The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

Multimodal Fusion in Visual Question Answering: A Comprehensive Review of Approaches, Datasets and Applications

Syed Ishak Sha Ali[1], A.S. Venkata Praneel[2]
[1,2]Department of Computer Science and Engineering,
GST, GITAM (Deemed to be University), Visakhapatnam AP India.
Syedishakshaali4588@gmail.com[1], praneelsri@gmail.com[2]
Corresponding Author *: praneelsri@gmail.com

Abstract

Multimodal fusion is, therefore, the backbone of all the breakthroughs in Visual Question Answering, in which models are able to mix visual and textual input for reasoning and prediction with tremendous effectiveness. This article comprehensively evaluates the environment of multimodal fusion in VQA through examining datasets, procedures, metrics, and applications together with normal hurdles. Key datasets such as VQA v1/v2, GQA, CLEVR, OK-VQA, and TVQA are assessed based on their size, diversity, and reasoning complexity to determine their sufficiency for creating multimodal research. The solutions covered include transformer-based models like LXMERT (Learning Cross-Modality Encoder Representations from Transformers) and ViLBERT (Vision language model), graph-based approaches such as VQA-GNN, and efficient multimodal fusion techniques, including bilinear pooling and attention mechanisms, which have considerably enhanced model performance. It will be scored using accuracy, consensus accuracy, BLEU, CIDEr, ANLS, and F1-score metrics in acquiring better knowledge of the utility of comparing models on various datasets. VQA will be further explored at the applications that function via healthcare, education, e-commerce, and scientific research; it also brings out disruptive power into real-world implementation. Although AI has made enormous achievements, challenges such as dataset biases, limited reasoning capabilities, poor fusion methodologies, lack of incorporation of external knowledge, and high processing costs remain important hurdles toward the development of completely robust and scalable models. This review intends to provide a full overview of the topic, so pushing the future research into overcoming these limits and realizing the actual potential of multimodal fusion for VQA and more.

Keywords:
VQA, Multimodal Fusion, Attention Mechanism, Transformer Based Model, Video Based Reasoning Models

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

1. Introduction

Multimodal fusion is the heart of the VQA challenge, where computer vision and natural language processing come together to deliver responses based on visual input. VQA was first proposed in the v1 dataset of VQA but has since become a benchmark for testing reasoning, comprehension, and integrating multimodal material [1]. The task is naturally complicated, requiring models to extract and align textual and visual information, reason over their relationships, and also yield consistent results. Data-sets like VQA v2 [2], CLEVR [3], GQA [4], OK-VQA [5], and TVQA [6] have, over time, enlarged the scope of the domain, bringing in varied difficulties ranging from compositional reasoning to knowledge-based inference. The core of multimodal fusion lies in merging visual and textual characteristics to generate a cohesive representation that enables successful thinking. Transformer-based models, such as LXMERT and ViLBERT, vastly expand this domain using the self-attention and the cross-attention process where interactions between modalities capture these interactions [7], [8]. Graph-based models like VQA-GNN use scene graphs, for modeling object relationships as this enables relational reasoning towards more complex questions [9]. This process is again advanced by the attention process that includes co-attention and self-attention wherein focus is dynamically drawn: important parts of the image and significant aspects of the text [10]. These advances have been accompanied by the advances in the efficient fusion schemes, like bilinear pooling and balancing computing efficiency with performance [11]. Datasets fuel VQA advancement. The development of the early datasets that include VQA v1 and COCO-QA provided large-scale collections focused on the general visual understanding of the image-question answer triplets [1], [12]. CLEVR introduced Synthetic scenes to analyze the compositional reasoning in manipulated environments. GQA reinforced the reasoning abilities further, using real images with their scene graphs and compositional queries. OK-VQA pushed the limits of these models by including external knowledge to answer questions that a content cannot solve. Instead, TVQA and MovieQA extended VQA in the temporal setting, meaning reasoning over sequences of video and related subtitles [6, 13]. Still, numerous challenges exist. Dataset biases, such as in VQA v1, help the models take advantage of statistical patterns instead of actual multimodal reasoning. Poor fusion methods paired with low-level reasoning capacity prevents more sophisticated tasks specifically in relational or commonsense reasoning [9, 10]. Again, another barrier to integrating external knowledge limits the domain, which is a limitation to implementing OK-VQA for the following domains contextual information. Advanced transformer-based and graph-based models have large computational costs [7, 9], restricting scalability. This review seeks to provide a full synthesis of the area, ranging from datasets to methodology, metrics, and applications. By

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

identifying present trends and obstacles, it is believed that future research can be steered toward bettering the efficiency, accuracy, and fairness of multimodal fusion for VQA.

DATASETS

1. VQA-v1

VQA v1 was one of the largest databases built with the goal of visual question answering. This real-world photo collection is paired with open-ended questions and their applicable responses. Therefore, it can be claimed that this is a basis set for testing VQA models. It can be divided into two classes: real images and abstract images. Real Images: These include complex and varied scenarios involving various objects, properties, and relationships. Abstract Images: Synthetic scenes designed using clipart software that enables the design of controlled experiments in a reduced framework. Questions in VQA v1 are human-written via sites like Amazon Mechanical Turk, where tasks were given to the employees to come up with hard questions that may potentially "trip up a smart robot." The dataset: questions on objects, activities, attributes, and counting. It also suffers from biases, including a high percentage of "yes" replies to yes/no questions and the “2” popularities for number-related issues. These biases let models leverage statistical trends in the data without understanding the images themselves.

