ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

A SEMANTIC WEIGHT ADAPTIVE MODEL BASED ON VISUAL QUESTION ANSWERING

Dr. M. V. NARAYANA[1], NITHIN REDDY[2], RAHUL[3], HARSHA VARDHAN REDDY[4], PRAVEEN [5]

[1]Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— This project presents a multilingual Visual Question Answering (VQA) web application built with Flask, integrating deep learning and NLP techniques. It utilizes the Salesforce BLIP-VQA model, which combines a Vision Transformer and text decoder to answer questions based on uploaded images or videos (up to 60 seconds, given sufficient GPU and memory). Users can interact in various Indian languages like Hindi, Tamil, and Telugu, with a mock translation service converting questions to English for processing and translating answers back to the original language. The system supports secure file uploads, validates inputs, and extracts keyframes from videos to generate context-aware answers. This real-time, multilingual application demonstrates the potential of transformer-based VQA models in accessible and scalable web environments, with room for future integration of services like IndicTrans2. A key highlight of the system is its multilingual support, allowing users to ask questions in a variety of Indian languages such as Hindi, Tamil, Telugu, and others. To achieve this, a mock translation service is integrated to simulate the translation of the user's input question to English (the primary language used by the VQA model) and then translate the model-generated English answer back to the user's original language. This simulates an end to-end multilingual pipeline and demonstrates the feasibility of extending the system with real translation services like in dictrans.

Index Terms – Visual Question Answering (VQA), BLIP-VQA, Vision Transformer (ViT), Natural Language Processing (NLP), multilingual question answering, Flask web application, image understanding, video question answering, transformer models, computer vision, multilingual AI, Indian languages.

I. INTRODUCTION

Visual Question Answering (VQA) is an important research area in Artificial Intelligence that combines Computer Vision (CV) and Natural Language Processing (NLP). The main goal of VQA is to enable computers to understand an image and answer questions related to that image in natural language. Unlike traditional image processing tasks such as image classification or object detection, VQA requires both visual understanding and language understanding at the same time.In a VQA system, the model analyzes the image content and the textual question provided by the user. The system then processes these two types of information and generates an appropriate answer. The questions can be open-ended (for example, "Where was this photo taken?") or closed-ended (for example, "Is there a dog in the image?"). Therefore, the system must learn complex relationships between objects in images and the meaning of the question. However, many existing VQA models face challenges in understanding the deep semantic meaning of questions and effectively combining image and text features. To solve this problem, researchers proposed a Semantic Weight Adaptive Model Network (SWAMN). This model improves the reasoning ability of the system by dynamically extracting important information from questions and guiding the fusion of multimodal features. The proposed approach integrates image features and textual features more effectively, allowing the system to produce more accurate answers. Experimental results on the VQA v2.0 dataset show that the model achieves high accuracy and improves the performance of visual question answering systems. A Multimodal Framework for Visual Question Answering and Multilingual Translation is an advanced Artificial Intelligence (AI) system that combines computer vision and natural language processing (NLP) to understand visual content and communicate information in multiple languages. Traditional AI systems often process either images or text independently. In contrast, a multimodal framework integrates both types of data, enabling the system to interpret images, answer questions related to them, and

translate the responses into different languages. Visual Question Answering (VQA) allows users to ask questions about an image, such as identifying objects, describing scenes, recognizing activities, or extracting text from images. The system analyzes the visual content using deep learning models and generates accurate answers based on the image and the user's question. Multilingual translation extends this capability by translating the generated answers into various languages while preserving their meaning and context. This feature improves accessibility, supports global communication, and enables users from different linguistic backgrounds to interact with the system effectively

II. RELATED WORK

A. Deep Modular Co-Attention Networks for Visual Question Answering

This paper introduces the Deep Modular Co-Attention Network (MCAN), a novel deep learning architecture designed for the task of Visual Question Answering (VQA). Unlike conventional attention models that only focus on one modality at a time, MCAN simultaneously models fine-grained interactions between vision and language through modularized co-attention layers. The architecture stacks multiple co attention modules that progressively refine the alignment between image regions and question words, enabling the model to capture deeper semantic relations and contextual cues. By leveraging residual connections and gating mechanisms, MCAN avoids information loss during multiple reasoning steps, ensuring more robust feature representations. Extensive experiments on benchmark datasets such as VQA v2.0 demonstrate that MCAN achieves state-of-the-art performance, outperforming previous attention-based methods. The paper highlights the importance of multi-level reasoning across modalities and establishes MCAN as a strong foundation for future research in VQA and related vision-language tasks.

