The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

The Comprehensive Review of VQA in Education: Challenges and Innovations

Sushmita Dokkari[1], Rakesh Kamireddy[2], Venkata Subhash Paritala[3*]
Abhishek Yalamanchili[4], A.S. Venkata Praneel[5]
[1,2,3,4,5] Department of Computer Science and Engineering,
GST, GITAM (Deemed to be University), Visakhapatnam AP India.
sushmitadokkari521@gmail.com[1], rakesh18kamireddy@gmail.com[2], subhashparitala22@gmail.com[3],
yalamanchiliabhishek09@gmail.com[4], pankired@gitam.edu[5]
Corresponding Author *: pankired@gitam.edu

Abstract

The integration of AI in education has revolutionized the interactive way of learning. For instance, in Visual Question Answering (VQA) systems. This work is a novel approach to implementing real-time, interactive VQA in bridging traditional text-based learning and image-driven multimodal education. VQA is an essential area of artificial intelligence that combines computer vision and natural language processing to interpret and respond to image-based queries. Here, users may upload or capture images, then automatically receive their descriptions, structure their questions, and get assured answers with an embedded verification mechanism for accuracy. Our study focuses on analyzing a VQA system specifically for educational applications by leveraging insights gained from an extensive review of historic research papers. Through this survey, we analyzed various datasets and models used in previous studies to identify the most effective approaches for educational use cases. From our study and understanding of various datasets, we used COCO-QA and VQA 2.0 as our primary datasets due to their extensive annotations, diverse question-answer pairs, and widespread acceptance in the VQA research community. Additionally, we integrated models such as BERT, ResNet50, Bottom-Up and Top-Down (BUTD) attention, Bilinear Attention Networks (BAN), and Multi-Scale Features to enhance performance. By systematically comparing multiple VQA studies, we identified common shortcomings, including inadequate dataset generalization and inefficient feature extraction.

Keywords: VQA, Image Captioning, Natural Language Processing, Computer Vision, Deep Learning for VQA, Multimodal Learning, Transformer Models in VQA, Attention Mechanisms.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

1. INTRODUCTION

VQA denotes the novel algorithmically integration of Artificial Intelligence into pedagogy, in which queries posed by users to images are answered through explanations and answers. The purpose of this current work of research is to attempt to design an education-based VQA-interactive system that can be applied to science, history, and geography, as well as to art-related questions in general and, among other examples, to improve the nature of the learning experience surrounding the subject that is under study. It allows students to manipulate images and then supply them with answers for contexts [1]. Thus, it bridges the gap between learning, which is traditionally learned by watching text and being equipped with cognitive benefits from visual interaction.

There is always a look around in the educational landscape for the best methods [2] of delivering knowledge and meaningful engagement with learners. The use of an interactive VQA system may be a step toward dynamic and intuitive AI development in the field of learning. A biology student can post an image of a cell drawing and get what components are explained to him. Such tools will stimulate curiosity and deeper levels of thinking, catering to all kinds of learning styles.

Furthermore, due to the real-time capability of our VQA [3] system, it accepts user-uploaded images as soon as they are available, or users can take photos with the camera before delivering the queries. Using the input image and query, the system retrieves an appropriate answer and also validates the truthfulness of the answer. This guarantees that students are provided with consistent and accurate data, thereby promoting deeper learning. The feature of being able to assess the correctness of the answer provides one more educational benefit, contributing to improving the system's strength and effectiveness for educational purposes [4]. Despite their promise, developing a VQA system specifically for education purposes is rife with challenges. The usual VQA models are in fact based on general datasets, not specifically designed for educational applications, where, due to the importance of domain-specific knowledge and accuracy, the quality of the available data can vary significantly. Further challenges encompass processing ambiguous and rich questions, supporting linguistic plurality amongst users, and facilitating the access of learners with different educational backgrounds.

This paper presents the design and implementation of an educational VQA system, describing its key features, underlying technologies and future issues associated with it. By addressing these aspects, this research aims to contribute to the growing integration of AI in education, offering insights into how interactive VQA tools can improve learning outcomes and promote engagement.

2. USES OF INTERACTIVE VQA

Contemporary methods of teaching are changing drastically and have altered how learning occurs in society by making these new methods fun, engaging, and effective ways for students. Of the vast contributions made recently to this subject matter, an advancement that must be mentioned would include interactive visualizing, as achieved through a set of imagery-using systems of this nature that greatly enhance both memory and facilitate learning of larger sets of content matter.

