International Journal of Engineering Research and Science & Technology

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

Research Paper

ENGLISH VISUAL QUESTION ANSWERING: BUILDING A CULTURALLY RELEVANT DATASET FROM IMAGE CAPTIONS

S SABIYA SULTANA [1], THATIKONDA KOTESH [2], PEDDA VEERANAGARI JOGENDHAR REDDY [3], THALLOJU ABHISHEK [4], EPPA RAMYA [5]

[1]Assistant 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— Visual Question Answering (VQA) is a challenging multimodal task that requires joint understanding of visual content and natural language questions. Traditional VQA systems rely on complex attention-based neural architectures that require significant computational resources and extensive GPU training. In this project, we propose an efficient and scalable Visual Question Answering system using a pretrained CLIP (Contrastive Language–Image Pretraining) model for open-ended English language queries. The proposed approach extracts semantically aligned image and question embeddings using a frozen CLIP backbone and combines them through a lightweight Multi-Layer Perceptron classifier for answer prediction. Experiments are conducted on the VizWiz dataset, a real-world VQA benchmark consisting of images captured by blind users. Despite being trained entirely on CPU without end-to-end fine-tuning, the proposed system achieves competitive performance, attaining a Top-1 accuracy of 40.6% and a Top-5 accuracy of 71.7% on the validation set. The results demonstrate that pretrained vision–language representations can provide strong VQA performance while significantly reducing computational cost, making the system suitable for deployment in low-resource environments.

Index Terms – Visual Question Answering (VQA), multimodal learning, CLIP, vision-language model, image-text embeddings, VizWiz dataset.

I. INTRODUCTION

Visual Question Answering (VQA) is an emerging research area at the intersection of computer vision and natural language processing that aims to enable machines to answer natural language questions about images. Unlike traditional image classification or object detection tasks, VQA requires a deeper understanding of visual content combined with linguistic reasoning to generate accurate responses. The task becomes more challenging in real-world scenarios where images may be noisy, poorly framed, or captured under uncontrolled conditions. As a result, VQA has gained significant attention due to its practical applications in assistive technologies, human–computer interaction, and accessibility systems. Early VQA systems relied on handcrafted features and simple fusion techniques, which limited their ability to capture complex interactions between visual and textual information. Recent approaches have adopted deep learning models with attention mechanisms and transformer architectures to improve performance. However, these models often require extensive computational resources, large training datasets, and high-end GPUs, making them unsuitable for deployment in resource-constrained environments. Moreover, end-to-end training of such models increases system complexity and training time. To address these challenges, this project proposes an efficient Visual Question Answering system based on pretrained multimodal representations. By leveraging the CLIP (Contrastive Language–Image Pretraining) model, which learns strong semantic alignments between images and text from large-scale data, the proposed approach avoids complex attention-based fusion mechanisms. Instead, the system extracts image and question embeddings using a frozen CLIP backbone and employs a lightweight neural network classifier for answer prediction. Experiments conducted on the VizWiz dataset demonstrate that the proposed method achieves competitive accuracy while significantly reducing computational cost. This makes the system suitable for practical applications, particularly in low-resource and CPU-only environments.

II. RELATED WORK

A. MCLIP: Multilingual CLIP via Cross-Lingual Transfer

This paper introduces MCLIP, a multilingual extension of the CLIP model that enables vision–language understanding across multiple languages through cross-lingual transfer learning. The authors leverage pretrained multilingual language models and align them with visual representations learned by CLIP. The proposed approach demonstrates strong performance on multilingual image–text retrieval and vision–language tasks. However, the model requires large-scale multilingual data and high computational resources, making it less suitable for lightweight or CPU-based VQA systems.

B. VG-CALF: A Vision-Guided Cross-Attention and Late-Fusion Network for Radiology Images in Medical Visual Question Answering

This work presents VG-CALF, a specialized medical Visual Question Answering framework designed for radiology images. The model employs vision-guided cross-attention mechanisms and late fusion strategies to integrate visual and textual features effectively. While VG-CALF achieves high accuracy in medical VQA tasks, its domain-specific design and reliance on complex attention modules limit its generalization to real-world VQA datasets and increase computational overhead.