B. Multimodal Unified Attention Networks for Vision-and-Language Interactions

This work presents the Multimodal Unified Attention Network (MUAN), a general-purpose framework that unifies different types of attention mechanisms for improved vision-language interaction tasks such as Visual Question Answering (VQA), image captioning, and cross-modal retrieval. MUAN addresses the limitations of previous attention models, which often treated visual and textual modalities separately or relied on shallow fusion techniques. The proposed

network introduces a unified attention mechanism that integrates self-attention and co-attention modules into a single architecture, enabling both intra-modality and inter-modality reasoning. This design ensures that textual and visual features are not only aligned but also enriched with contextual dependencies. The authors further emphasize scalability and modularity, making MUAN adaptable across multiple vision-language applications. Experimental evaluations demonstrate that MUAN achieves competitive or superior results compared to specialized models, validating the effectiveness of unified attention for broad multimodal learning. This paper significantly contributes to the understanding of how shared attention architectures can generalize across diverse multimodal problems.

C. *Multimodal Encoder–Decoder Attention Networks for Visual Question Answering*

In this paper, the authors propose a Multimodal Encoder–Decoder Attention Network (MEDAN) for the Visual Question Answering (VQA) task. Unlike traditional models that rely on shallow feature fusion, MEDAN adopts an encoder decoder structure where both image and text modalities are encoded into high-level feature spaces and then decoded through attention mechanisms to generate accurate answers. The encoder integrates convolutional features from images with semantic embeddings of questions, while the decoder leverages multi-step attention to iteratively refine the alignment between visual regions and linguistic tokens. This design ensures that the model can perform complex reasoning, handle long textual dependencies, and extract meaningful cross-modal correlations. The authors validate the approach on standard VQA datasets, where MEDAN outperforms several baseline methods, showing improved interpretability and reasoning capabilities. The study underscores the potential of encoder decoder attention frameworks for advancing multimodal AI and highlights how sequence to-sequence principles can be extended beyond text-based applications to vision-language tasks.

D. *An Effective Dense Co-Attention Network for Visual Question Answering*

This paper introduces an Effective Dense Co-Attention Network (DCN) designed to improve performance in Visual Question Answering (VQA). The dense co-attention mechanism extends beyond simple one-to-one feature alignment by densely connecting multiple layers of visual and textual representations, thereby enabling richer

interactions across modalities. Each co-attention layer simultaneously attends to fine-grained image regions and question words, while dense connectivity ensures that information flows across different levels of abstraction. This design allows the model to capture both low-level details and high-level semantic relationships, which are crucial for complex reasoning tasks. The authors emphasize efficiency by proposing optimized attention operations that reduce computational overhead without compromising accuracy. Experimental results on VQA v2.0 and related benchmarks show that DCN significantly improves answer prediction accuracy and interpretability compared to traditional attention networks. The research highlights the effectiveness of dense connectivity in multimodal learning and provides valuable insights for building more robust VQA systems.

III. PROPOSED METHODOLOGY

The core algorithm behind this VQA system is based on the BLIP (Bootstrapping Language-Image Pretraining) model, specifically the Salesforce/blip-vqa-base, which is a state-of-the-art multimodal transformer model. This model is designed to understand and answer natural language questions about images by learning visual and textual information jointly. It comprises two main components: a Vision Transformer (ViT) used to extract high-level semantic features from input images, and a Transformer-based text decoder that interprets the image features in the context of a user's question to generate an appropriate answer. For generating textual answers, the BLIP model utilizes beam search decoding, a common algorithm in sequence generation tasks. Beam search helps generate coherent and high probability answers by exploring multiple candidate sequences at each decoding step and selecting the best one based on cumulative likelihood.

