INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

REVIEW ARTICLE

A Review of Recent Advances in Visual Question Answering: Capsule Networks and Vision Transformers in Focus

Miranda Surya Prakash[1]*, **S N Devananda**[2]

1 Research Scholar, Department of ECE, PES Institute of Technology and Management, Visvesvaraya Technological University, Shimogha, Karnataka, India
2 Professor, Department of ECE, PES Institute of Technology and Management, Visvesvaraya Technological University, Shimogha, Karnataka, India

OPEN ACCESS

Received: 18-10-2023

Accepted: 28-10-2023

Published: 30-12-2023

Citation: Prakash MS, Devananda SN (2023) A Review of Recent Advances in Visual Question Answering: Capsule Networks and Vision Transformers in Focus. Indian Journal of Science and Technology 16(47): 4525-4546. https://doi.org/10.17485/IJST/v16i47.2643

* **Corresponding author**.

mirashc23@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: Multimodal deep learning, incorporating images, text, videos, speech, and acoustic signals, has grown significantly. This article aims to explore the untapped possibilities of multimodal deep learning in Visual Question Answering (VQA) and address a research gap in the development of effective techniques for comprehensive image feature extraction. **Methods:** This article provides a comprehensive overview of VQA and the associated challenges. It emphasizes the need for an extensive representation of images in VQA and pinpoints the specific research gap pertaining to image feature extraction and highlights the fundamental concepts of VQA, the challenges faced, different approaches and applications used for VQA tasks. A substantial portion of this review is devoted to investigating recent advancements in image feature extraction techniques. **Findings:** Most existing VQA research predominantly emphasizes the accurate matching of answers to given questions, often overlooking the necessity for a comprehensive representation of images. These models primarily rely on question content analysis while underemphasizing image understanding or sometimes neglect image examination entirely. There is also a tendency in multimodal systems to neglect or overemphasize one modality, notably the visual one, which challenges genuine multimodal integration. This article reveals that there is limited benchmarking for image feature extraction techniques. Evaluating the quality of extracted image features is crucial for VQA tasks. **Novelty:** While many VQA studies have primarily concentrated on the accuracy of answers to questions, this review emphasizes the importance of comprehensive image representation. The paper explores recent advances in Capsules Networks (CapsNets) and Vision Transformers (ViTs) as alternatives to traditional Convolutional Neural Networks (CNNs), for development of more effective image feature extraction techniques which can help to address the limitations of existing VQA models that focus primarily on question content analysis.

Keywords: Feature Extraction; Visual Question Answering; Multimodal Deep Learning; Capsule Networks; Vision Transformer; Datasets

1 Introduction

Multimodal deep learning has witnessed significant advancements in recent years. It combines multiple modalities like images, texts, videos, speech, and even acoustic signals. Multimodal deep learning models include different tasks like Visual Question Answering, Image Captioning, Visual Question Generation, Visual Storytelling, Image-Text matching etc. These tasks combine information from different modalities like Image, Video and Text. Researchers have developed a range of innovative approaches to tackle challenges associated with multimodal tasks. This survey contributes to visual question answering. VQA is a multidisciplinary field that aims to develop systems capable of comprehending visual content and responding to questions posed in natural language. It faces unique challenges, requiring models to not only perceive and interpret visual information but also understand and generate textual responses.

While numerous studies have concentrated on developing fusion techniques to facilitate the integration of visual and textual information in VQA systems, there exists a conspicuous research gap pertaining to the comprehensive extraction of image features and development of effective techniques for extracting information from images. Most of the VQA research has focused on ensuring that answers accurately match the given questions, but they focus less on understanding image details. Typically, the model predicts answers primarily by analyzing the content of the question, with less emphasis on comprehensive image representation or without even fully examining the image.

This review article aims to address this gap by exploring the potential of multimodal deep learning for VQA, with a particular focus on enhancing image feature extraction. By emphasizing the extraction of detailed image features, we endeavor to refine the VQA process, ultimately enabling the generation of more precise and context-aware answers to questions posed about images.

The role of images in multimodal understanding is very crucial. Despite the challenges and concerns raised, images do offer benefits in multimodal understanding. They aid in disambiguation and provide contextual information to resolve language ambiguities. In cases of text imperfections or missing information, images can serve as crucial references. This survey reveals a tendency for systems to overlook or even neglect one modality, whether it be visual or textual, while placing excessive emphasis on the other. There exists a problem in neglecting one modality, particularly the visual one, in multimodal systems which pose a significant challenge to achieving true multimodal integration. While there are challenges to address, images continue to play a valuable role in enhancing our understanding of the world, even in the context of language processing.

Even with multimodal transformers and pre-trained models, the textual modality continues to dominate in decision-making [1]. This highlights the persistence of the inequality between modalities and the influence of data biases, which is an ongoing concern. To address these challenges, this survey paper explores recent breakthroughs in image feature extraction for VQA, with a specific focus on two promising alternatives to traditional CNNs: CapsNets and ViTs. This paper identifies and reviews the state-of-the-art multimodal deep learning models for VQA and explores emerging deep learning approaches for image feature extraction in VQA. A general framework of VQA is shown in Figure 1.

