International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

English Visual Question Answering: Building a Culturally Relevant Dataset from Image Captions

B. Dinesh kumar[1], Bhoomika Mandadi[2], C. Anil kumar[3], Dr. Md Sirajul Huque[4]

[1,2,3]UG Scholars, Department of Computer Science and Engineering Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India.

[4]Assistant Professor, Department of Computer Science and Engineering Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India.

Abstract - Visual Question Answering (VQA) is a challenging multimodal task requiring joint understanding of visual content and natural language questions. Traditional VQA systems rely on complex attention-based architectures demanding significant computational resources and GPU training. This paper proposes an efficient and scalable VQA system using a pretrained CLIP (Contrastive Language–Image Pretraining) ViT-B/32 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 (MLP) classifier for answer prediction. Experiments on the VizWiz dataset — a real-world benchmark of images captured by visually impaired users — demonstrate competitive performance, achieving a Top-1 accuracy of 40.6% and Top-5 accuracy of 71.7%, trained entirely on CPU without end-to-end fine-tuning. A Flask-based web application supporting user authentication, image upload, and real-time Top-5 predictions with confidence scores is also demonstrated.

Key Words: Visual Question Answering, CLIP, VizWiz Dataset, MLP Classifier, Multimodal Learning, Deep Learning

1.INTRODUCTION

Visual Question Answering (VQA) is an emerging research area at the intersection of computer vision and natural language processing that enables machines to answer questions about images. Unlike traditional image classification or object detection, VQA requires deeper visual understanding combined with linguistic reasoning, becoming especially challenging when images are noisy or captured under uncontrolled conditions as in accessibility-oriented datasets.

Early VQA systems relied on handcrafted features and simple fusion techniques, limiting their ability to capture complex visual-language interactions. Recent deep learning models with attention mechanisms and transformer architectures improve performance but require extensive computational resources and high-end GPUs, making them unsuitable for resource-constrained deployment.To address these challenges, this paper proposes an efficient VQA system using a frozen CLIP (Contrastive Language–Image Pretraining) ViT-B/32 backbone for image and question embedding extraction, combined with a lightweight MLP classifier for answer prediction. Deployed as a Flask web application, the system achieves a Top-1 accuracy of 40.6% and Top-5 accuracy of 71.7% on the VizWiz dataset, operating entirely on CPU.

2. Body of Paper

2.1 Literature Survey

Visual Question Answering has gained significant attention due to rapid advances in computer vision and natural language processing. Early VQA systems used Convolutional Neural Networks (CNNs) for image feature extraction and Long Short-Term Memory (LSTM) networks for question processing. Although these models achieved reasonable performance, they struggled to capture complex relationships between visual objects and language queries. Recent transformer-based architectures have significantly improved multimodal understanding across vision-language tasks including image captioning, visual reasoning, and question answering.

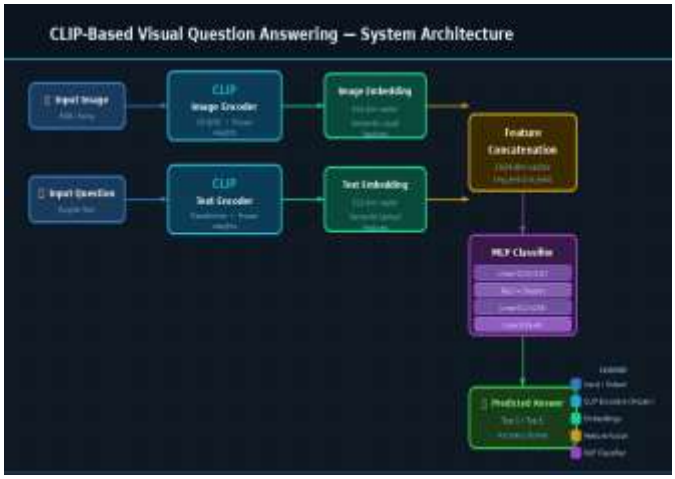
Chen et al. proposed MCLIP, a multilingual CLIP extension for cross-lingual vision-language understanding, requiring large-scale GPU resources unsuitable for lightweight deployment. Lameesa et al. introduced VG-CALF, a vision-guided cross-attention framework for medical VQA, limited to domain-specific tasks with high computational overhead. Joshi et al. proposed a multi-head self-attention architecture for medical VQA that improves reasoning but demands extensive GPU training. Li et al. presented BLIP, a unified vision-language pretraining framework supporting both

understanding and generation, though computationally expensive for CPU deployment.