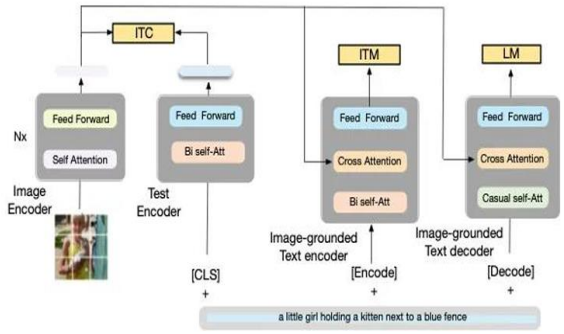


Fig. 1: System Overview

1. File Upload Module

Manages secure upload and storage of image files. Ensures that only supported formats (PNG, JPG, JPEG, GIF, WebP) are accepted. Implements file size validation (16 MB max) and generates unique filenames using

UUIDs. Technologies: Flask, werkzeug. utils. secure_filename, OS file handling.

2. VQA Processing Module

Core module that handles image-question processing. It loads the Salesforce BLIP-VQA model (Blip Processor and Blip For Question Answering) to process visual inputs with natural language queries. If the model fails to load, a demo response generator ensures system continuity. Technologies: Hugging Face Transformers (Blip Processor, Blip For Question Answering), PyTorch, PIL for image preprocessing.

3. Translation Module

Provides multilingual support by translating user queries from Indian languages to English and translating answers back into the original language. Currently implemented as a mock service with predefined word mappings to simulate multilingual processing. Future integration with IndicTrans2 or Google Translate is possible.

4. API and Web Interface Module

Provides RESTful APIs for communication between frontend and backend. Exposes endpoints for file upload (/api/upload), VQA processing (/api/vqa), language retrieval (/api/languages), and file serving (/uploads/). The frontend (index.html) allows users to interact with the system in real time.

5. Error Handling and Logging Module

Ensures robustness by handling exceptions and errors gracefully. Implements logging for debugging and monitoring model loading, file processing, and translation errors. Custom error handlers manage cases such as file-too-large (413), not found (404), and server errors.

6. Security and Validation Module

Protects the system by validating file types, limiting file size, and using a secret key for Flask sessions. Ensures only valid inputs are processed to prevent misuse or server crashes.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed BLIP-VQA-based semantic adaptive model accurately answers questions related to uploaded images and video keyframes while supporting multilingual user queries. The combination of Vision Transformers, semantic understanding, and the multilingual translation pipeline improves contextual accuracy and user accessibility across Indian languages. The Flask-based implementation demonstrates

reliable real-time performance with secure file handling and efficient inference. Overall, the proposed framework provides a scalable, accurate, and user-friendly solution for multilingual Visual Question Answering applications.