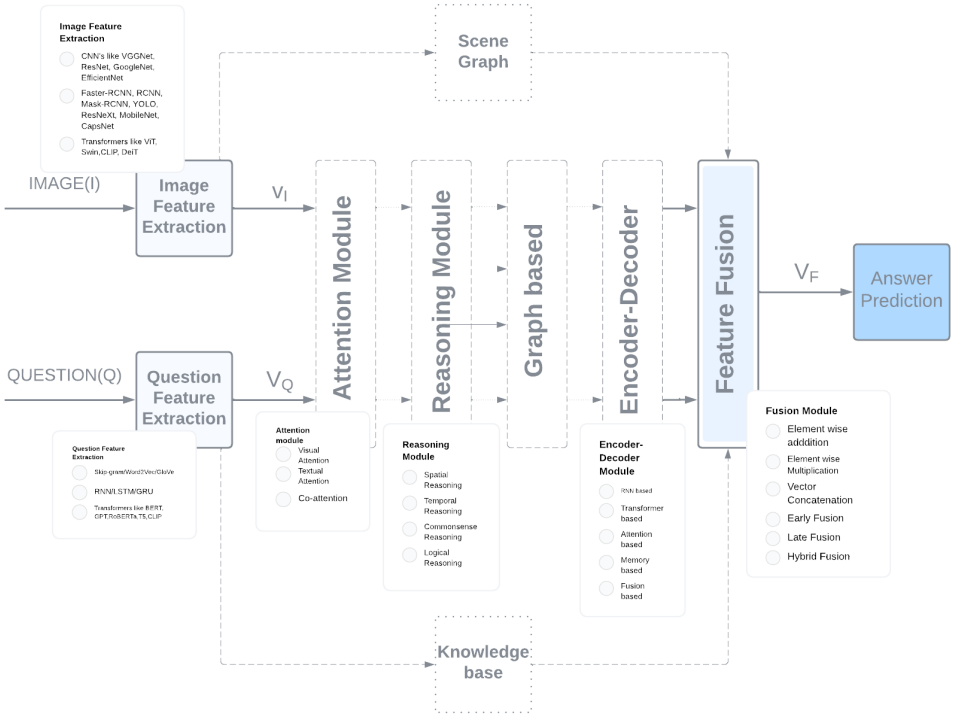

Fig 1. General framework of a VQA model

To obtain better image representations, we delve deeper into VQA models which use transformers. Over the years, CNNs have been the dominant approach for image feature extraction in VQA, achieving state-of-the-art results. Recent advances in deep learning methods offer comparable or better performance than CNN. Researchers are increasingly using transformers as an alternative to traditional CNNs for VQA tasks. Traditional methods, such as CNNs, have limitations in VQA, such as not being well-suited for capturing long-range dependencies in images and being difficult to train for tasks requiring reasoning about object relationships.

The main contributions of our review article are listed below.

- We introduce a novel taxonomy that categorizes various VQA techniques, applications, and challenges in the field of VQA.
- We provide a comprehensive overview of state-of-the-art multimodal deep learning approaches for VQA.
- We conduct an in-depth analysis of the image feature extraction techniques employed by various VQA models, datasets, and evaluation metrics used.

1.1 Essential techniques for Visual Question Answering

In this section, we present an overview of the major techniques employed in VQA tasks. The taxonomy of various techniques used for VQA is depicted in Figure 2.

1.1.1 Image featurization

Image featurization is the process of extracting features from an image that represents its visual content. These features can be employed to identify objects, attributes, and their respective relationships. The ability to encode visual information is essential for VQA models. It helps VQA models to gain the ability to ”see” and understand the contents of the image. VQA models use image features extracted from transformers like ViT or Swin and CNNs like Faster-RCNN, VGG, ResNet to represent the visual content of an image.

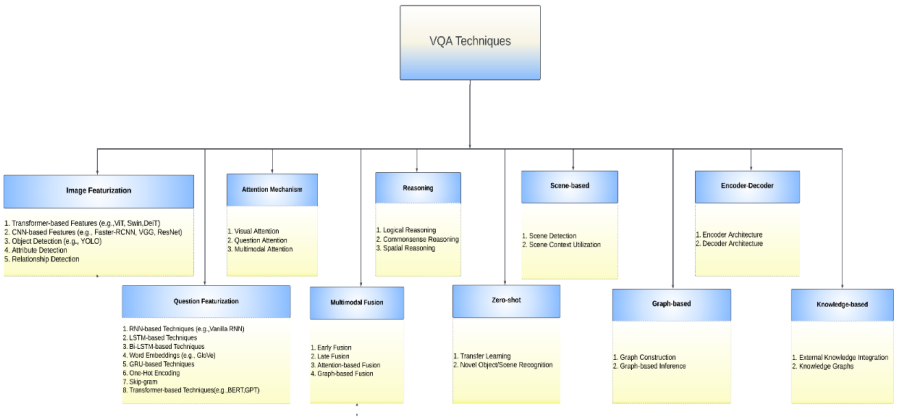

Fig 2. Taxonomy of various VQA Techniques

1.1.2 Question featurization

Question featurization is the process of extracting features from a question that represents its semantic content. These extracted features can be utilized to identify the entities, relations, and actions provided in the question. To understand the semantic content of the question, VQA models also need to convert the text-based question into a numerical representation The features can be extracted using a variety of NLP techniques which utilize Recurrent Neural Network (RNN) based models, Long short-term memory (LSTM), Bidirectional LSTM (Bi-LSTM), Global vectors for word representation (GloVe), Gated recurrent unit (GRU), One-hot embedding, skip-gram, and transformers.

1.1.3 Attention Mechanism

The attention mechanism is a technique which enables VQA models to focus on important parts of an image and question when answering a question, assigning weights to these relevant parts based on the question. This attention helps improve the accuracy of model by emphasizing important visual and textual contexts while answering the question.

1.1.4 Multimodal fusion

Multimodal fusion is a technique that allows VQA models to combine image and question features in a meaningful way. Fusion techniques like concatenation, element-wise addition, and bilinear pooling enable models to effectively integrate visual and semantic information. This would be particularly useful for questions requiring a joint understanding of image and question.

1.1.5 Reasoning

Reasoning is a technique that allows VQA models to reason about the content within an image and question to answer a question They can go beyond simple feature extraction and attention mechanisms to perform reasoning about content for both image and question. Techniques such as natural language inference (NLI) and commonsense reasoning enable models to infer the answer based on logical reasoning.

1.1.6 Zero-shot

Zero-shot approaches use transfer learning to answer questions about objects or scenes not in the training data. This involves using a trained model on a large dataset to generalize knowledge from similar tasks to unseen concepts. This feature is particularly useful for handling unexpected questions, as it allows models to answer questions about new objects or scenes that were not included in the training data.

1.1.7 Scene-based

Scene-based approaches identify the scene within an image using scene detection techniques. These techniques identify objects and relationships characteristic of a particular scene, providing context for answering questions. Scene-based VQA approaches also identify the scene within the image, by categorizing the image content for more accurate answers.

1.1.8 Graph-based

Graph-based approaches represent images and questions as graphs, where nodes represent objects and relationships, and edges represent connections between objects. These graph structures enable models to reason about image and question in a graph-based framework, facilitating more complex and structured reasoning processes. Graph-based VQA approaches represent both image and question as a graph structure, allowing more detailed and efficient answers to questions.

1.1.9 Encoder-decoder

Encoder-decoder is the most used architecture for VQA, where both images and questions are encoded into a shared latent space. The model then decodes this shared representation to produce the answer, allowing for seamless integration of visual and textual information, promoting accurate and context-aware answers. This approach is crucial for VQA applications.