2.2 Proposed Methodology

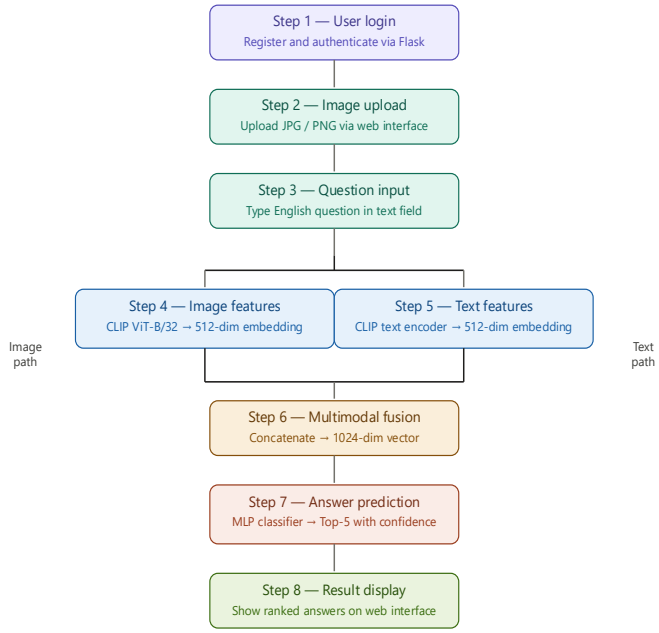
The proposed system develops an efficient VQA model using a pretrained CLIP ViT-B/32 architecture. Unlike complex transformer fusion designs, this approach leverages CLIP's joint vision-language embedding space to align images and questions naturally without cross-attention mechanisms. The VizWiz dataset, consisting of real-world images captured by visually impaired users with crowd-sourced answers, is used for training and evaluation. The top-500 most frequent answers are selected as the classification vocabulary to handle the highly imbalanced answer distribution. The frozen CLIP image encoder and text encoder independently produce 512-dimensional embeddings for each image and question respectively. These embeddings are concatenated into a unified 1024-dimensional multimodal vector and passed through a lightweight MLP classifier consisting of Linear(1024→1024) with BatchNorm and Dropout(0.4), Linear(1024→512) with Dropout(0.3), and Linear(512→500).

3. System Architecture



The system architecture consists of two parallel input pipelines. An input image is processed by the frozen CLIP ViT-B/32 image encoder producing a 512-dimensional visual embedding, while the input question is processed by the frozen CLIP text encoder producing a 512-dimensional text embedding. Both embeddings are concatenated to form a 1024-dimensional multimodal feature vector, which is passed through the MLP classifier to generate Top-5 predicted answers with confidence scores. The MLP is the only trainable component, keeping the CLIP encoders frozen throughout. The system is deployed via Flask, allowing users to upload images, submit questions, and receive real-time predictions through a web interface.

3.1Workflow of the Proposed System



The workflow of the proposed Visual Question Answering system consists of multiple stages that process the input image and user query to generate accurate predictions.

Step 1 – User Registration and Login The user registers and logs in through the secure Flask-based authentication interface. Only authenticated users are granted access to the VQA system.

Step 2 – Image Upload The authenticated user uploads an image in JPG or PNG format through the web interface. The uploaded image is stored temporarily and prepared for further processing.

Step 3 – Question Input The user types a natural language question in English related to the uploaded image through the web interface text input field.

Step 4 – Image Feature Extraction The uploaded image is preprocessed and passed through the frozen CLIP ViT-B/32 image encoder, which produces a 512-dimensional semantic embedding capturing visual features such as objects, colors, and spatial relationships.

Step 5 – Question Feature Extraction The input question is tokenized and passed through the frozen CLIP text encoder, which produces a 512-dimensional text

embedding capturing the semantic meaning and linguistic structure of the question.

Step 6 – Multimodal Feature Fusion The 512-dimensional image embedding and 512-dimensional text embedding are concatenated to form a unified 1024-dimensional multimodal feature vector representing both visual and textual information jointly.

Step 7 – Answer Prediction The 1024-dimensional vector is passed through the trained MLP classifier, which generates probability scores for all 500 answer classes and returns the Top-5 predicted answers with confidence scores.

Step 8 – Result Display The Top-5 predicted answers along with their confidence scores are displayed to the user through a structured result table on the web interface.