2. VQA-v2

VQA v2 is actually built in order to overcome all the biases as well as limits observed from the VQA v1 approach. The progress here for the result presented is developed by the utilization of complement question-image pairings. Now here, for two photographs that are similar yet look quite same, simply a single question is answered however two viable lawful replies will exist. This structure ensures that the model has to look at the image to reply and therefore supplies VQA v2 with a more complicated set of images for evaluation tied to reasoning about images. Apart from balancing the dataset, VQA v2 shares the same scale and diversity as VQA v1 by using real-world scenarios with rich visual content. The architecture and updates of the dataset make it a gold standard for training and testing VQA models.

3. GQA

GQA is a dataset that focuses on compositional reasoning and mixes real-world images with rich structural representations to evaluate a model's ability to conduct multi-step reasoning. It aims to assess the compositional understanding of VQA models by presenting questions that demand the integration of many reasoning techniques. The collection contains 22 million questions associated with 113,000 photos obtained from real-world sources such as the Visual Genome dataset. GQA uses scene graphs to represent the structural representation of the image, including objects, their properties, and relationships. GQA also gives functional programs for each question, which specifies the step-by-step reasoning process required to achieve the solution. For instance, a query such as "What is the color of the thing to the left of the table? would entail the steps to identify the table, point to the object to the left of the table, and determine its color.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

Unlike synthetic datasets such as CLEVR, GQA uses real photos to bring greater diversity in classes, connections, and attributes to make it richer and complicated. GQA questions are generated automatically yet remain comparable to normal speech and incorporate numerous skills such as pointing to things, spatial thinking, and relationship comprehension. GQA It presents itself pretty effectively with a structured architecture for the evaluation of VQA models, seeking to enhance the skill of reasoning while working towards the relief of concerns like ambiguity and dataset bias.

4. COCO-QA

COCO-QA is a question-and-answer image dataset based on the corpus of pictures of the MS COCO database. The dataset contains 123,000 images accompanied by artificially created questions and their answers. This is created by converting image captions to questions using pre-existing templates. Questions in COCO-QA revolve around simple visual actions like object existence, such as "What is seen in the image?" Number Questions: For example, "How many people are there?" Color Questions: For example, "What is the color of the car?" Location Questions: For example, "Where is the person standing?" COCO-QA's simplicity lies in its few question types and, in that respect, is good for testing core capabilities in visual reasoning. However, the reliance on automated template-based question production limits the diversity and complexity of the collection compared to more current datasets like VQA or GQA. Despite its simplicity, COCO-QA is a useful resource for evaluating early-stage VQA models due to its clear architecture and concentration on foundational VQA activities.

5. CLEVR

CLEVR is a synthetic dataset intended exclusively to evaluate a model's capacity to execute compositional reasoning. It provides a controlled environment where questions involve logical, spatial, and relational thinking about simple geometric objects in computer-generated settings. By focusing on compositional reasoning, CLEVR examines whether models can integrate several reasoning stages to arrive at the correct answer. The dataset includes 70,000 images and 1 million questions produced using preset templates. Each image features simple forms (e.g., spheres, cubes, cylinders) with varied sizes, colors, materials, and spatial layouts. CLEVR questions are designed to examine specific reasoning talents, such as counting, such as “How many spheres are in the image?” Comparison: “Is the red object larger than the green object?” Spatial Reasoning: “What is the color of the object to the left of the cube?” Logical Reasoning: “Are there more cubes than metallic objects?” A fundamental feature of CLEVR is that it avoids biases typical in real-world datasets by assuring a fair distribution of question types and replies. Moreover, each question is accompanied by a functional program describing the reasoning steps needed to answer it. For instance, solving a spatial query might involve object location, neighbor detection, and attribute analysis for an object. CLEVR is helpful, especially in the context of reasoning-focused models, as the set of isolated tasks cancels out the complexities and biases that arise in natural datasets of pictures. It is an important dataset for developing and evaluating such VQA models based on logical reasoning and multi-step problem-solving.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

6. OK-VQA

OK-VQA is a knowledge-based visual question-answering dataset that evaluates capabilities for answering questions that require outside facts that are not observable from the image. OK-VQA differs from most VQA datasets, which are image-only reasoning, by allowing questions whose answers cannot be derived through mere reasoning over the contents of the image but actually require retrieval of pertinent information from resources external to itself, like an encyclopedia or commonsense knowledge base. The collection contains 14,031 questions over 14,031 images drawn from the MS COCO dataset, including topics that include history, geography, science, and cultural background. OK-VQA questions include facts and common sense, thus necessitating models to introduce dynamic multimodal fusion with external knowledge retrieval. For example, a question such as "What country uses this currency?" involves the model identifying the currency in the picture and retrieving geographical information, whereas "Why would this dog be wearing a vest?" includes the common-sense application with service animals. This distinctive approach highlights the burden on the dataset beyond the sheer interpretation of images. This dataset is essential, as it points to a much-needed gap in traditional VQA models, inspiring the creation of systems that could integrate visual recognition with the integration of external knowledge. It makes the model acquire and use external information dynamically and acts as a major benchmark for furthering research in knowledge-aware reasoning. It also strengthens the generalization skills of VQA systems.