1.1.10 Knowledge-based

Knowledge-based approaches use external knowledge from databases like Wikipedia to answer questions about images. This uses a technique, known as knowledge retrieval. It retrieves relevant information from knowledge base to answer the respective question. Knowledge-based VQA models use this external knowledge to provide answers that require a broader understanding beyond image and question itself.

1.2 General Challenges in Visual Question Answering

VQA requires a robust fusion of visual and textual information, handling ambiguous questions, addressing biases, and attaining consistent performance across various domains. Researchers are exploring novel approaches, such as multi-modal architectures, attention mechanisms, pre-trained language models, combining CNNs for image understanding with RNNs or transformers for language processing, reinforcement learning, knowledge incorporation, and data augmentation techniques to enhance VQA model's robustness and generalization capabilities. Let us go through a few related works and challenges in VQA. Despite significant progress that has been made, there remain several challenges that need to be addressed in VQA. Figure 3 illustrates several key challenges that arise in VQA tasks.

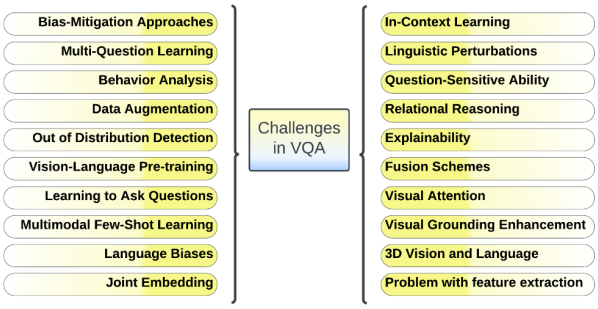

Fig 3. Key challenges in Visual Question Answering Tasks

In-Context Learning enhances VQA models by integrating strategies that enable them to produce more contextually relevant answers. Linguistic perturbations have been explored to evaluate and enhance the language understanding capabilities of models. For question-sensitive ability, few researchers have focused on enabling VQA models to exhibit question-sensitive reasoning abilities, leading to more accurate and relevant responses. Relational reasoning investigates relational reasoning capabilities which enables VQA models to understand and reason about complex visual scenes.

Explainability in VQA develops models that has led to improved user trust and understanding of model decisions. In bias-mitigation approaches, many efforts have been made to mitigate and reduce biases present in models, ensuring fairness and avoiding undesirable biases in responses. Multi-question learning approaches have improved the model's ability to handle multiple questions and contexts effectively. Behavior analysis analyzes the behavior of VQA models which has provided insights into their limitations, strengths, and potential areas for improvement. Utilizing data augmentation strategies has proven to be effective in enhancing the robustness and generalization capabilities of various models.

Developing models capable of detecting out-of-distribution samples has improved the reliability and robustness of systems. Integrating 3D vision and language understanding has expanded the scope of VQA to more immersive and complex scenarios.

Pre-training models on large-scale vision and language data has become a common practice, resulting in improved performance in different question answering tasks. Learning to ask questions explores the ability of VQA models to generate meaningful and relevant questions. It has opened new opportunities for research. The recent advancements in multimodal few-shot learning have enabled VQA models to generalize better with limited data. Analyzing and addressing language biases present in VQA models has become a critical aspect of research in the field.

The joint embedding techniques have facilitated the seamless integration of visual and textual representations in VQA. Few research is focused on utilizing different fusion schemes for combining visual and textual features, resulting in enhanced performance. Investigating visual attention mechanisms in different models has led to more accurate answers. In visual grounding enhancement, research is aimed at improving the ability of models to accurately ground visual elements in their answers. In addition to all these challenges, there are still problems with feature extraction. VQA models struggle to understand the relationship between the images and questions due to the differences in feature extraction formats.

1.3 Applications of Visual Question Answering

VQA has got a wide range of applications, including medicine, education, and robotics. Figure 4 illustrates diverse applications of VQA. It can assist doctors and medical professionals in interpreting complex medical images like X-rays, MRIs, and CT scans, producing accurate diagnoses and treatment strategies. By analyzing biomedical images, such as cellular structures, VQA can enhance the understanding of disease causes. In architecture, VQA can help create 3D designs and structures for improved design processes.

In robotics, VQA can enable robots to respond to queries about their surroundings, leading to better interactions between robots and humans. In the educational domain, VQA can help students to ask questions about instructional content, tables, diagrams, and visual materials. It can generate descriptions and explanations of historical sites, artworks, and cultural artifacts, enhancing tourist's experiences and safeguarding cultural heritage.

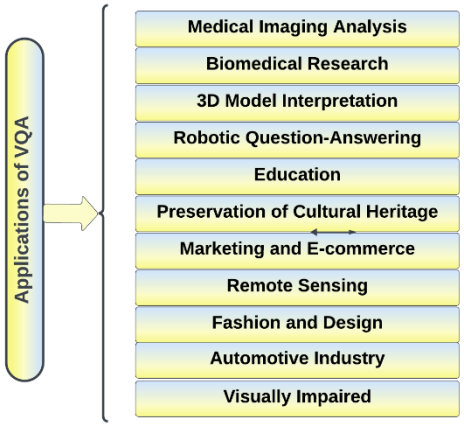

Fig 4. Versatile applications of VQA

VQA can also improve user experiences on e-commerce platforms by providing information about products based on images. It helps in identifying remote sensed images related to object detection and identification tasks. Through image-based inquiries, it can assist fashion enthusiasts in identifying clothing items, suggesting designs, and providing fashion advice.

VQA can also enhance driving assistance systems by providing users with information about traffic conditions, obstacles, and traffic signs. It can empower individuals who are blind or visually impaired by allowing them to interact and better understand their surroundings through spoken questions related to visual information, such as images and videos.

2 Approaches and Techniques in Visual Question Answering

2.1 Transformer based approaches for Visual Question Answering

Transformer-based approaches have revolutionized VQA by introducing various frameworks for tackling complex visual questions. These approaches include an attention mechanism, multi-modal fusion, pre-trained models, and attention

visualization. Transformers are effective in VQA by focusing on relevant regions of the image and words, capturing interactions between the two modalities, and enhancing performance on VQA-specific datasets. There are few drawbacks in transformer-based approaches like handling ambiguous questions, addressing biases in training data and handling large-scale visual information. In Figure 5, a transformer-based architecture is illustrated.