Images are also essential in learning as they provide significant enhancement to understanding and remembering. Studies show that people learn and remember much better when images are used together with the information. Interactive visualization systems allow the student to engage actively with the images, relating answers to a specific context. This capability fosters a deeper understanding of abstract concepts, making subjects like history, science, and geography more accessible and exciting. For instance, students can compare historical artifacts, analyze scientific diagrams, or explore geographical maps in a dynamic way, which enriches their learning experience.

The flexibility of interactive visualization systems lends them to application in a vast array of education disciplines, from science to history, geography, and even art. This applies the tool to varied curricula, allowing educators to implement the tool appropriately for different subjects. Through this, the systems ensure students from different educational backgrounds utilize it as a means of elevated learning levels.

Active learning is one of the essential features of successful education, and interactive visualization systems encourage it. The system will stimulate the learner to actively interact with the material and, therefore, turn passive receivers of information into active participants in their education. Interactive and personalized, it lets students get more involved in topics, leading to critical thinking and independent exploration.

One of the most attractive features of interactive visualization systems is that they can support individualized learning paces and styles. They provide personalized feedback and explanations, which support a wide range of learner requirements, from beginners to advanced students. This ensures that every learner gets the support he or she needs to succeed, making education more inclusive and effective. Interactive visualization systems offer contextual explanations and step-by-step procedures in addition to answering questions. This would ensure that not only are correct answers provided to learners but that the reasoning is also understood by them. The more

comprehensive knowledge and critical thinking skills acquired make students understand more profoundly the content.

Continual improvement through the inclusion of feedback mechanisms is essential to interactive visualization systems. Teachers and learners both are facilitated, as time progresses in this periodic process-where such a system should refine and get closer to accuracy. With feedback incorporated, user needs will be met through these systems, which keep evolving to stay relevant and actually become more effective educationally.

Interactive visualization systems make up a package of excellence in resource-scarce environments where more quality educators and teaching materials are scarce, and it democratizes access to more quality educational resources that enlighten learners to eventually bridge knowledge gaps-a capability much needed in bringing equitable education to every student regardless of environment.

In that regard, a modern piece of interactive visualization systems takes that transformative role. Besides enhancing learning through imagery, supporting multiple disciplines, allowing for active engagement, personalizing education, encouraging offering contextual explanations, and enabling continuous improvement, these systems greatly enhance the educational experience. As we continue embracing technological advancements, it is huge potential with interactive visualization for democratizing and elevating education, thus opening doors for a more informed and actively engaging group of learners.

3. SHORTCOMINGS

3.1 Limitation of Dataset:

Domain Agnostic Datasets: The datasets like VQA 2.0, Visual Genome, and COCO have not been prepared exclusively for the educational domain. Hence, these make VQA not generalizable, but on the other hand specific for educational purposes.

Limited Diversity: The dataset may not hold diverse subjects, languages and educational levels that would make the dataset representative for more people.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

3.2 Question Difficulty Level

Vague Query: Questionnaire that is vague or open-ended and requires deeper inference or previous knowledge is still challenging.

Contextual Understanding: Educational questions, which require multi-hop reasoning in terms of retrieving information from multiple sources, cannot be well achieved with the current VQA systems.

3.3 Accuracy and Relevance of Answer:

Model Limitation: Sometimes, pre-trained models do not provide accurate and relevant answers, especially for very specific or subtle educational-type questions.

Error Propagation: Error in text recognition or image feature extraction may cause the wrong answers to be displayed, which might have a bad effect on the learning process.

3.4 User Interface Challenges:

Accessibility: Creating an intuitive interface that can be accessed by everyone, regardless of age and technical proficiency, is very challenging.

Cognitive Load: The overloading of information, for example, too many annotations or overly detailed explanations, might impede learning.

3.5 Feedback and Adaptation:

Slow Learning Adaptations: The time to add feedback that improves the model's accuracy is too long. It may not be able to adapt fast enough to the expectations of the users.

3.6 Resource Intensity:

High Computational Costs: Training and fine-tuning large VQA models, especially those that make use of multimodal transformers, is computationally expensive.