Table 1: Comparison of Models using datasets and their parameters

Model	Dataset	Test-dev	Test-std	Y/N	Num	Other	All
LXMERT	VQA v2	87.0	87.1	89.3	54.2	63.5	73.8
ViLBERT	VQA v2	86.5	86.6	88.9	53.7	62.9	73.3
BAN	VQA v2, GQA	82.3	82.4	87.1	52.4	61.8	72.0
VQA-GNN	GQA, CLEVR	84.1	84.0	86.7	56.2	63.7	74.5
MoReVQA	TVQA	74.5	74.8	-	-	-	74.8
HCRN	TVQA	72.4	72.7	-	-	-	72.7
OK-VQA Baseline	OK-VQA	38.6	38.7	-	-	-	38.7

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

VQA-Large Language Model	OK-VQA	41.3	41.4	-	-	-	41.4
MFB+ATT	VQA v2	84.0	83.3	84.0	39.8	56.3	65.9
MFH+ATT	VQA v2	85.0	85.5	85.0	39.7	57.4	66.9
HieCo-ATT	VQA v2	79.7	-	79.7	38.7	51.7	61.8

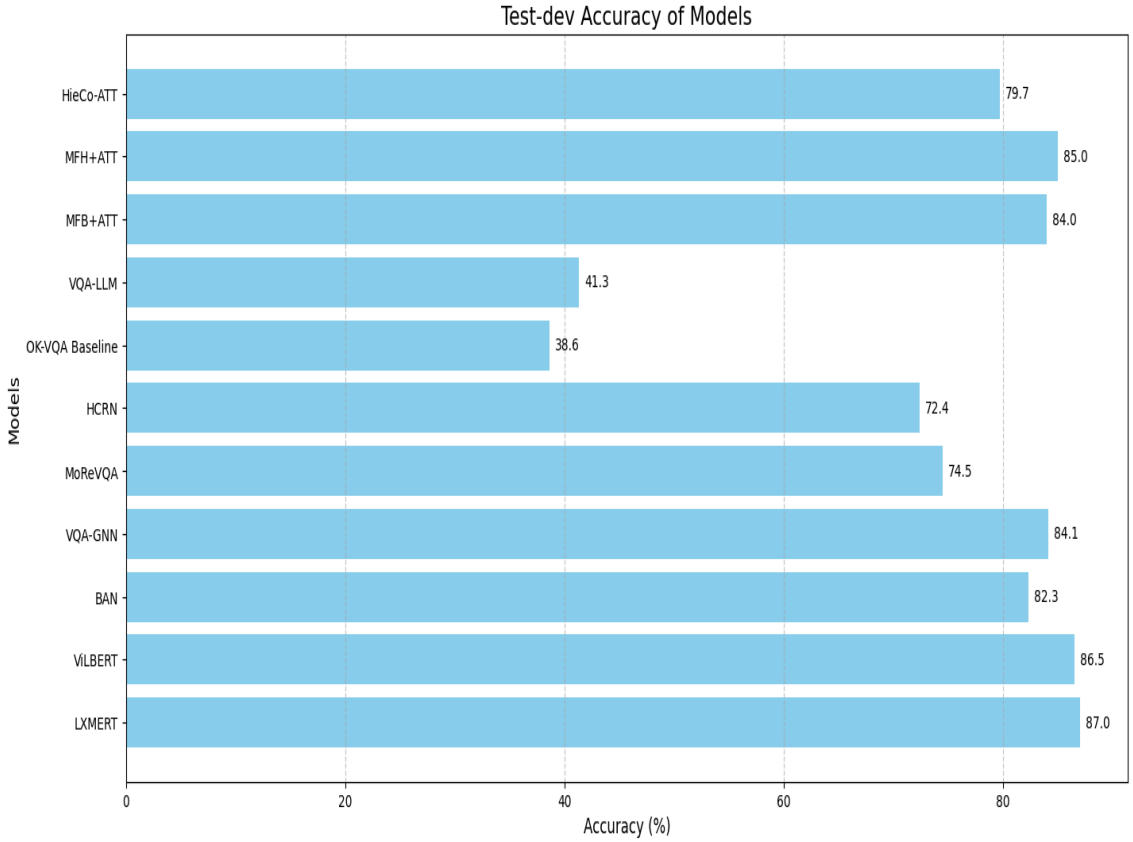

Fig 1: Illustrates the graphical representation of the Test-dev Accuracy of the Models.

APPROACHES:

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

1. Feature Extraction Method in VQA

Visual Question Answering systems rely on feature extraction as a key method. The system rapidly extracts key features from input-visual and input-textual to enable collaborative reasoning. Translation of raw input data into a structured representation from which crucial information would later be extracted for processing is sometimes involved in feature extraction. In the case of VQA, this technique is very important because it reveals just how well a model would read and assimilate multimodal input, that is, visual features from images and language features from questions.

1..1 Visual Feature Extraction

Visual feature extraction means that meaningful information is being drawn out of the images, which are later used to generate answers to questions. This method mostly begins with feeding raw pixel input into convolutional neural networks or, in newer contexts, Vision Transformers, to analyze the input. Those networks are trained to recognize a variety of visual features, including objects, shapes, textures, and spatial relationships between different objects within an image. For instance, the earlier versions, VQA v1 and VQA v2 models, used CNN-based models, either ResNet or Inception, to extract visual features that focus on item categories, locations, and their interaction. Then, these features are passed on to successive layers for additional processing or mixed with textual features. Scene Graphs, another basic idea, is the manner of hierarchical representation of visual features in which relationships between things within an image are specified. Scene graphs have been applied in numerous datasets, such as Visual Genome and GQA, to capture more in-depth knowledge regarding the image by identifying items and their interaction. In parallel, feature extraction from text works on turning natural language inquiries into representations that machines can handle. The traditional techniques were word embeddings such as Word2Vec or GloVe, but contemporary systems used transformer-based models like BERT or GPT [10]. These models capture the semantics of the words and context relation within a sentence, making them grasp the question's underlying meaning [11]. In VQA v2, textual features can be obtained using models like LXMERT or ViLBERT that simultaneously process both image and text data to give an overall representation [12]. This functionality integrates visual information with the text, efficiently representing it [13]. This is the point where features are obtained from sources, whether pictorial or text, and the most critical phase is a multimodal fusion [14]. It involves such a mode of fusion that these traits can be collectively reasoned to yield an answer [15]. Many multimodal fusion techniques, such as bilinear pooling, attention mechanisms, or graph-based fusion, have been studied to capture picture interactions with text elements [16]. For example, LXMERT and ViLBERT are systems that employ cross-attention strategies to align and merge visual information with textual information [17]. This merging allows the model to understand how different visual features relate to words in the question and to generate an appropriate answer [18].