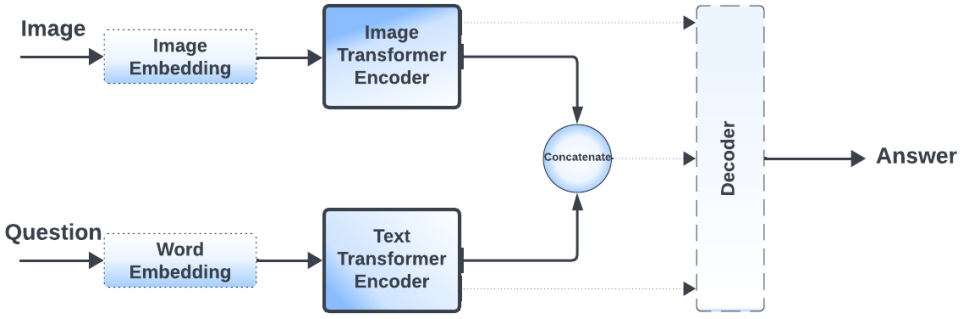

Fig 5. General architecture of a transformer-based approach

VisualBERT [2], is the earliest transformer-based approach. It first encodes the image using a CNN and then feeds the encoded image features into BERT and achieves state-of-the-art results on the Visual Genome dataset. [3]proposes a transformer-based approach, MQAT, for medical visual question answering. It combines self-supervised pre-training with finetuning and achieves good performance on medical VQA datasets. This helps the model to learn general features that can be used for the VQA task.

Transformer Module Network (TMN) [4] introduces modularity to Transformers by using Neural Module Network (NMN) concepts. The work of[5] investigates the efficacy of co-attention transformer layers in VQA. Co-attention transformer layers allow the model to attend to both image and question simultaneously. This can help to improve the accuracy of the answer, as the model can learn the interactions between the two modalities. BST framework of[6] integrates transformer based VQA models, generating slimmed sub models for efficient deployment. Hypergraph Transformer of[7] encodes question-knowledge semantics, captures associations, and effectively infers answers in multi-hop reasoning. This would be helpful for VQA, as the task often requires a model to reason about the relationships between different entities in image and question.

The model of[8] employs vision-text transformers and a residual MLP-based VisualBERT encoder to improve performance in classification-based answering. A multi-modal transformer-based architecture, VB-Fusion, is proposed by[9] to learn joint representations by combining modality-specific features with multi-modal transformer layers. The two encoded representations are then combined using a novel modal fusion method.

The approach of[10] utilizes the Contrastive Language Image Pretraining (CLIP) network for embedding image patches and question words. The CLIP network can be used to learn embeddings for image patches and question words. The proposed Re-Conv Attention in the transformer module (CAT) [11] combines self-attention and depth wise separable convolution to capture both global and local relationships. Self-attention is a mechanism which enables the model to use different parts of the input sequence. Depth wise separable convolution is a type of convolution that makes the model to learn local relationships. Semantic Aligned Multi-modal Transformer [12] is utilized to enhance the alignment mechanism by incorporating image scene graph structures as a bridge between vision and language. [13] proposed MMFT-BERT. It employs a novel transformer-based fusion method to combine different modalities effectively. MMFT-BERT can be utilized to enhance the accuracy of different VQA models by learning the interactions between the image and question through both transformers and BERT.

2.2 Attention based approaches for Visual Question Answering

Attention mechanism is a technique that enables neural networks to focus on specific parts of input, such as images and questions, to increase the accuracy of answers. In VQA, the attention mechanism encodes images and questions into vector representations and later computes an attention weight for each region by combining the image and question representations. The combined representations are then weighted. Benefits of using attention mechanism for VQA include enhancing the accuracy of answer, learning long-range dependencies between the image and question. The attention mechanism allows the model to better align visual features with corresponding textual elements, enhancing its ability to capture relevant information and context. An overview of attention mechanism applied for VQA task is shown in Figure 6.

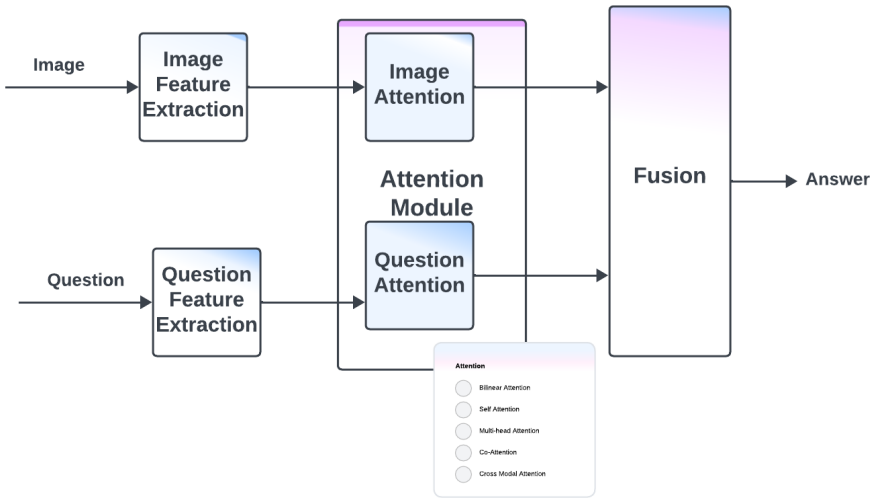

Fig 6. An overview of attention mechanisms employed in the context of VQA

CLIP-guided VideoQA introduces[14] a visual-text attention mechanism that uses CLIP to guide cross-modal learning. DAQC-VQA[15] introduces a dual attention mechanism for enriched cross-domain representation and a question categorizer subsystem to reduce answer search space. The AMAM[16] model addresses the issue by incorporating an attention-based approach to align text-based and image-based attention in medical VQA.

[5] investigates the efficacy of co-attention transformer layers in VQA and evaluates the impact of critical components on visual attention. The Cubic Visual Attention (CVA)[17] model applies spatial attention on object regions instead of the pixels. This can improvise the accuracy of the answer by focusing on the most relevant parts of the image. Word and Sentence Dual-Attention Network (WSDAN) proposed by[18] uses a dual-attention learning network with word and sentence embedding for Medical VQA. This allows the model to learn the interactions between words and sentences in the question, which can improve the accuracy of the provided answer. A dual-attention learning (DAL) module consists of self-attention and guided attention models intensive intramodal and intermodal interactions. Here self-attention and guided attention are models that learn interactions between different parts of image and the question, enhancing the accuracy of the answer by learning the rich interactions between visual and language streams. The model proposed by[19] introduces a multi-hop attention alignment method that enriches surrounding information when using self-attention.

