensure compatibility with deep learning models.

The classification module uses ResNet50 to categorize images. It processes the preprocessed image and outputs the predicted class label.

The caption generation module uses the BLIP model to generate descriptive captions. This module enhances interpretability by providing contextual information about the image.

The VQA module processes user queries using the FLAN-T5 model. It generates responses based on the image context and user input, enabling interactive analysis.

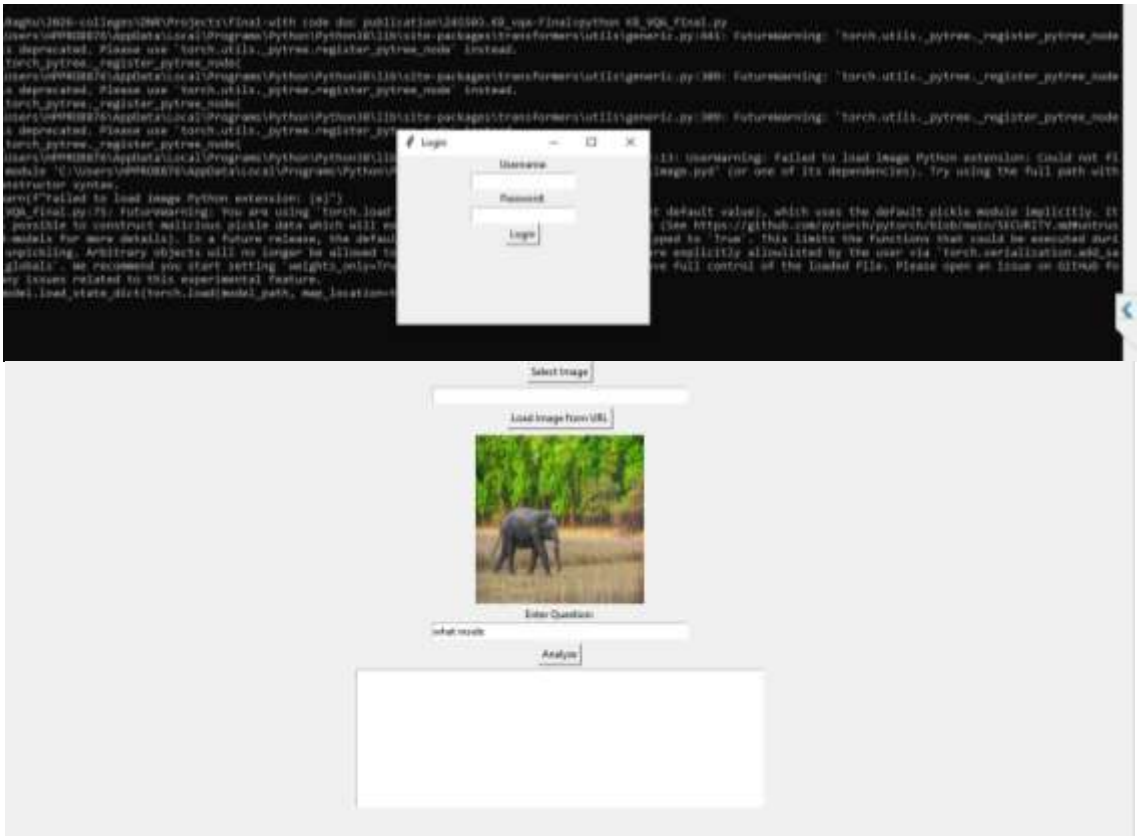
The GUI module integrates all functionalities into a single interface. It provides buttons for image selection, input fields for queries, and display areas for results.

The database module stores user data and analysis results. This ensures data persistence and allows users to retrieve past records.

The overall system follows a pipeline architecture:
User Input → Image Processing → Classification → Caption Generation → Question Answering → Output Display → Database Storage

This design ensures efficient data flow, modularity, and scalability.

SYSTEM DESIGN IMAGES



VIII. CONCLUSION

The rapid evolution of artificial intelligence has enabled the development of advanced systems capable of understanding and interpreting visual data in a meaningful way. This work presented an intelligent image analysis and Visual Question Answering (VQA) system that integrates deep

learning and transformer-based models to deliver a comprehensive and interactive solution. By combining image classification, caption generation, and question answering into a unified framework, the system overcomes the limitations of traditional image processing approaches.

The use of a ResNet50-based convolutional neural network ensures accurate image classification through effective feature extraction, while the BLIP model enhances interpretability by generating context-aware captions. Furthermore, the integration of the FLAN-T5 transformer model enables natural language interaction, allowing users to query images and receive meaningful responses. This multimodal capability significantly improves user engagement and system usability.

The implementation of a graphical user interface using Tkinter makes the system accessible to users with varying levels of technical expertise. Additionally, the incorporation of a SQLite database ensures secure user authentication and efficient storage of analysis results. The modular architecture of the system allows for easy scalability and future enhancements.

Overall, the proposed system demonstrates the potential of combining computer vision and natural language processing to create intelligent, user-friendly applications. It can be effectively applied in domains such as healthcare, education, surveillance, and wildlife monitoring. Future work may focus on improving model accuracy, incorporating real-time video analysis, and deploying the system on cloud platforms for enhanced scalability and accessibility. The integration of more advanced multimodal models could further enhance the system's ability to understand complex visual and textual relationships.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016.

[2] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.

[3] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019.

[4] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, 2020.

[5] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training," International Conference on Machine Learning (ICML), 2022.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep

Convolutional Neural Networks," NeurIPS, 2012.

[7] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," ICLR, 2015.

[8] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, 2015.

[9] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and VQA," CVPR, 2018.

[10] S. Antol et al., "VQA: Visual Question Answering," ICCV, 2015.

[11] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations," EMNLP, 2019.

[12] Z. Lu et al., "VilBERT: Pretraining Task-Agnostic Visiolinguistic Representations," NeurIPS, 2019.

[13] PyTorch Documentation, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2023.

[14] Hugging Face, "Transformers Library Documentation," 2023.

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," ICLR, 2015.