2. Attention mechanism

Attention processes are crucial to VQA models as they can dynamically focus on the right portions of a picture and related textual pieces to make for correct multimodal reasoning [19]. Emphasis on specific visual characteristics or written words lowers irrelevant information and

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

thus increases prediction accuracy [20]. Earliest stages used discrete attention maps to highlight particular locations, whereas recent extensions covered continuous attention distributions, smoothing down the transition across the environment and capturing subtle interactions. Multimodal attention aligns elements from both modalities such that matching objects in images are aligned to words or phrases in questions, and, more recently, question-guided attention makes the focus of the model contingent upon query context. Such approaches help tackle challenges such as bias reduction due to less reliance on the linguistic priors, enabling complicated reasoning via the relationship-centric mappings, and improved interpretability by giving insights to decision-making. Applications include scene understanding and object detection with tasks like TextVQA, where attention is paid to regions with text for OCR-based reasoning.

3. Scene graph-based reasoning

Scene graph-based reasoning in VQA incorporates scene graphs and structured representations of images that reflect object-object relations, their attributes, and spatial knowledge [21]. The things are represented as nodes and the relations among them as edges. Thus, it allows models to reason about items in an image or the manner in which objects may be interacting with each other or the spatial arrangement. This approach is especially fruitful for compositional thinking and requires understanding the relationships amongst objects and their attributes. For example, a question like, "What color is that object to the left of the car?" requires the model to identify the car and find the object to its left, then determine its color. In scene graph-based reasoning, the models use either graph neural networks or other relational models for propagating information throughout the graph. This will enable the model to make inferences based on the relationships between the elements visually extracted from the image. Scene graphs give a comprehensive, structured manner of capturing visual context, especially important for more complicated reasoning tasks beyond simple object recognition. Models like GQA and VQA-GNN use scene graphs to describe the visual relationships in photographs of real-world situations, thus simplifying multi-step reasoning and compositional questions. By incorporating scene graphs, VQA systems can reason more sophisticatedly about visual scenes as interpreted by humans.

4. GNN

The concept graphs and scene graphs give organized representations, which create a framework for the strengthening of reasoning in VQA. These graphs consist of the objects, properties of that object, the relations existing within an image, as well as the knowledge being external but related to context of query and answer. In these graphs, information is transmitted by GNNs between nodes and edges, allowing the model to describe dependencies and interactions of discrete elements. For example, with VQA-GNN, unstructured multimodal data that contains image embeddings and the textual context can be aligned with information learnt from the graphs to build an integrated representation for reasoning. That is achieved through bidirectional fusion, where information moves between unstructured knowledge; that is, textual question-answer context and structured knowledge, such as object or relationship graphs [17]. This interaction is enabled by a central QA-context node, which links the question-and-answer context to the visual and conceptual nodes [18]. GNNs allow models to accomplish complex reasoning tasks like relational understanding, compositional reasoning, and answering questions that involve

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

accessing external commonsense knowledge by iteratively propagating information within the graph [19]. These qualities make GNNs particularly ideal for datasets that stress reasoning over multiple objects and relations, such as GQA and CLEVR [20]. The multimodal attention mechanisms further enhance the process by pointing out the relevant nodes and connections in question answering, boosting the interpretability and performance of these VQA models.

5. Transformer-based models

Transformer-based models are vital to modern Visual Question Answering systems because they can handle multimodal input by merging visual and textual data. Examples of such transformers are LXMERT and ViLBERT in which the picture and the question data are processed using two distinct encoders, and then their representations mixed for reasoning. These perform quite well in massive datasets tackling large-scale generalization and Completing difficult reasoning problems. The VQA-GNN concept introduces a fresh approach to the integration of transformers with GNNs. Textual data is encoded using RoBERTa, while multimodal reasoning is conducted using GNNs. Structured and unstructured knowledge would bridge this technique from scene graphs to text data. Though they are durable models, transformer-based models are nevertheless expensive in computation and require fine-tuning upon individual tasks. However, they remained vital in the purpose of advancing the boundaries of VQA and advanced reasoning in both visual and external knowledge environments.

6. Multimodal fusion

In the Visual Question Answering setup, multimodal fusion is a significant strategy in the fusing of features between the visual and textual modalities to answer a given question. Such a process usually occurs at the input level early, late level, or hybrid. Early fusion integrates visual and linguistic elements at the input level, whereas late-level fusion blends results later in the process. Hybrid fusion combines both tactics and exploits each's power. Co-attention and bilinear pooling were heavily used to increase multimodal fusion in VQA tasks. It can co-attentively focus on relevant picture regions, question phrases, and the bilinear pooling of fine-grained interactions between image and text features [13]. Transformer models, LXMERT and ViLBERT, adopted an approach with self-attention and cross-attention: they can process and fuse the features from images and text very rapidly, leading to state-of-the-art performance on complex VQA tasks [14]. Such models have been beneficial in tasks requiring compositional reasoning, knowledge-based reasoning, and bias mitigation on datasets like VQA v2, GQA, and OK-VQA [15]. The primary hurdles in multimodal fusion include the representation gap between images and text, enhancing the features' alignment, and the models' interpretability [16]. Despite these limitations, multimodal fusion does not stand still, and current research focuses on offering improved alignment approaches, dynamic integration of external knowledge, and greater generalization in many domains like scientific diagrams or video-based VQA [17,21].