Deployment Challenges: Real-time execution on edge devices, like tablets or mobile phones, is not easy because of hardware constraints.

3.7 Educational Relevance:

Trade-off between Complexity and Detail: The system should be suitable for all levels of education, ranging from primary school students to college students, making it challenging to design responses and explanations.

Narrow Use Cases: This means that the usefulness of the system could be limited to narrow topics or school content that decrease its impact generally.

3.8 Ethics and Bias Issues:

Bias in the Datasets: Preexisting bias in datasets results in giving inappropriate answers with some form of cultural bias or more.

Privacy Issues: The users could upload pictures on the sites; thus, pictures may end up being a data collection process which is improper for educational setting purposes.

3.9 Adoption and Implementation:

Teacher and student acceptability: The educators and students do not readily accept new technology due to un-familiarity with new technology and apparent complexity.

Curriculum integration: Alignment of curricula with the VQA system's capabilities is a very complex procedure.

4. EXISTING METHODS IN VQA FOR EDUCATIONAL PURPOSES

4.1. Multimodal Learning Techniques

Multimodal learning is learning data from other sources-well, not only text from which a VQA system should extract for its functionality. The earlier techniques were using CNNs to feature an image and then using RNNs, particularly LSTM networks to process textual input questions. That

ensures the system is aware both of what the image represents as well as the linguistic structure of a question.

In the latest architecture, transformers replace RNNs because of the efficient processing of sequential data. BERT and ViL apply the mechanism of attention so that it describes the interaction between words and, further, regions of an image. Hence, these methods contribute improvements toward understanding the system with an aspect of parts of the image in which the linkages with suitable words of the question are developed. In turn, multimodal transformers such as ViLBERT reduce the process complexity because it runs text features parallel with image ones in order that its more accurate response could be developed. Multimodal learning [5,6] enables the use of fusion methods to better understand questions. Techniques, such as concatenation or bilinear pooling fuse image features with text-based features into one single representation. State-of-the-art systems involve advanced attention mechanisms that dynamically focus on different image regions about the query during answering.

4.2. Dataset Utilization

Dataset utilization is one of the major aspects in training and fine-tuning VQA models. The most extensively used general-purpose datasets [7,8,9,10,11] for comprehensive annotations and diverse question-answer pairs are VQA 2.0, COCO, and Visual Genome. For instance, VQA 2.0 is comprised of over 80,000 images and 1.1 million questions with other benchmarks being more specialized. For example, there is COCO with object detection and captioning annotations or even more intimate relationships between elements in images such as the case of Visual Genome that assists in complex reasoning tasks.

The datasets are often also altered according to the research findings to add educational content related to domains. The sets of data from the system EDUVQA [12], or EDUVI [13], were gathered from books from NCERT or any such source that incorporates education. In fact, all images related to the topic or questions based on different levels of educations are parts of the same dataset. The synthetic dataset of CLEVR is mainly used because the structured compositional reasoning tasks indicate a very complex question-answering situation in a controlled environment. The quality of the dataset decides the accuracy and applicability of the model. A specialized dataset of domains is highly critical for educational VQA since the curricula answers are so significant to have precision and relevance.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

4.3. Model Architectures

The architecture of a VQA system determines what the capability to process and understand images and questions is. Its earlier models followed the CNN-LSTM architecture where CNNs were used for extracting visual features from images and LSTMs explained the sequential nature of questions. It was effective but not scalable for complex reasoning tasks.

Transformers have revolutionized this space. VisualBERT and ViLBERT make use of the architectures of transformers while handling multimodal data. This can be applied over both the composite inputs of images and texts to boost their contextual understanding through the self-attention mechanism. The best results about compositional models trained on the CLEVR dataset have been achieved for outstanding performance in multi-step reasoning. These break up the complex queries into simpler logical operations.

The attention mechanism further enriches these architectures since it dynamically pays attention to important regions in the image and to words, which allows it to focus on relevant details such as whether an object exists in some particular image or if a word appears in some question that allows it to be more responsive to it.

4.4. Educational VQA Approaches

Educational VQA systems are developed to be more accessible and applicable in the teaching and learning environment since it brings closer and engages complex AI processes. EDUVQA is a model dealing with relatively simple question types such as Yes/No, or even multiple-choice type questions that could be well applied to young students or beginning-level learners. Simplicity becomes an issue because it tends to make the answers understandable and fitting to any education context.