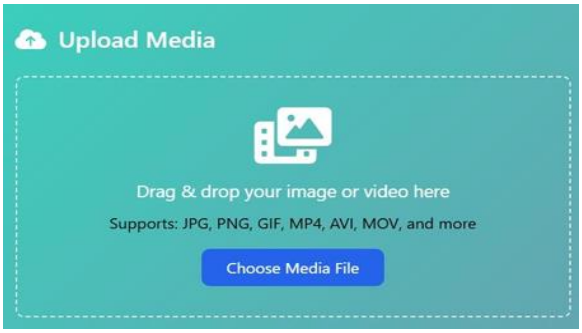


Fig. 2: Upload File

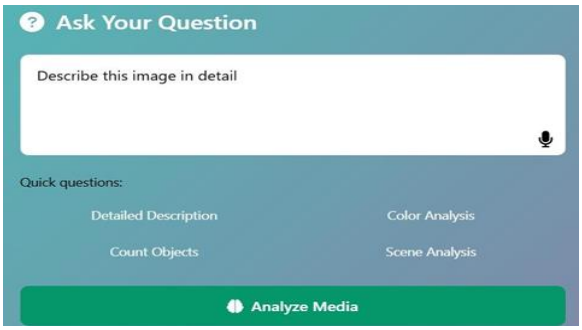


Fig. 3: Ask Your Question



Fig. 4: Select Language

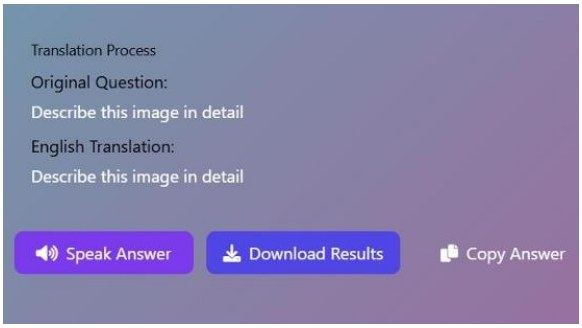


Fig. 5: Result

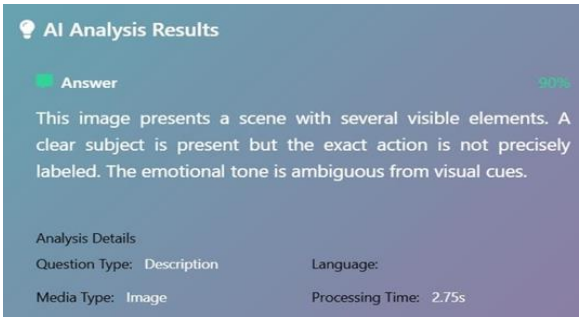


Fig. 6: AI Analysis

V. CONCLUSION

In conclusion, this project demonstrates the potential of automated translation systems in bridging the language gap and enabling seamless communication across different linguistic backgrounds. By leveraging Natural Language Processing (NLP) techniques and translation models, the system effectively translates text between multiple languages with improved accuracy and fluency. The implementation highlights the importance of combining linguistic rules with machine learning approaches to achieve context-aware translations. While the current system provides a strong foundation, future

integration of advanced Neural Machine Translation (NMT) methods can further enhance its efficiency and adaptability. Overall, this work contributes toward developing a reliable, scalable, and intelligent translation framework that can be extended to meet the growing demands of global communication.

REFERENCES

[1] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, ‘‘Deep modular co-attention networks for visual question answering,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6274–6283.

[2] Z. Yu, Y. Cui, J. Yu, D. Tao, and Q. Tian, ‘‘Multimodal unified attention networks for vision-and-language interactions,’’ 2019, arXiv:1908.04107.

[3] C. Chen, D. Han, and J. Wang, ‘‘Multimodal encoder–decoder atten tion networks for visual question answering,’’ IEEE Access, vol. 8, pp. 35662–35671, 2020.

[4] S. He and D. Han, ‘‘An effective dense co-attention networks for visual question answering,’’ Sensors, vol. 20, no. 17, p. 4897, Aug. 2020.

[5] K.Xu,‘‘Show,attendandtell: Neural image captiongeneration with visual attention,’’ in Proc. Int. Conf. Mach. Learn., 2015, pp. 1–24.

[6] C. Kolling, J. Wehrmann, and R. C. Barros, ‘‘Component analysis for visual question answering architectures,’’ in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2020, pp. 1–8.

[7] N. Kim, D. Kim, S. Kwak, C. Lan, and W. Zeng, ‘‘ReSTR: Convolution free referring image segmentation usingtransformers,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 1–23.

[8] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, and Q. Tian, ‘‘Deep multimodal neural architecture search,’’ in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 1–30.

[9] C. Wu, ‘‘Differential networks for visual question answering,’’ in Proc. AAAI Conf. Artif. Intell., 2019, pp. 1–23.

[10] Y. Liu, X. Zhang, F. Huang, Z. Zhou, Z. Zhao, and Z. Li, ‘‘Visual question answering via combining inferential attention and semantic space mapping,’’ Knowl. Based Syst., vol. 207, Nov. 2020, Art. no. 106339.

[11] S. Pandhre and S. Sodhani, ‘‘Survey of recent advances in visual question answering,’’ 2017, arXiv:1709.08203.

AUTHORS Profile



Dr. M. V. NARAYANA received his Bachelor of Engineering (B.E.) degree in Computer Science in 2004 from JNTU Kakinada, India. He completed his Master of Technology (M.Tech) degree in Computer Science in 2010 from JNTU Hyderabad, India, and was awarded his Ph.D. in Computer Science in 2017 from JNTU Kakinada, India. With over 20 years of dedicated experience in academia and research, he currently serves as a Professor and Head of the Department (HoD) in the CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India. His research interests encompass various areas within computer science, and he has contributed significantly to the field throughout his career.



Mr. NITHIN REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Cam-pus affiliated to JNTUH in 2022-2026.



Mr. RAHUL pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Inst-itutions Technical Cam-pus affiliated to JNTUH in 2022-2026.



Mr. HARSHA VARDHAN REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. PRAVEEN pursuing B.Tech degree in Computer Science and Engineering at GuruNanak Institutions

Tech-nical Campus affiliated to JNTUH in 2022-2026.