7. Video-based reasoning model

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

Video-based reasoning models in VQA are different from typical image-based VQA systems since they combine temporal reasoning and knowledge of many frames to handle queries regarding dynamic settings. The model evaluates video sequences by their frames and audio or subtitles, and spatiotemporal elements are retrieved for reasoning purposes regarding visual and temporal contexts. Techniques applied to minimize the complexity of visual data include event parsing, frame grounding, and temporal attention procedures. For example, HCRN and EgoVQA follow hierarchical structures to track alignments between frames that are consistent with textual inputs. On a similar vein, modular reasoning approaches split the process of reasoning into interpretable sub-tasks such as key event extraction, grounding inquiries to the relevant frames, and reasoning about temporal linkages in MoReVQA (Modular Reasoning for Video Question Answering) [5]. These techniques have been examined on datasets like TVQA, MovieQA, and ActivityNet-QA focused on evaluating interactions, activities, and temporal sequences [6]. The video-based reasoning challenges include a considerable compute cost in handling big sequences, insufficient fine-grained capture of the underlying correlations between temporal values and, consequently, ensure well-generalizing models over diverse sources of videos [7]. Video-based reasoning models evolve further with new approaches deployed, such as transformer models for spatiotemporal fusion and pre-trained video-language models for comprehending complex scenarios in intricate dynamic scenes [8].

APPLICATIONS:

1. Healthcare

Visual Question Answering is one of the newest technologies that combines NLP and computer vision to answer questions based on visual content. These two modalities have allowed systems to perceive the world in ways that are remarkably close to human interpretation, employing visuals coupled with language. VQA models have demonstrated excellent potential in many sectors. Such sectors include health, education, accessibility, and so many more. The section highlights the diverse applications of VQA across various industries and its changing influence.

2. Education

In educational environments, VQA systems provide opportunities for students to learn interactively, whereby the student can query any instruction information, diagrams, charts, and scientific images. In doing so, interaction is enabled, accelerating their learning of hard concepts as they may ask questions and examine visual content [2]. For example, in biology class, chemistry class, or physics class, the complicated diagrams and VQA systems would be able to answer questions such as, "Which part of a cell generates energy?" [3]. In fact, VQA can improve individualized tutoring. It develops intelligent systems that respond to the student's demands for learning. Based on the context developed from visual aids and textbooks, the system generates domain-specific answers [6]. The relevance of VQA in education is that it bridges the

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

gap between visual and textual information, thereby making students highly engaged with knowledge expansion and interactive learning [7]. Models such as FigureQA, which were designed for scientific diagram interpretation, have demonstrated the potential of VQA to increase learning in professions that rely heavily on visual resources, such as medicine and engineering.

3.E-Commerce and Retail

In the e-commerce sector, VQA models are deployed in order to deliver an improved customer experience by enabling the customer to query the product by its image, boosting user delight and minimizing decision time. These models perform best with huge e-commerce websites that incorporate many products. The relationship makes shopping more fun, and it helps clients make more sensible purchasing selections. The significance of VQA in e-commerce resides in its capacity to enable buyers to interact with products through both visual and textual inquiries, so making the overall shopping experience much more interesting and driving more conversion rates in online retail. Models like ViLBERT and LXMERT, which specialize in processing multimodal data, would be great for this firm as they can assess and answer queries based on product images and descriptions.

4. Security and Surveillance

In the e-commerce industry, VQA models make the shopping experience better for customers. They can ask questions about products by looking at images, hence raising user happiness and assisting in decision-making, particularly in large online stores with vast inventory. Such types of questions may include "what colour is this chair?" and "what material is the handbag made of?". The VQA looks at the images of the product to come up with fitting answers. VQAs can enable virtual try-on in fashion e-commerce. That is, one can question how certain clothing items may look on oneself. This is of great benefit to the experience of customers while browsing, and they will make great purchasing decisions with better conversion rates. The VQA system, therefore, will allow customers to both visually and in writing communicate with the items so that their purchase becomes interesting and effective. Models such as ViLBERT and LXMERT have shown high performance when dealing with data from multiple modes. These can be embedded within the e-commerce website and will answer queries concerning a product based on pictures and descriptions. VQA is one of the best tools in security and surveillance systems, which can evaluate information from cameras to answer questions about the occurrence of an event or danger. The VQA system is used for the analysis of surveillance video, and it makes possible the tracking down of people or things. For example, it can answer questions like "Is there a suspicious person in this area?" or "What is the object near the door?" These VQA models will be helpful in reporting crimes and analyzing pictures or videos of break-ins or unauthorized access. Such VQA systems enhance situational awareness, thereby enabling quicker

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

decision-making and response to security threats [10]. Models such as VQA-GNN, which support relational reasoning using scene graphs, have proven to be pretty useful in the evaluation of complex interactions in surveillance feeds. As such, they are extremely valuable in applications where understanding spatial relationships and the interaction between objects is at the core [11].