Others are EDUVI, merging the image captioning functionality with VQA. This means the functionality can also give a user a descriptive caption along with answers, thereby enhancing the learning process. For instance, a student asks about the cycle of life in a plant. In return, the system produces both a descriptive explanation and answer to specific questions about the picture. These models are generally trained on datasets sourced from educational resources, so they are curriculum aligned and learning goal aligned. They are all user-focused, approach detail and simplicity giving rise to powerful learning.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

4.5. Feedback and Adaptive Learning

The feedback pipeline is the most crucial component of fine-tuning VQA systems. It becomes much easier to identify what to improve as it collects ratings and comments from users on the responses of the system. This is an iterative process that makes developers fine-tune models much better.

This dynamic adaptation, based on real-time reinforcement feedback, may be integrated with reinforcement learning within such systems. The adaptive behavior enhances the performance of the system, more strongly aligning it with the expectations and needs of its user over time. In effect, the pace and preferences of the learner are utilized for delivering a more personalized and efficient learning experience to the user.

Feedback loops are also helpful in rectifying errors or fine-tuning the model of ambiguous queries. If a user answers and marks a response as incorrect, for example, the system would retrain on near-queries to get better response in future for that query. Adaptive learning keeps an innovation and effectiveness as changes take pace with the evolving what is learned over time.

4.6. Interactive Interfaces

The developed VQA system includes a web-based interface to make it user-friendly and accessible to the user, especially students. The use of HTML, CSS, and JavaScript technologies in the interface provides an intuitive and seamless interaction with the system. Features such as drag-and-drop functionality for image uploads have been implemented to make the process of providing visual input easier. This application also has a text box wherein the questioning process will be facilitated so that the user can easily input his queries. These thoughtful design elements come together to create an environment friendly to users, thereby creating an engagement and ease in use. Advanced interfaces give explanations with visual justifications of the regions of the image supporting the answers given by the system. For instance, if a student wants to know where a specific organ is located on a diagram, the system can actually draw markers around the corresponding part of the image. Such features make the system more interactive and appealing so as to ensure its usability in multiple educational settings.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

5. DATASETS

Table 1: Comparison of datasets on some features.

Feature	VQA v2 Dataset	CLEVR Dataset	COCO QA Dataset	Visual Genome
Purpose	General-purpose VQA with reduced answer bias	Compositional reasoning and structured QA	Enhances AI's ability to understand images and answer questions	Deep semantic understanding of images
Image Type	Real-world images (Microsoft COCO)	Synthetic images (geometric objects)	Real-world images (COCO dataset)	Real-world images with detailed annotations
Question Type	Yes/No, multiple-choice, open-ended	Logical reasoning, spatial relationships	Object recognition, actions, relationship	Complex scene understanding
Dataset Size	200k+ images, 1M+ Q&A pairs	100k+ images, 1M+ Q&A pair	Extensive collection of images with various question types	Millions of images with detailed annotations
Strengths	Reduces bias, diverse human-annotated answers, suitable for benchmarking	Strong in multi-step logical reasoning, structured question-answering	Bridges the gap between vision and NLP, supports contextual scene understanding	Advances AI perception, deep semantic knowledge
Limitations	Limited domain-specific applicability	Synthetic images may not reflect real-world complexity	Focuses only on COCO images, limiting diversity	Complexity may require advanced models to utilize effectively
Educational Use	General-purpose VQA training, useful for students engaging with visual content	Useful for teaching logical reasoning and spatial understanding	Supports AI research in contextual image understanding	Enhances AI models for deeper semantic comprehension of visual data

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

6. EXISTING MODELS

6.1. BUTD MODEL:

The Bottom-Up and Top-Down Attention (BUTD) model [14] is among the most acknowledged and influential deep learning frameworks that are specifically built for the task of VQA. The model was first proposed by Anderson et al. in 2018, which marked a complete revolution in the way visual data was being processed and interpreted by artificial intelligence systems in association with natural language queries. The core innovation of the BUTD model lies in the attention mechanism that is conceptually different from traditional VQA models, which rely on a global image. This particular model utilizes the object detection method, which, in turn, allows a model to draw the attention of only some objects existing in an image. This enables the model to not only effectively extract relevant features from the input image but do so in a structured approach and in an interpretable way. This is by identifying and paying attention to different objects within it, hence making it answer more accurately and contextually relevant to questions posed about the image.