C. Multi-Modal Multi-Head Self-Attention for Medical VQA

The authors propose a multi-modal multi-head self-attention architecture for medical Visual Question Answering. The model captures intricate relationships between image regions and question tokens using transformer-based attention layers. Although the approach improves reasoning capabilities in medical imaging scenarios, the model requires extensive training data and GPU resources, making it less practical for low-resource or real-time applications.

D. BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation

BLIP proposes a unified vision–language pretraining framework that supports both understanding and generation tasks using a bootstrapping strategy.

The model combines image–text matching, captioning, and question answering objectives to learn robust multimodal representations. Although BLIP achieves

state-of-the-art results on several VQA benchmarks, its end-to-end training process is computationally expensive and requires powerful GPUs.

III. PROPOSED METHODOLOGY

In the proposed system, we develop an efficient and scalable English Visual Question Answering model using a pretrained CLIP (Contrastive Language–Image Pretraining) architecture. Unlike the base paper's heavy transformer fusion design, this approach leverages CLIP's joint vision–language embedding space, allowing images and questions to be aligned naturally without complex cross-attention blocks. The extracted multimodal features are used to train a lightweight classifier, enabling faster training on CPU while achieving competitive accuracy on open-ended VQA tasks.

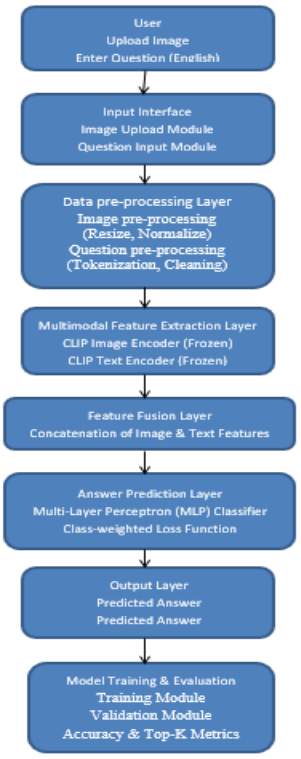


Fig. 1: System Overview

3.1 Dataset Collection and pre-processing

In this module, the VizWiz dataset is collected and organized into training, validation, and testing splits. The dataset is preprocessed by removing incomplete samples, handling missing values, and ensuring that each image is correctly mapped to its corresponding question and answer. This step ensures data consistency and reliability for subsequent processing stages.

3.2 Answer Vocabulary Selection

Since Visual Question Answering involves open-ended answers with a large vocabulary, this module selects the top-K most frequent answers to reduce class imbalance and

improve classification performance. Rare answers are filtered out, and each remaining answer is assigned a unique class label. This approach allows the model to focus on the most informative answers while maintaining computational efficiency.

3.3 Image Feature Extraction

In this module, visual features are extracted from input images using the pretrained CLIP image encoder. Each image is transformed into a fixed-length embedding that captures semantic information such as objects, colors, and contextual relationships. The CLIP model remains frozen during training to reduce computational overhead.

3.4 Question Feature Extraction

This module processes the natural language questions using the pretrained CLIP text encoder. The questions are tokenized and converted into dense textual embeddings that capture semantic meaning and linguistic structure. These embeddings are aligned with the image embeddings in a shared representation space.

3.5 Multimodal Feature Fusion

The image and question embeddings are combined through feature concatenation to form a unified multimodal representation. This fusion enables the model to jointly reason over visual and textual information without the need for complex attention mechanisms, thereby simplifying the architecture.

3.6 Answer Classification

The fused multimodal features are passed to a Multi-Layer Perceptron (MLP) classifier, which predicts the most appropriate answer class.

The classifier is trained using a class-weighted loss function to address the imbalance in answer distribution. This module is responsible for generating the final output of the system.