5. Scientific Research

In the scientific sector, VQA models play a very essential role in assessing complicated visual data, including graphs, charts, and scientific diagrams, which enables researchers to form relevant conclusions and spot trends or abnormalities. Such systems allow for data analysis by answering inquiries concerning visual representations, making them particularly significant for studying massive datasets. For example, employing VQA models to analyze graphs and charts may reveal patterns or connections that could otherwise be time-consuming and labor-intensive to find manually. VQA systems can also be beneficial in grasping the nuances of scientific diagrams, such those in textbooks on physics or biology, by addressing particular questions concerning pieces and how they interact. This property accelerates the research cycle since it auto-interprets data; consequently, scientists may make intelligent conclusions much faster. Models like FigureQA, intended to cope with scientific diagrams, indicate an area of practicality for VQA in academic and research domains where interpretation of complicated visual data is vital, particularly in fields that mostly depend on ordered visual information.

METRICS:

1. Accuracy

Accuracy is one of the most common metrics used to evaluate VQA, which also remains the primary method of evaluating model performance. Accuracy is defined as the number of questions answered correctly divided by the total number. It is commonly used in datasets such as VQA v1 [2], VQA v2 [3], CLEVR [4], and OK-VQA [5]. The general approach of dealing with the evaluation of a model is usually found within its response compared to ground truth. In VQA v2, accuracy is dealt with by using consensus-based accuracy, where an answer would be considered correct if it aligns with at least three responses from human annotators to avoid linguistic biases. Accuracy is the most basic metric; however, it has some limitations. For instance, models could also have obtained high accuracies based on linguistic priors rather than true visual understanding, especially in datasets like VQA v1 [7]. In addition to accuracy, when it comes to more complicated reasoning tasks, for instance, OK-VQA, achieving accuracy alone is probably not the best measure for capturing the ability to reason because depth requires accuracy and other measures as well, such as Consensus Accuracy. So, accuracy is very important for general evaluation, and most of the time, some other metrics need to be supported to have finer

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

insight into how it works, especially when this task involves profound cognition or information from outside.

2. Consensus

Accuracy Consensus Accuracy: This statistic measure is widely utilized in VQA to check the agreement of a model's expected responses with human annotations. Here, unlike traditional accuracy, which only compares the predicted response with a single ground truth answer, consensus accuracy also considers replies that align with the majority view of human annotators. In datasets like VQA v2, an answer is valid if it matches up with three or more responses annotated by humans. This metric was designed to reduce the effect of biases present in datasets where some questions may contain multiple plausible answers. With consensus accuracy, the model is evaluated based on its response with respect to how much it agrees with an even more extensive scope of probable accurate answers that would aid in assessing whether it's truly anchored within the content of the image and not based on statistical linguistic priors. For example, in questions that may happen to be subjectively interpreted or have divergent answers, consensus accuracy obligates the model to follow the general human consensus, thereby making it a much more stringent and reliable evaluation metric, especially in datasets like VQA v2, for which bias mitigation should be the primary objective.

3. BLEU

BLEU (Bilingual Evaluation Understudy) is a fairly typical measure in VQA that examines the quality of machine-produced responses by assessing the n-gram overlap between the answers generated and the reference answers. It computes precision for various n-gram sizes—unigrams, bigrams, and trigrams and then averages them using the geometric mean [15]. In addition to these factors, it further limits the occurrence of answers that are too concise, which might pay in terms of precision but would suffer when considering completeness by employing a Brevity Penalty (BP)[16]. BLEU becomes a top contender in cases with high tens of thousands of viable formulation alternatives, favoring solutions that share n-grams with human-supplied answers. It does serve to quantify the quality of text generation. Still, BLEU cannot fully account for the semantic correctness or the deeper reasoning required by commonsense reasoning or knowledge-based VQA tasks, and so is usually used in conjunction with other metrics, including F1-Score and Plausibility, for complete evaluation [17].

4. CIDEr

CIDEr (Consensus-based Image Description Evaluation) is a metric used to measure the quality of captions machine-generated for images and natural language answers in applications such as Visual Question Answering (VQA) [18]. CIDEr measures how similar n-grams, sequences of words, in the automatically generated answer are to those in a reference set of responses, with an emphasis on consensus in the responses of people. Unlike other metrics, such as BLEU, which only look at precision, CIDEr also analyzes the relevance of n-grams by taking a weighted approach that favors often-used terms in the reference replies. This ensures that answers with

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

words and phrases that are more common and semantically meaningful are selected. CIDEr is especially beneficial in VQA tasks that involve image-grounded reasoning and natural language generation, since it penalizes models for providing poorly matched responses with human-written replies in terms of content and phrasing.

5. F1 Score

The F1-Score is amongst the most frequently used metrics to evaluate VQA systems and considers both precision and recall; it is given in Ref. [2]. Precision is simply the number of accurately predicted answers. Recall tells how many accurate answers the model actually predicts. The F1-Score combines the two by calculating the harmonic mean, which is a single value that may represent the accuracy as well as the completeness of the model's predictions. This is crucial in VQA tasks since there could be more than one correct solution, and just picking on correctness could capture less about everything in model performance. This is useful when the datasets are unequal or when some kinds of errors, like false positives or false negatives, are worse than others. In VQA, the F1 Score checks that the model provides correct answers without missing important answers, so it is good for tasks that have many correct answers or complex reasoning.