So, in summary, this BUTD model might be considered a giant leap forward for the field of VQA since it conceptualizes extracting visual features with a much more structured and interpretive method. Its success in many scenarios does not serve to negate its inability to address matters of text processing or complex reasoning and leaves room for investigation and fine-tuning of methodologies in VQA systems.

6.2.BAN MODEL:

It is the Bilinear Attention Networks model, an impressive step in the task of VQA which involves the fusion of both visual and textual information to answer questions about images. It was developed by Kim et al. in 2018 [15] and has emerged as a highly prominent model for multi-modal learning tasks because the ability to understand and interpret the intricate relationships between visual content and textual queries is crucial in generating accurate responses. One of the key innovations of the BAN model is its use of a bilinear attention mechanism. Unlike the common attention models which used to focus on one feature of the input data, this bilinear approach focuses on a more subtle interaction between multiple regions of an image and also between different words in the question, capturing some complex dependencies and interactions critical in

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

understanding the context of a question relative to its visual content. For instance, at the time when the question mentions a certain object in an image, the bilinear attention would be able to highlight the correct regions of the image, mainly paying attention to the specific words used in the question that point toward those regions.

The benefits of the BAN model are quite numerous. Through the use of bilinear attention, it is able to provide more context-sensitive answers that would better represent an understanding of the visual and textual inputs. This leads to better performance in VQA tasks since the model is hence strong in its determination of the relationships between the elements within the image and their linguistics, which are the corresponding parts in the question. Consequently, it has been revealed that the BAN model outperforms many current traditional models in VQA, especially where the posed questions are complicated, bringing about an imperative need for a deep understanding of the visual scene.

6.3. MULTI-SCALE FEATURE EXTRACTION MODEL:

Multi-scale feature extraction [16] is a sophisticated technique within the world of computer vision, extensively utilized by VQA models to make significant strides in terms of comprehension of images. This new feature enables the model to study features at various degrees of detail, capturing features at levels of granularity and larger structures, namely global features, simultaneously. These diversified levels of information get integrated within the model and mark an enhancement in the correct answering ability for complex visual questions. Challenges of VQA, which requires answering based on the content in visuals, are somewhat demanding for models. Many of the VQA models suffer from problems like partial occlusion in an image in which objects could be hidden partly from view, variation in the size of the objects, causing them to interpret things incorrectly, and details with subtle features in a scene requiring subtle knowledge. The extraction of multi-scale features solves all such problems as these models are now able to analyse images at any resolution. This capability helps the model maintain awareness of the larger context of the image as well as the finer details that might be necessary to answer particular questions. Practical applications of multi-scale feature extraction are very vast and varied. For example, in education, VQA models can be applied to interactive learning environments by giving detailed explanations based on visual aids. These can be particularly helpful in medical imagery for boosting medical-diagnostic procedures by correctly interpreting MRIs or CT scans where both the small anomaly and larger anatomical structure need to be made sense of. In autonomous systems, recognizing and responding to a wide variety of visual stimuli in real time is needed for

safety and efficiency in an application such as self-driving cars. In general, the incorporation of multi-scale feature extraction into VQA models not only enhances their performance but also expands their applicability across numerous domains, making them invaluable tools in both research and practical scenarios.

6.4. BERT MODEL:

Bidirectional Encoder Representations from Transformers (BERT) model. Devlin et al. made a remarkable achievement in NLP when it published its BERT model in 2018 [17], gaining so much attention and great recognition through their new innovative approaches [18] toward the comprehension and processing of human language. Unlike most models, BERT is different, as it typically processes the text in a bidirectional mode of transformation. This has helped BERT capture most of the relationships between different words in the text, leading to a better understanding of most linguistic nuances, which ultimately allows BERT to achieve a high level of performance in many tasks of NLP.BERT has greatly helped improve the performance of QA systems by doing justice to the interpretability of context-related queries and responses within that context. It has also shown great progress in sentiment analysis, so that text may not only be able to reflect a surface expression of emotions but also to understand deeper expressions of emotion. BERT has been useful in the creation of more accurate categorizations of documents and contents for text classification. It has also made great strides in VQA, which integrates visual data with textual queries to produce contextually appropriate responses. The paper delves deep into BERT's complicated architecture. It explains in considerable detail the different components of BERT and their mechanisms. In particular, the contribution BERT brought to the VQA capabilities was to incorporate processing and evaluation of visual input along with the textual one. Still, as with these improvement gains, it faces a certain set of implementations and real applications, coupled with limitations, from which its applications suffer in real-life situations.