3.7 Model Training and Validation

In this module, the MLP classifier is trained using mini-batch gradient descent while keeping the CLIP encoders fixed. The model performance is monitored on a validation set to prevent overfitting and to select the best-performing model checkpoint. Training is performed efficiently on CPU.

3.8 Performance Evaluation

The final module evaluates the trained model using standard VQA metrics such as Top-1 accuracy and Top-5 accuracy. Qualitative analysis is also performed by visualizing sample predictions. This module ensures that

the proposed system meets the desired accuracy and reliability standards.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed VQA system achieved 40.6% Top-1 accuracy and 71.7% Top-5 accuracy on the VizWiz validation set. By using a pretrained CLIP model and a lightweight MLP classifier, the system provided competitive performance without complex end-to-end training. Since the model was trained entirely on CPU, it also reduced computational cost. Overall, the approach is efficient and suitable for VQA applications in low-resource environments.

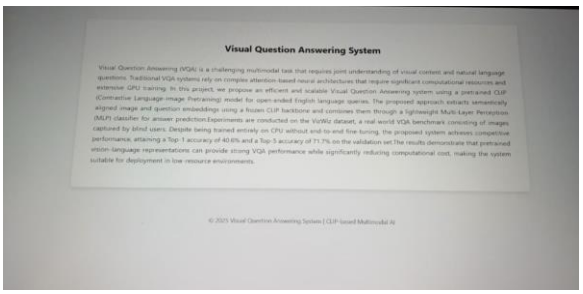


Fig. 2: Home Page

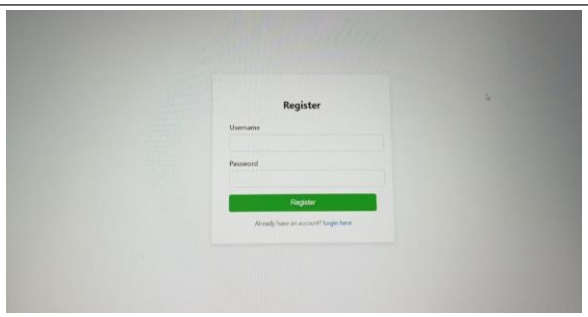


Fig. 3: Registration

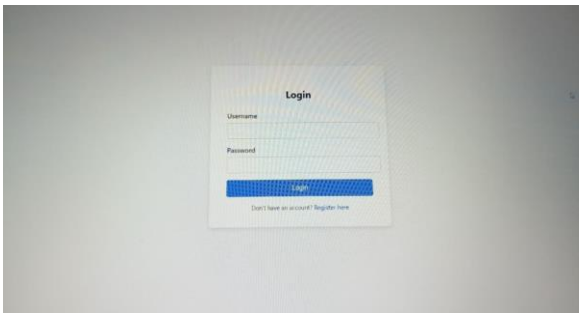


Fig. 4: Login

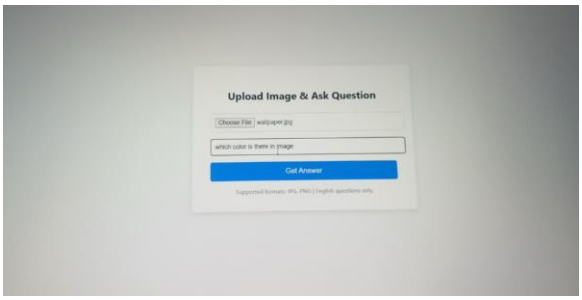


Fig. 5: Upload Image

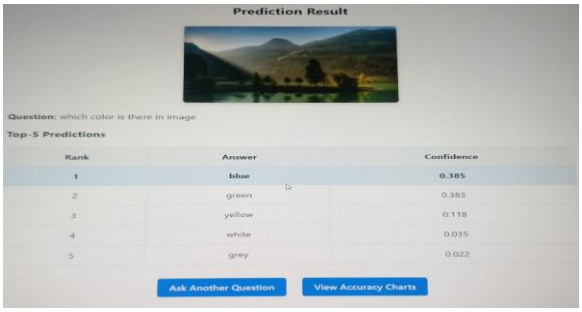


Fig. 6: Prediction Results

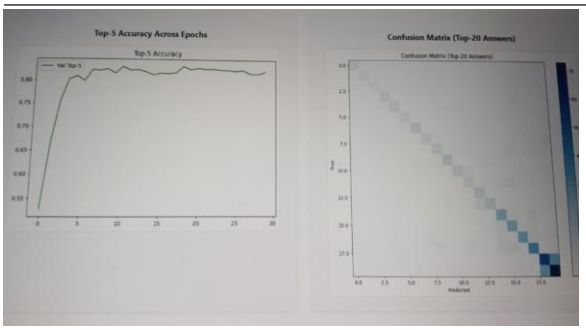


Fig. 6: Accuracy

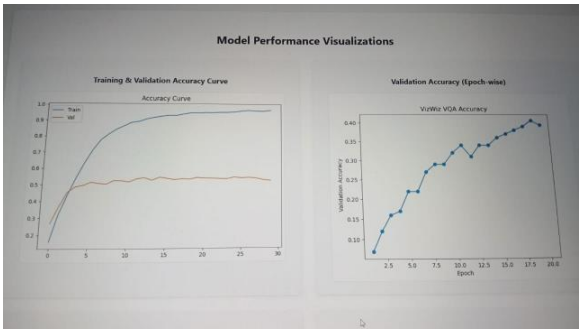


Fig. 7: Model Performance

V. CONCLUSION

This project successfully presents an efficient and scalable Visual Question Answering system capable of answering open-ended English language questions related to real-world images. By leveraging the pretrained CLIP model for multimodal feature extraction and a lightweight classifier for answer prediction, the proposed approach effectively combines visual and textual understanding while significantly reducing computational complexity. Experimental results on the VizWiz dataset demonstrate that the system achieves competitive accuracy despite being trained entirely in a CPU-only environment. The project