6. Plausibility

The Plausibility evaluation metric for Visual Question Answering measures the semantic accuracy or reasonableness of the model-generated answers. Unlike accuracy or BLEU, which measures only if the answer matches a reference answer, plausibility measures how reasonable and rational the answer is about the question and the image. A plausible answer is one that, with general knowledge and common-sense reasoning, makes sense and is logically reasonable in the light of the context provided by the question and the visual content. Plausibility is especially helpful in tasks such as knowledge-based reasoning and commonsense reasoning, where the model needs to provide responses that are not only syntactically correct but also contextually and logically sensible. For datasets such as OK-VQA, such answers typically contain some external knowledge or logical inference that cannot be inferred from what is apparent in the image.

7. ANLS

ANLS, or Average Normalized Levenshtein Similarity, is one measure employed to score the similarity of machine-generated answers with reference answers, mainly in a VQA context, specifically in datasets such as TextVQA [11]. A computation of Levenshtein distance, or the least number of insertions, deletions, and substitutions that must be done on each to turn one string into the other, ANLS normalizes this distance by the length of the reference answer. This will help control differences in the length of answers and maintain fairness while comparing answers with varying lengths. The ANLS score is averaged over all question-answer pairs in a corpus to determine how well the responses generated by the model mirror the reference answers. This statistic is extremely beneficial when the model discovers infrequent exact matches, and answers may be worded differently, hence preferring models that generate answers

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

with high semantic similarity with human-provided answers.

SHORTCOMINGS OF THE METHODS:

1. Database Bias

One of the major difficulties facing VQA systems is that of dataset biases because most models leverage inherent patterns in datasets rather than acquiring a real comprehension of multimodal material. For example, by overrepresenting answers such as "yes" for yes/no questions or "2" for counting-related queries, the model learns to rely on statistical shortcuts rather than image-grounded reasoning in VQA v1. This artificial inflation of accuracy scores, dependent on linguistic priors, masks the inadequacy of such algorithms to integrate relevant visual and textual data. Models generally fail to generalize in better-balanced datasets such as VQA-CP v1 and VQA-CP v2, in which the distributions of answers for every type of question are purposely altered between training and test splits. These biases dramatically skew the assessment of model performance and greatly limit their utilization in real-world scenarios with high visual diversity and complexity, where solid reasoning is the most crucial. As such, minimizing dataset biases has become a crucial part in constructing dependable multimodal reasoning abilities inside a VQA system; many studies focus especially on developing new datasets or models that are designed to mitigate such biases.

2. Limited Reasoning Capabilities

Limited reasoning is still an immense challenge with VQA models, as many of the proposed models fail to perform tasks that demand multiple steps and compositional or even temporal reasoning. Datasets like CLEVR and GQA demand a relational understanding of capability in sequential processing involved, such as the detection of objects and spatial relation analysis of interactions. While some models have used approaches such as scene graphs or graph neural networks (for example, VQA-GNN) to deal with these challenges, they generally fail when posed with more complex or compositional queries. Similarly, in video-based reasoning datasets like TVQA and MovieQA, where questions embody temporal dynamics over several frames, models like HCRN and MoReVQA indicate the intrinsic difficulty of correlating textual queries with spatiotemporal information. These activities necessitate deeper integration of visual and textual characteristics and the ability to persist context between frames, which most design failures currently lack. This creates limitations for the models in dealing with the complexity of real-life scenarios, which may come along with complicated reasoning interconnects or temporal dependencies, thereby requiring more significant levels of reasoning in more complex VQA systems.

3. Inefficient Multimodal Fusion

Inefficient multimodal fusion is a key difficulty in VQA models since it directly affects the system's capacity to align and integrate visual and textual information effectively. Traditional fusion techniques, such as early or late fusion, sometimes fail to capture subtle interactions between the two modalities. Early fusion will combine the characteristics at the input level, which could result in noisy representations, and the late fusion aggregates information after separate processing, therefore suffering from not being as deeply integrated. Advanced

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 1, January 2025
PP 119–137

approaches, such as bilinear pooling applied in BAN and cross-attention mechanisms used in LXMERT and ViLBERT, have greatly enhanced multimodal alignment by enabling models to focus on specific picture regions and dynamically on text tokens. These approaches grasp finer-grained relationships but increase computational complexities. They require significant amounts of resources for training and inference. This tradeoff constrains the ability to scale and even deploy it in resource-limited environments. Inefficient fusion especially excludes tasks that require precise multimodal reasoning, such as answering compositional questions in CLEVR or knowledge-based queries in OK-VQA, where delicate alignment of visual and textual properties is crucial. The solution to this problem involves the development of more efficient yet potent fusion algorithms that combine computational expense with performance.

4. Lack of External Knowledge Integration

The most fundamental limitation of current VQA models is their inability to include any sort of external information, which becomes an incredibly critical element when the task in hand involves thinking beyond the content of the input image or text. For example, datasets such as OK-VQA feature questions that demand commonsense thinking or unique knowledge of culture, history, or science that cannot be derived from the image or text alone. Since most of these models are not able to tap into external knowledge bases such as encyclopaedias, knowledge graphs, or commonsense resources, their responses are frequently rather limited in terms of accuracy. This weakness makes the applications of VQA systems very unsuccessful in real-world circumstances where contextual or encyclopedic knowledge is important for logical reasoning. Where as other methods attempt to introduce dynamic external knowledge using retrieval-based methods or pre-trained knowledge-aware word embeddings, they tend to be computational in nature and generally cannot rival their counterparts in more demanding knowledge-rich tasks. Therefore, this constraint must be solved to make the VQA systems move to more efficient and adaptive applications.