2.3 Knowledge and Graph based Visual Question Answering

Researchers have developed knowledge-based approaches to enhance VQA models by embedding external knowledge sources, such as textual corpora or structured databases, into the VQA framework. This approach allows models to tackle complex questions that demand a broader perspective and understanding. Graph-based approaches transform the VQA problem into a graph, where nodes represent entities and edges represent relationships between them. These graph structures capture semantic relationships, contextual dependencies, and interconnections between visual and textual components, providing better reasoning and improving comprehension of complex questions involving multiple entities. Knowledge-based approaches are used to answer questions that require common knowledge or go beyond what is visible in the image. Graph-based approaches represent the relationships between entities of image and question, allowing for better comprehension of complex questions involving multiple entities. Figure 7 represent a framework which includes a knowledge and graph-based approach.

The work addressed in[20] constructs the concepts graph by inferring images and question entity concepts, using language model to calculate correlation scores, and form a visual graph from visual and spatial features of filtered image entities. The proposed baseline method[21] constructs spatial, semantic, and implicit relationship graphs on image regions, questions, and semantic labels, enabling answer and graph reasoning paths for different questions. Multi-modal Semantic Graph Knowledge Reasoning Model (MSG-KRM)[22] approaches to unify the representation of heterogeneous data and diverse types of knowledge. The model performs reasoning and deep fusion of image-text information and external knowledge sources using

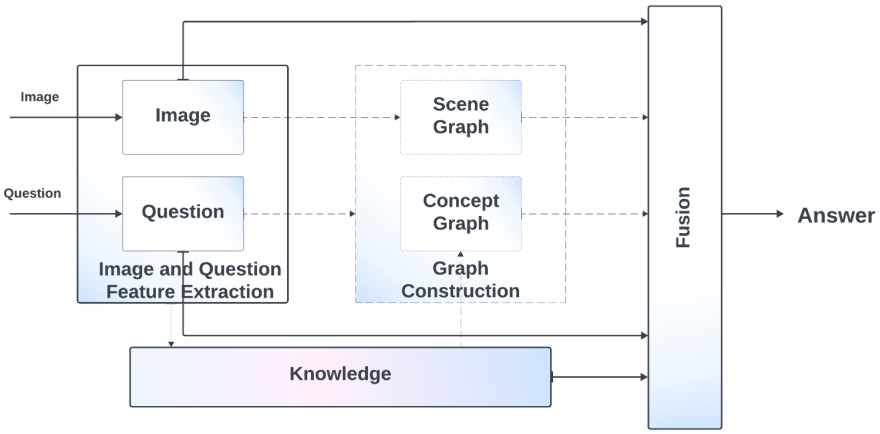

Fig 7. Architecture for knowledge and graph based VQA

a semantic graph knowledge reasoning model.

The work proposed in [23] models semantic, spatial relations in images using three graph attention networks. The DMMGR model [24] performs explicit and implicit reasoning over a knowledge memory module and a spatial-aware image graph, respectively. DM-GN [25] introduces a dual message-passing mechanism to properly encode multi-scale scene graph information. Hierarchical Graph Neural Module Network (HGNMN) [26] introduces a hierarchical approach that reasons over multi-layer graphs with neural modules. It encodes the image using multi-layer graphs from visual, semantic, and commonsense views, enabling multi-step reasoning within and between different graphs.

VQA-GNN [27] proposes a VQA method that unifies image-level information and conceptual knowledge through a multimodal semantic graph. The text-only model [28] use the implicit knowledge of pretrained language models through captioning of images. Multimodal Interpretable VQA [29] incorporates information from outside knowledge and multiple image captions to improve the rationality and diversity of generated explanations. The method in [30] treats sentences as pseudo-questions and their contexts as pseudo-relevant passages, extending to multimodal documents by considering images near texts. Prophet [31] is a framework that enhances the capacity of GPT-3 for knowledge-based VQA by prompting it with answer heuristics.

2.4 Reasoning based approaches for Visual Question Answering

Reasoning-based VQA is a significant advancement in Artificial Intelligence and Computer Vision. It focuses on understanding the relationships and contexts in visual and textual data, using advanced reasoning mechanisms to simulate human cognitive processes. These models can answer complex questions that demand logical inference, spatial comprehension, and contextual reasoning. Reasoning-based VQA systems provide more accurate and contextually relevant answers to questions about images. These models exhibit more power than traditional models but require more training data. They are more complex than traditional approaches. A framework for reasoning is shown in Figure 8.

CMQR [32] introduces an event-level visual question reasoning framework that explicitly discovers temporal causal structures and mitigates visual spurious correlations. Visual Commonsense Reasoning (VCR) is a challenging extension of VQA that focuses on high-level visual comprehension. To address this issue, [33] proposes framework which couples question answering and rationale inference processes by introducing a novel branch as a bridge to conduct processes connecting. [26] extends visual reasoning from one graph to more, tracing the reasoning process according to module weights and graph attentions.

The approach in [34] introduces 3D geometric information into the spatial reasoning process for TextVQA. [35] proposes a model which incorporates spatial relationship reasoning along with visual object semantic reasoning using a sparse attention encoder and a graph neural network attention mechanism. [36] proposed model incorporates multi-modal relation reasoning in multi-scales. It makes use of a regional attention scheme to focus more on informative and question-related regions, enabling better answering.