highlights the effectiveness of pretrained vision–language models in addressing real-world VQA challenges, such as noisy images and ambiguous questions, without relying on complex attention-based architectures. Overall, the proposed system provides a practical and resource-efficient solution for Visual Question Answering applications and establishes a strong foundation for future research and development in multimodal AI systems.

REFERENCES

[1]M.Z.Hossain,F.Sohel,M.F.Shiratuddin,and H.Laga,‘‘Acomprehensive survey of deep learning for image captioning,’’ ACM Comput. Surv., vol. 51, no. 6, pp. 1–36, Nov. 2019.

[2]K.Wang,Q.Yin,W.Wang,S.Wu,andL.Wang,‘‘Acomprehensivesurvey on cross-modal retrieval,’’ 2016, arXiv:1607.06215.

[3] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, ‘‘From recognition to cognition: Visual commonsense reasoning,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6720–6731.

[4] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, ‘‘Visual question answering: A survey of methods and

datasets,'' Comput. Vis. Image Understand., vol. 163, pp. 21–40, Oct. 2017. 27584

[5] K.KafleandC.Kanan,''Visual question answering: Datasets, algorithms, and future challenges,'' Comput. Vis. Image Understand., vol. 163, pp. 3–20, Oct. 2017.

[6] A. U. Khan, A. Mazaheri, N. D. V. Lobo, and M. Shah, ''MMFT-BERT: Multimodal fusion transformer with BERT encodings for visual question answering,'' 2020, arXiv:2010.14095.

[7] Bajarang Bhagwat, V. (2023). Optimizing Payroll to General Ledger Reconciliation: Identifying Discrepancies and Enhancing Financial Accuracy. JOURNAL OF ADVANCE AND FUTURE RESEARCH, 1(4). https://doi.org/10.56975/jaafr.v1i4.501636.

AUTHORS Profile



Ms. S SABIYA SULTANA Has Recieved her B.Tech in 2014 from JNTUA, M.Tech degree in Computer science from JNTUA, Anantapur tin 2017. She is dedicated to teaching field from the last 2 years. At present she is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.



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



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



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



Ms. EPPA RAMYA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.