5. Computational Costs and Scalability

The main challenge to deploying such sophisticated VQA models is that they incur high computational costs and scalabilities. Models like LXMERT and ViLBERT, based on the transformer architecture, and graph-based reasoning models in the form of VQA-GNN are excellent models that exploit a complicated architecture demanding considerable processing resources. Such models require significant hardware support, such as GPUs or TPUs, for training and inference, infeasible for many academics and impossible for deployment in resource-constrained situations like mobile devices, embedded systems, or edge computing platforms. That was a particularly important limitation if applications need to have near real-time reactions: think autonomous systems or even providing accessibility assistance for the visually handicapped. Finally, a high computational cost limits scalability to much more extensive

datasets or much complicated reasoning tasks, like video-based VQA where processing temporal sequences increases the computational cost. Hence, there is a need for new architectures that are lighter models or approaches like model pruning, quantization, or distillation to balance performance with computing efficiency, thus making wider implementation of VQA systems possible in real world applications.

CONCLUSION:

VQA is a breakthrough technology that bridges the gap between visual perception and natural language understanding to enable major applications across multiple disciplines. In healthcare, diagnostic workflows are modernizing as VQA models allow immediate real-time analysis of medical pictures for diagnosis, enhancing accuracy and minimizing the time necessary for diagnosis. Similarly, in education, VQA makes learning interactive and personalized. Thus, it improves understanding of complex concepts since it can analyze and answer visual educational content. In e-commerce, these models have improved the experience of customers through intuitive image-based product interactions where customers make informed purchasing decisions. Second, security and surveillance depend on VQA to enhance situational awareness and incident response, from the analysis of video feeds to actionable insights. Lastly, VQA systems enhance data interpretation in scientific research so researchers can assess visual data, including graphs and diagrams, much better with speed and accuracy. Despite these advancements, some obstacles, including biases of the dataset, computational inefficiencies, and the inability to reason appropriately, continue to impede innovation in the VQA area. Such barriers will be eliminated, leading to reaching its complete potential. The increasing models for managing complex reasoning, the multimodal fusion method, and the efficiency of adding external knowledge would produce a system of VQA that is durable and scalable. Such diversifications of datasets, methods, and applications in various sectors shall continue to preserve VQA as a fascinating subject promising to be one of the pillars for multimodal AI research and real-world utility.

References

1. 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.

2. Kabir, R., Haque, N., & Islam, M. S. (2024). A Comprehensive Survey on Visual Question Answering Datasets and Algorithms. arXiv preprint arXiv:2411.11150.

3. de Faria, A. C. A. M., Bastos, F. D. C., da Silva, J. V. N. A., Fabris, V. L., Uchoa, V. D. S., Neto, D. G. D. A., & Santos, C. F. G. D. (2023). Visual question answering: A survey on techniques and common trends in recent literature. arXiv preprint arXiv:2305.11033.

4. Wang, Y., Yasunaga, M., Ren, H., Wada, S., & Leskovec, J. (2023). Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21582-21592).

5. Kafle, K., & Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding, 163, 3-20.

6. Ishmam, M. F., Shovon, M. S. H., Mridha, M. F., & Dey, N. (2024). From image to language: A critical analysis of visual question answering (QA) approaches challenges and opportunities: information Fusion, 102270.

7. Yusuf, A. A., Feng, C., Mao, X., Ally Duma, R., Abood, M. S., & Chukkol, A. H. A. (2024). Graph neural networks for visual question answering: a systematic review. Multimedia Tools and Applications, 83(18), 55471-55508.

8. 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.

9. Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., ... & Zhang, S. (2024). A comprehensive review of multimodal large language models: Performance and challenges across different tasks. arXiv preprint arXiv:2408.01319.

10. Lu, S., Liu, M., Yin, L., Yin, Z., Liu, X., & Zheng, W. (2023). The multi-modal fusion in visual question answering: a review of attention mechanisms. PeerJ Computer Science, 9, e1400.

11. Wang, Z., Chen, L., You, H., Xu, K., He, Y., Li, W., ... & Chang, S. F. (2023). Dataset bias mitigation in multiple-choice visual question answering and beyond. arXiv preprint arXiv:2310.14670.

12. Yigit, G., & Amasyali, M. F. (2024). From text to multimodal: a survey of adversarial example generation in question answering systems. Knowledge and Information Systems, 66(12), 7165-7204.

13. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2901-2910).
14. Zakari, R. Y., Owusu, J. W., Wang, H., Qin, K., Lawal, Z. K., & Dong, Y. (2022). Vqa and visual reasoning: An overview of recent datasets, methods, and challenges. arXiv preprint arXiv:2212.13296.

15. Hudson, D. A., & Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6700-6709).

16. Lei, J., Yu, L., Berg, T. L., & Bansal, M. (2019). Tvqa+: Spatio-temporal grounding for video question answering. arXiv preprint arXiv:1904.11574.

17. Khurana, K., & Deshpande, U. (2021). Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: a comprehensive survey. IEEE Access, 9, 43799-43823.

18. Yusuf, A. A., Chong, F., & Xianling, M. (2022). An analysis of graph convolutional networks and recent datasets for visual question answering. Artificial Intelligence Review, 55(8), 6277-6300.

19. Ren, M., Kiros, R., & Zemel, R. (2015). Exploring models and data for image question answering. Advances in neural information processing systems, 28.

20. Kahou, S. E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., & Bengio, Y. (2017). Figure: An annotated figure dataset for visual reasoning. arXiv preprint arXiv:1710.07300.

21.Dunga, S. V. P. R., Venkata Praneel, A. S., & Chowdary, P. R. (2024). Spatio-Temporal Attention Mechanisms in Video-Based Visual Question Answering: A Comprehensive Review. The Journal of Computational Science and Engineering, 2(10), 8–25. ISSN: 2583-9055.