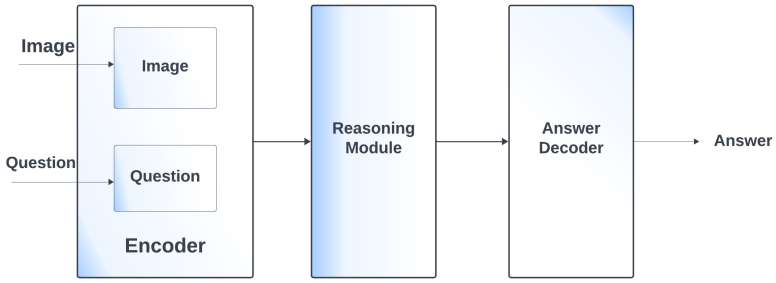

Fig 8. A framework of Reasoning based VQA

2.5 Capsule Networks and Vision Transformers for Advanced Visual Question Answering

2.5.1 Capsule Networks for Visual Question Answering

Capsule networks are neural networks that represent features as capsules instead of neurons, capturing more information about the features they represent, including their pose and spatial relationships. They are well-suited for tasks such as image classification and object detection. The architecture of a capsule network consists of two main parts: an encoder and a decoder. The encoder extracts feature from the input image and represents them as capsules, which are groups of neurons whose activity vector represents parameters of a specific entity. The encoder typically consists of a CNN followed by a capsule layer. The CNN extracts low-level features from the input image, such as edges and corners, and groups these features into capsules, representing each capsule as a vector. The decoder reconstructs the input image from the capsule representations generated by the encoder, typically consisting of a series of fully connected layers.

Capsule Networks (CapsNet) are deep learning architectures that overcome limitations in traditional neural networks, improving performance in multimodal tasks like VQA by efficiently capturing spatial relationships and interactions between modalities such as images and text, leading to more accurate results.

Semantic-aware modular capsule routing framework (SUPER)[37] proposes a framework that uses capsule networks for VQA. It introduces specialized modules and dynamic routers in each layer to better capture instance-specific vision-semantic characteristics and improve representations for prediction. Linguistically driven Graph Capsule Network[38] develops a hierarchical compositional reasoning model for VQA, where the compositional process is guided by a linguistic parse tree. It incorporates capsule networks to perform a compositional reasoning process inside a CNN for VQA. Dual capsule attention mask network with mutual learning[39] introduces a dual capsule attention mask network for VQA, which uses capsule-based attention mechanisms for processing both coarse-grained and fine-grained features. The proposed approach achieves state-of-the-art performance on the VQA-v2 dataset.

Dynamic Capsule Attention (CapsAtt)[40] proposes CapsAtt, an alternative to static multi-layer attention architectures for VQA. CapsAtt treats visual features as capsules and employs dynamic routing to update attention weights, making the model more efficient while achieving competitive results. Capsules are leveraged to transform visual tokens into capsule representations in the visual encoder, and language self-attention layers are used as a text-guided selection module to mask capsules in [41]. The work focuses on weakly supervised grounding in transformers for VQA. It introduces capsules for visual token representation and uses text-guided selection to mask capsules for improved grounding.

2.5.2 Vision Transformers for Visual Question Answering

A Vision Transformer is a neural network that uses self-attention mechanisms to process images. It is like a standard Transformer but with key changes to make it better suited for image processing. The input to a ViT is a two-dimensional image divided into fixed-size patches, flattened into vectors, and embedded into a lower-dimensional space. The embedded patches are passed to a stack of Transformer encoder blocks, each consisting of a multi-head self-attention layer and a feed-forward layer. The multi-head self-attention layer allows the model to learn long-range dependencies between patches, which is crucial for image classification. The feed-forward layer applies a non-linear transformation to the output of the self-attention layer, helping the model learn complex relationships between patches. The output of the final transformer encoder block is a sequence of vectors representing the features of the image.

The work of[42] introduces a Gated Vision-Language Embedding (GVLE) to fuse heterogeneous modalities (text and image) efficiently and uses a Language Vision Transformer (LViT) with a detection head for localization. By using the ViT architecture,

this approach enables the model to perform vision-based tasks with high performance, achieving notable success in many vision tasks because of its ability to capture the long-range dependencies and complex visual patterns through self-attention. The[43] introduces the Language-Vision GPT (LV-GPT) model, which incorporates vision input for VQA in surgery. It utilizes ViT, ResNet18 and Swin for feature extraction.

The work of[44] proposes the Cross-Modal Question Reasoning (CMQR) framework for event-level visual question reasoning. It focuses on discovering temporal causal structures and aligning visual and linguistic content for robust question reasoning. The proposed work of[45] introduces the MedVidCL and MedVidQA datasets, aiming to understand medical videos and provide visual answers to medical language questions. The datasets support cross-modal understanding and have potential applications in medical domains. Figure 9 shows an illustration of utilizing Vision Transformer for VQA.

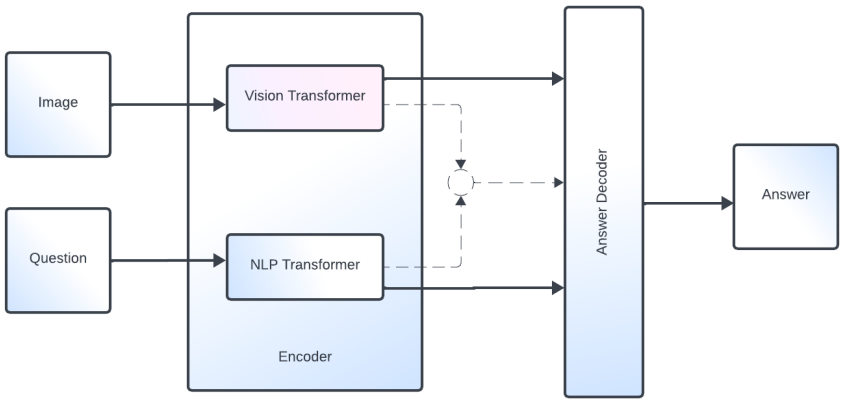

Fig 9. Overview of the model architecture of Vision Transformer for VQA task

The work of[46] discusses a transformer based VQA system for medical images, utilizing the ViT and textual encoder transformer. The model exhibits good results on two VQA datasets comprising radiology images. The Layout-Aware Transformer (LaTr)[47] introduces for Scene Text VQA. It focuses on reasoning over different modalities and proposes a pre-training scheme that incorporates layout information. They have used a vision transformer to remove external object detectors. The[48] proposes the Plug-and-Play VQA framework, a modular approach for zero-shot VQA. It leverages a pre-trained language model and generates question-guided informative image captions to achieve state-of-the-art results on zero-shot VQA tasks. The model utilizes a vision transformer to extract the features from the given image.