3.2 Methodology

The proposed system is implemented using Python as the primary programming language. The Flask framework is used to develop the web-based interface supporting user authentication, image upload, and real-time result display. The CLIP ViT-B/32 model is loaded using the HuggingFace Transformers library with the model identifier openai/clip-vit-base-patch32. The CLIPProcessor handles all image resizing, normalization, and question tokenization required before feature extraction. The MLP classifier is built and trained using the PyTorch deep learning library. The VizWiz dataset annotations are processed using the Pandas library, and the LabelEncoder from Scikit-learn is used to encode the top-500 answer classes. The trained MLP model is saved as vizwiz_clip_mlp_best.pth and loaded at inference time for real-time predictions. Image processing operations are handled using the Python Imaging Library (PIL). The entire training and inference pipeline runs on CPU without requiring any GPU hardware, making the system accessible for deployment in resource-constrained environments.

3.3 Implementation

The system is implemented in Python using Flask for the web interface supporting user authentication, image upload, and result display. The CLIP ViT-B/32 model is loaded via HuggingFace Transformers using openai/clip-vit-base-patch32, with CLIPProcessor handling image preprocessing and question tokenization. The MLP classifier is built using PyTorch, and dataset annotations are processed using Pandas. The LabelEncoder from Scikit-learn encodes the top-500 answer classes. The trained model is saved as vizwiz_clip_mlp_best.pth and loaded at inference time for real-time predictions.The entire pipeline runs on CPU without any GPU hardware requirement.

4. Results and Discussion

The developed application successfully processes uploaded images and generates Top-5 predicted answers with confidence scores in real time. It demonstrates a sample prediction result where the system correctly identifies the answer "cat" with a confidence score of 0.982 for the input question "what is this", demonstrating strong performance on clear object recognition queries. The high confidence score indicates that the CLIP embeddings effectively capture semantic visual features, enabling accurate answer prediction through the lightweight MLP classifier.

Table 1: Performance Comparison of VQA Models

Model	Approach	Training Device	Top-1 Accuracy	Top-5 Accuracy
ResNet-152 + LSTM	CNN + Attention + LSTM	GPU	~31%	—

CLIP + Linear Layer	Frozen CLIP + Linear Layer	GPU	54%	—
CLIP + Linear Layer	Frozen CLIP + Linear Classifier	GPU	60.15%	—
Proposed CLIP + MLP	Frozen CLIP + MLP + Focal Loss	CPU Only	40.6%	71.7%

Description:
This table compares the proposed system against existing VQA models on the VizWiz dataset. While GPU-trained CLIP-based systems achieve higher Top-1 accuracy, the proposed system achieves 40.6% Top-1 and 71.7% Top-5 accuracy operating entirely on CPU without end-to-end fine-tuning. This demonstrates that the proposed approach delivers competitive multimodal understanding while significantly reducing computational requirements, making it suitable for low-resource deployment.

5. CONCLUSIONS

Conclusion

This paper presented an efficient Visual Question Answering system using a frozen CLIP ViT-B/32 backbone combined with a lightweight MLP classifier trained on the VizWiz dataset. The system achieves a Top-1 accuracy of 40.6% and Top-5 accuracy of 71.7% operating entirely on CPU without end-to-end fine-tuning. Focal Loss with gamma=2.0 and Adam optimizer effectively handle the imbalanced answer distribution across 500 answer classes. The Flask-based web application demonstrates practical deployability with user authentication, image upload, and real-time Top-5 predictions with confidence scores. The results confirm that pretrained vision–language representations deliver competitive VQA performance in low-resource environments.

FUTURE SCOPE

Although the proposed system performs effectively on CPU, several enhancements can be explored in future research. End-to-end fine-tuning of the CLIP backbone using GPU resources could further improve Top-1 accuracy. The system can be extended to support generative answer models using large language models, enabling free-form responses beyond predefined answer classes. Multilingual question support would enhance accessibility for non-English speaking users. Integration of explainable AI techniques such as attention visualization would improve model interpretability.

ACKNOWLEDGEMENT

The authors express sincere gratitude to Dr. Md Sirajul Huque, Assistant Professor, Department of Computer Science and Engineering, Guru Nanak Institutions Technical Campus, Hyderabad, for providing valuable guidance and continuous encouragement throughout the development of this project. The authors also acknowledge the institution for providing the necessary infrastructure and computational resources.The authors are also grateful to the open-source communities behind PyTorch, HuggingFace Transformers, and Flask whose tools supported the entire development process.

REFERENCES

[1]. M.Z.Hossain, F.Sohel, M.F.Shiratuddin, and H.Laga, "A comprehensive 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, and L.Wang, "A comprehensive survey 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.

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