6.5. ResNET50 MODEL:

This is one of the advanced deep convolutional neural networks [19] (CNNs) that have made significant contributions to all kinds of computer vision tasks such as image classification, object detection, and VQA. ResNet-50 is a deep neural network introduced by Kaiming He and his team in 2015 [20]. This has 50 layers, and the main drive behind its construction was to eliminate the vanishing gradient problem faced by deep networks, which delays the training as the depth of the network increases. To overcome this challenge, ResNet-50 introduced the concept of residual

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

learning. The concept allows deep CNNs to effectively train a network that had residual functions referencing its layer inputs instead of output; thus, the very deep degradation of the deep nets could easily be overcome. Based on that ResNet architecture includes convolution layers in combination with some batch normalization, some activation functions alternately connected and placed in sequence; it applies to ResNet50 architecture. These skip connections facilitate the flow of gradients during backpropagation, ensuring that the network can learn effectively even as it scales in depth. ResNet-50 improves both the training procedure and the entire performance of the model on other benchmarks for the VQA setting. In addition to improving training, ResNet-50 has been used greatly for feature extraction in many state-of-the-art approaches for ResNet-50. The primary requirement for an image is being processed and a high-level feature of visual for multimodal. The model is required to know and incorporate knowledge from the both visual and text inputs. Better and robust performance towards question answers with regard to the images by use of ResNet-50 because of the highly potent capabilities it uses to do feature extraction. Though this network has its numerous strengths, on the other hand, there is still room for some shortcomings such as ResNet-50 also faces vanishing gradient problems despite managing to cope up well and solving some instances but failing when faced with the situation and scenario types requiring an entirely much more contextual sense. In addition, the higher complexity of the model can easily increase the demand for computation that may be too high for certain real-time applications or devices. In short, ResNet-50 is a landmark in deep learning and computer vision, which provides a strong framework for solving complicated tasks like VQA. Its novel architecture along with residual learning opened up pathways for deeper nets and current research into shortcomings and areas it can be further improved within encourage further advancement in the field.

7. LIMITATIONS OF THE METHODS USED

Table 2: Comparison of datasets on some features.

MODELS	LIMITATIONS
BAN (Bilinear Attention Network)	High computational cost due to bilinear pooling operations
	Requires powerful GPUs and large memory for training
	Difficult to deploy in real-time applications due to latency
	Large-scale datasets make training and deployment complex.
	Requires model pruning and optimization for feasibility
	Hard to implement optimizations due to architectural complexity

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

Multi-Model Approaches	Extremely high computational cost due to multiple deep learning architectures
	Requires powerful GPUs, large memory, and long training times
	Prone to overfitting, especially on domain-specific datasets
	Struggles with long and complex questions, particularly multi-step reasoning
	Issues with bias and data imbalance, leading to unreliable answers
	Dependency on large-scale annotated datasets for effective training
	Lack of external knowledge integration, limiting real-world applicability
	Poor interpretability and explainability, making debugging difficult and reducing trust in critical applications
BERT	Cannot process images, requires a vision model for integration.
	Struggles with aligning text and visual features
	High memory usage and slow inference.
	Answering varies with slight changes in question phrasing.
	Limited ability to answer questions requiring real-world knowledge.
	Struggles with complex logical and sequential reasoning.
	Requires fine-tuning for specific VQA domains (e.g., medical, education).
	Acts as a "black box," limiting answer interpretability.
	Slow training and inference limit real-time applications.
RoBERTa	Lacks direct visual processing, focused only on text.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