The[49] introduces the Change Detection-based Visual Question Answering (CDVQA) task on multitemporal aerial images. It creates a CDVQA dataset and devises a baseline CDVQA framework, exploring different backbones and fusion strategies. The[50] presents VT-Transformer, an approach for Answerability on VQA which achieves competitive results on the VizWiz 2020 dataset. The[51] describes the AliceMind-MMU system, which achieves human-level performance on VQA by pre-training with comprehensive visual and textual feature representation and using specialized expert modules for different types of visual questions. The[52] introduces the M2I2 self-supervised method for pretraining and finetuning on medical image VQA, achieving state-of-the-art performance on medical VQA datasets.

2.5.3 Advantages of CapsNet and Vision Transformer in Visual Question Answering

CapsNet can capture spatial relationships and interactions between modalities like images and text, leading to more accurate results. Capsules, which represent hierarchical features, are suitable for complex reasoning tasks in VQA. Dynamic routing can adaptively update attention weights, improving efficiency and competitive performance. Capsule networks have been shown to achieve better results on VQA tasks than traditional neural networks.

ViTs can handle various vision tasks and its self-attention mechanism allows for the capture of long-range dependencies in both images and text, which makes it difficult to understand complex relationships between visual and textual information. ViTs architecture is flexible and can be adapted to various tasks, by fine-tuning pre-trained models. It can leverage pre-trained representations on large-scale image datasets like ImageNet, improving performance on downstream tasks. ViT can handle images of varying resolutions by dividing them into patches, making it scalable to different image sizes without significant modifications. Parallel processing of image patches in ViT can lead to faster inference times.

2.5.4 Analyzing Limitations of CapsNet and Vision Transformer in Visual Question Answering

Capsule Network based VQA has few drawbacks. They are complex, less computationally efficient than traditional neural networks, and have challenges in fine-tuning and adapting to specific VQA datasets due to their less widespread adoption compared to ViT.

ViT has several drawbacks,like limited understanding of local information, high computational cost, and reliance on pre-trained models. These models focus on global relationships in images. They require large training data which makes fine-tuning for specific tasks challenging. Additionally, they are less robust to noise in images than traditional CNNs and require more training data than traditional CNNs to achieve optimal performance.

2.5.5 Enhancing Vision Transformers with Capsule Network for Improved feature extraction

Capsule Networks (CapsNet) and Vision Transformers (ViT) can be combined to address some of the limitations of ViT while leveraging their strengths. This combination can help overcome specific ViT limitations by capturing local information effectively, enhancing data efficiency, enhancing domain adaptability, improving computational efficiency, and enhancing robustness to occlusions. CapsNet's routing mechanisms effectively preserve fine-grained details in images, making it suitable for extracting detailed local features while still leveraging global attention for long-range dependencies. By using CapsNet for initial feature extraction, ViT can require fewer data samples during fine-tuning, making it more adaptable to scenarios with limited training data.

CapsNet's ability to be adapted and fine-tuned for specific domains or tasks enhances its domain adaptability. For example, if a VQA task involves medical images, CapsNet can be pre-trained on medical data to provide specialized features, enhancing ViT's domain adaptability. Incorporating CapsNets robustness to occlusions can improve ViT's ability to handle obscured or partially visible objects in VQA tasks. Capsule networks are robust to noise present in the image than ViTs, making them less sensitive to noise.

Reducing the need for training data is another benefit of combining CapsNet with ViT for image feature extraction. Capsule networks can learn complex representations of the image with less data, reducing the training data required for optimal performance. Combining CapsNet with ViT for image feature extraction can offer several benefits, like improved local information capture, enhanced data efficiency, enhanced robustness to occlusions, reduced training data requirements, and improved long-range dependency capture.

2.6 Visual Feature Extraction Techniques

Feature extractors are crucial in VQA systems to transform raw pixel data into high-level visual representations. The choice of feature extractors impacts the performance of a VQA model, with discriminative features improving accuracy. These feature extractors represent objects, scenes, and other visual elements in images. Table 1 shows various image feature extractors used to extract visual contents from images by different VQA models.

2.6.1 ResNet[53]

ResNet (Residual Network) is a deep convolutional neural network which uses residual connections to enhance the training stability and its performance. It is a popular image feature extractor for VQA and has shown appropriate results on various VQA benchmarks. It has different variants like ResNet-152, ResNet-101, ResNet-10, and ResNet-50. ResNet enhances training stability and performance by using residual connections, which skip over some network layers, to prevent the vanishing gradient problem, a common issue in deep learning. This problem occurs when the loss function's gradients become small, making learning difficult for the network.

2.6.2 RCNN[120]

RCNN (Region-based Convolutional Neural Network) is a region-based CNN that proposes regions of interest (ROIs) in an image. It classifies each ROI as an object or background, despite being a slow algorithm with high accuracy. ROIs are likely to contain objects.

2.6.3 Faster RCNN[54]

It is an improved version of RCNN that uses a region proposal network to generate ROIs to make it faster and more accurate than RCNN. It enhances accuracy by utilizing a CNN that predicts bounding boxes for objects in an image.

Table 1. Image Feature Extractors used in different VQA models.

Visual Feature Extractors	VQA Model
Faster-RCNN [54]	CPDR [55], MulFA/UFSCAN [56], Bilinear Graph [57], AttReg [58], AMAM [16], Scene-text using PHOC [59], MGRF [60], Bottom-Up and Top-Down [61], DCAMN [39], Skill Concept [62], PGM [63], SR-OCE [64], RAMEN [65], CSST [66], Coarse-to-Fine [67], GMA [68], BLOCK [69], CapsAtt [32,40], Re-attention [70], CRN [71], CAT [11], shortcut [72], DAQC [15], MGFAN [73], MMMH [19], MSG [74], Fair-VQA [75], Attention map [5], SAVQA [76], MGAVQA [77], MuKEA [78], ACVRM [79], QD-GFN [23], Swap-Mix [80], CVA [17], HGNMN [26], SUPER [37], Uncertainty based [81], CLG [82], WSQG [83], VLR [84], LXMERT [85], SceneGATE [86], MMGLM [87]
Resnet [53]	MMRR [36], TPT [88]VQA [49], multi-image [89], NEWSKVQA [90], Answer-Me [91], SCAN [92], V-MODE [93]
ResNet-10	MQAT [3]
ResNet50	CAN [94], DET [95]
ResNet-101	LG-Capsule [38], LiVLR [96], DulVGR [97], GVQA [41]
Resnet-152	GPC [98], MTAN [99], Zero Shot [100], MFB [101], RSVQA [102], ReasonNet [103], MFB and MFH [101], FTDD [104], STA [105], WSDAN [18]
Faster R-CNN with ResNet-50	Seeing is Knowing [106], MULAN [107]
Faster R-CNN with ResNet-101	GAT [108], ATH [109], DMMGR [24], MCLN [110], MCAN [111], F-SWAP [112], SRRN [35], TVQA [113]
Faster R-CNN with Resnet-152	RA-MAP [114], MASN [115], Anamoly based [114], Vocab based [116], DA-Net [117]
ResNet CNN within Faster R-CNN	MuVAM [118]
FasterR-CNN with ResNext-152	CBM [119]
RCNN [120]	Multi-image [89]
VGGNet [121]	VQA-AID [122]
EfficientNetV2 [123]	RealFormer [124]
YOLO [125]	Scene Text VQA [126]
CLIPViT-B	CCVQA [14]
Resnet NFNet [127]	Flamingo [128]
ViT [129]	VLMmed [46], ConvS2S+ViT [130], BMT [10], M2I2 [52]
XCLIP with ViT-L/14	CMQR [32]
RsNet18, Swin, ViT	LV-GPT [43]
GLIP [131]	REVIVE [132]
CLIP [133]	KVQAE [30]