	Poor fusion of textual and visual information.
	Computationally expensive due to larger training datasets.
	Performance fluctuates based on question complexity and domain.
	Relies on pre-trained datasets, limiting external knowledge integration.
	Lacks deep logical reasoning capabilities.
	Overfits to specific pretraining data, reducing adaptability.
ResNet-50	Extracts image features but lacks deep semantic understanding.
	Does not integrate with text-based features naturally.
	Requires large computational resources for real-time VQA
	Struggles with small variations in visual features.
	Cannot infer relationships between objects without explicit annotations.
	Cannot process multiple object relationships efficiently.
	Fails to generalize well beyond its training dataset.
	Hard to explain feature extraction decisions.
	High computational demand makes real-time VQA difficult.
	May misinterpret culturally diverse images.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 5, May 2025
PP 62– 81

8. Results and Analysis

Table 3: Accuracy (%) of models against datasets

Dataset	BERT [17]	ResNet50 [20]	RoBERTa [21]	BUTD [14]	Multi-Scale Feature [16]	BAN [15]	Mixed Attention [22]
VQA	75	62	78	70	80	72	85
Visual Genome	65	60	68	0	72	0	77
COCO-QA	68	58	70	65	74	67	79

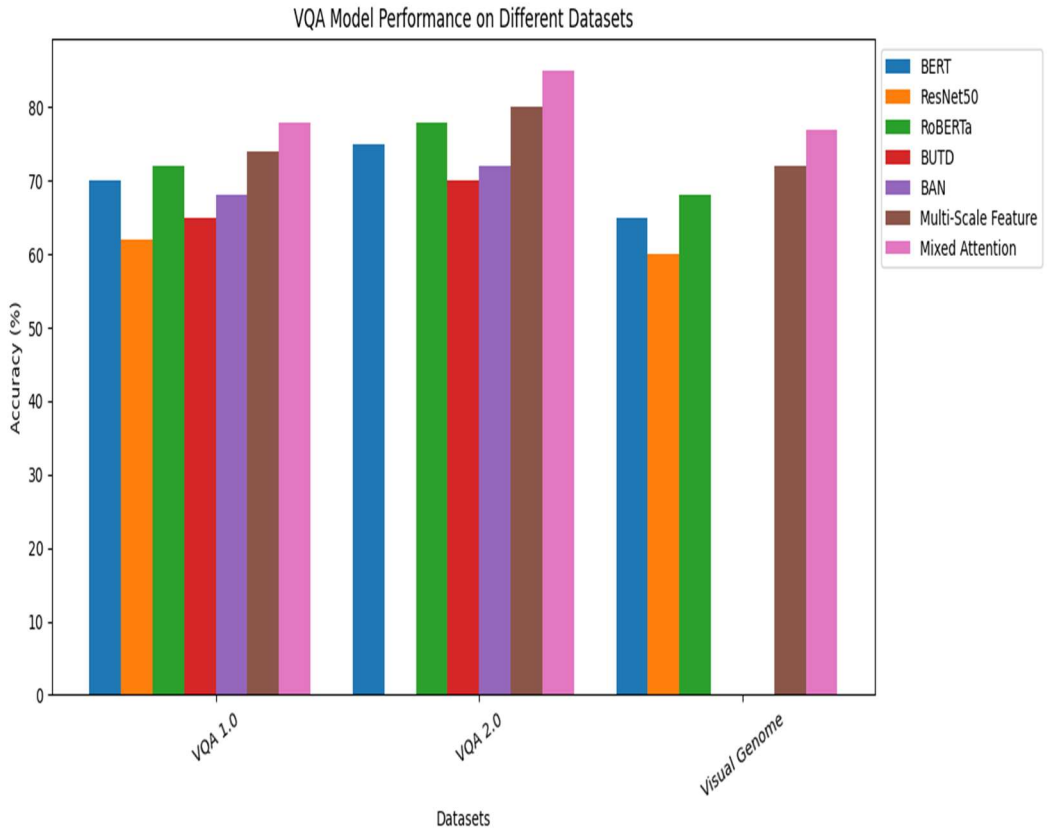

Figure 1: Graph of the Accuracy (%) of models against datasets

Figure 1 gives the graphical representation of Table 3, where we are giving the values of the accuracy (%) measured for different models through datasets.

8. CONCLUSION

The development of an AI-backed interactive VQA system is very advantageous for VQA systems engaged in the educational sector. While the objective statement of this paper will discuss the system's advantages in enhancing engagement, learning, and interfaces across a range of disciplines in ever-changing contexts for images, a unique aspect of our VQA model is that it works in real-time [23], allowing users to upload images or instantly capture pictures, pose questions, and get accurate answers. An added feature of the system is an accuracy-checking mechanism that verifies the reliability of generated responses for quality educational assistance. However, the system assistant is faced with challenges like poor dataset availability, interface accuracy and usability difficulties, limited computational resources, and optimization for engagement. Some existing VQA models like BUTD, BAN, and multimodal systems seem promising but share issues of poor reasoning, executing capabilities, and low generalization levels. Nonetheless, the performance of VQA models in the education sector stands to benefit from the adoption of effective architectures using multimodal transformers and attention mechanisms. Full realization of AI in educational use would require a rethink of the effectiveness of VQA systems at all education levels with respect to the following inclusivity issues: the approach will have to shift away from one of enhancing dataset variety and multilingual support towards one focused on robust feedback systems with real-time adaptability, therein making the systems more inclusive, scalable, and accurate for diverse learning environments.

9. REFERENCES

1. Lin, F. (2023). Research on the Teaching Method of College Students' Education Based on Visual Question Answering Technology. International Journal of Emerging Technologies in Learning (iJET*)*, *18*(22), 167-182.
2. J.Unni, S. J. (2023). Towards Robust VQA: Evaluations and Methods (Master's thesis, Arizona State University).
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425-2433).
4. Li, J., Thota, M. K., Gokhman, R., Holik, R., & Zhang, Y. (2024, December). SparrowVQE: Visual Question Explanation for Course Content Understanding. In *2024 IEEE International Conference on Big Data*

(BigData) (pp. 1814-1823). IEEE.

5. Lee, G., & Zhai, X. (2025). Realizing visual question answering for education: GPT-4V as a multimodal AI. *TechTrends*, 1-17.
6. Ishak Ali, S. I. S., Praneel, A. V., & India, V. A. Multimodal Fusion in Visual Question Answering: A Comprehensive Review of Approaches, Datasets, and Applications.
7. Agrawal, M., Jalal, A. S., & Sharma, H. (2023, October). A Review on VQA: Methods, Tools and Datasets. In 2023 International Conference on Computer Science and Emerging Technologies (CSET) (pp. 1-6). IEEE.
8. Zou, Y., & Xie, Q. (2020, December). A survey on VQA: Datasets and approaches. In 2020 2nd International Conference on Information Technology and Computer Application (ITCA) (pp. 289-297). IEEE.
9. Patadia, D., Kejriwal, S., Shah, R., & Katre, N. (2021, December). Review of vqa: Datasets and approaches. In 2021 International Conference on Advances in Computing, Communication, and Control *(ICAC3)* (pp. 1-6). IEEE.
10. Kafle, K., & Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, *163*, 3-20.
11. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & Van Den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, *163*, 21-40.
12. Koshti, D., Gupta, A., Kalla, M., Kanjilal, P., Shanbhag, S., & Karkera, N. (2024). EDUVQA–Visual Question Answering: An Educational Perspective. Journal of Advanced Research in Applied Sciences and Engineering Technology, *42*(1), 144-157.
13. Gupta, M., Asthana, P., & Singh, P. (2023). EDUVI: An Educational-Based Visual Question Answering and Image Captioning System for Enhancing the Knowledge of Primary Level Students.
14. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6077-6086).
15. Kim, J. H., Jun, J., & Zhang, B. T. (2018). Bilinear attention networks. Advances in neural information processing systems, 31.
16. Ma, Y., Lu, T., & Wu, Y. (2021, January). Multi-scale relational reasoning with regional attention for visual question answering. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 5642-5649). IEEE.
17. Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, No. 2).
18. Ishmam, M. F., Shovon, M. S. H., Mridha, M. F., & Dey, N. (2024). From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. Information Fusion, 102270.
19. Cheng, Y. (2023). Application of a Neural Network-based Visual Question Answering System in Preschool Language Education. IEIE Transactions on Smart Processing & Computing, 12(5), 419-427.
20. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
21. Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 364.

22. Yu, D., Fu, J., Mei, T., & Rui, Y. (2017). Multi-level attention networks for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4709-4717).
23. Ritvik, D. S. P., Praneel, A. V., & Ramaiah, P. (Year). Spatio-Temporal Attention Mechanisms in Video-Based Visual Question Answering: A Comprehensive Review.