2.6.4 VGGNet[121]

VGGNet (Visual Geometry Group Network) is a CNN with a small number of layers, achieving good performance in image classification tasks. It is basically known for its simplicity and generalizability to new datasets. It uses a 3x3 convolution kernel for efficient convolutional layers.

2.6.5 YOLO[125]

YOLO (You Only Look Once) is a real-time object detection system which utilizes a single CNN to predict the class labels and bounding boxes for objects in an image. YOLO works by dividing an image into a grid and later predicts the class labels and bounding boxes for each grid cell. YOLO performs all object detection tasks, including region proposal, classification, and bounding box regression.

2.6.6 EfficientNet[123]

EfficientNet is a family of CNNs that are efficient and scalable, achieving advanced results in object detection and image classification benchmarks. EfficientNets are based on the ResNet architecture. Techniques used to improve efficiency include

compound scaling factor, bottleneck design, and dynamic convolution kernel size, which increase network size while maintaining accuracy and adapt to input image size.

2.6.7 CLIP[133]

CLIP (Contrastive Language-Image Pre-training) is a text-image encoder that uses CNNs as a vision encoder and a transformer as a text encoder. It can match images and text or generate text descriptions of images. CNN extracts feature from images, while transformer extracts feature from text. The features from both are compared to determine their similarity.

2.6.8 GLIP[131]

GLIP is a text-image encoder that combines phrase grounding and object detection for pre-training, enhancing the performance of VQA models. It uses CNN to extract the necessary features from images and a transformer to extract the features from text, predicting object location and answering image-related questions using these features.

2.6.9 ViT[129]

In 2020, there has been a growing interest in using ViTs for Computer Vision (CV)tasks. ViTs can process visual data more globally than CNNs and has shown its effectiveness for various tasks, including VQA and Image Captioning. ViTs can overcome limitations of convolutional neural networks in image feature extraction, such as locality, data efficiency, and interpretability. CNNs are limited by their locality, limiting their ability to learn relationships between neighboring pixels. They also require large amounts of labeled data, making them costly and time-consuming to train. ViTs can learn relationships between any two pixels in an image, making them better suited for tasks requiring long-range dependencies. They are also more data-efficient and easier to interpret.

2.6.10 CapsNet[134]

CapsNets are a type of neural network that was introduced in 2017. CapsNets can capture long-range dependencies in images, and they are effective for tasks which require reasoning. CapsNets preserve positional information by using capsules, vector representations of objects that encode both the identity and pose of the object. They exhibit greater invariance to changes in the image using dynamic routing, which allows capsules to focus on different image regions depending on the task. CapsNets have enhanced interpretability due to their use of capsules, which are more meaningful and easier to understand than the features extracted by CNNs. D

2.7 Drawbacks of using CNN approaches for image feature extraction in Visual Question Answering

CNN-based approaches have achieved state-of-the-art results on many VQA benchmarks, but still, they have some limitations. The fixed spatial resolution is the most common issue with CNN-based architectures, as it may not capture the fine-grained details necessary for accurate question answering. Fixed-size receptive fields may not capture the entire context of an image, especially for questions that require global understanding or reasoning over distant image regions. CNNs are not good at reasoning about spatial relationships between different objects in an image. For example, CNN may be able to identify that there is a dog and a cat in the given image, but it may not be able to tell you which animal is in front of the other. This is because CNNs typically operate on local image patches, and they do not explicitly encode spatial information.

The semantic gap between visual and textual data is another challenge for CNNs, as VQA requires understanding the semantics of both images and questions. Traditional CNN-based approaches focus solely on image features, which makes it difficult to integrate textual information effectively. Training data bias can also lead to biased or unfair responses in VQA systems. Interpretable features can be challenging to provide explanations for answers generated by VQA systems.

To address the mentioned limitations, alternative approaches can be utilized for VQA which include attention mechanisms, transformers, and combining vision and language models like ViTs and BERT-based models, to enhance context modeling, multimodal integration, and generalization.

2.8 Datasets for Multimodal Question Answering

A model requires datasets for training and for the purpose of evaluation. The effectiveness of models does rely on the quality, type of datasets being considered. It is necessary to consider the size of dataset, diversity of questions, and difficulty of the questions. A larger dataset provides more training data, while a diverse dataset helps the model learn to answer various questions. Table 2 shows the visual question answering datasets, Table 3 shows the general domain specific question answering datasets and Table 4

represents different text-based question answering datasets.

Table 2. Visual Question Answering datasets

Domain	Dataset
Image	VQA, DAQUAR, VQA 1.0, VQA 2.0, RGQA, Polar-VQA, Visual Madlibs, Visual7W, VizWiz, VQA-HAT, VQA-Rephrasings, VQA-P2, IV-VQA, VLQA, CV-VQA, Fashion-VQA, Mini-FashionVQA, VQA_CE, MQA, ST-VQA, VQA-MHUG, HowMany-QA, TextVQA, CLEVR, CLEVR-CoGenT, COCO-QA, VQ2A-COCO, VQ2A-CC3M, GBQA, YNBQD, AQUA, GQA, GQA-OOD, CDVQA, GQA-SGL, CLOSURE, RecipeQA, PointQa-Local, VQA-GENDER, CRIC, HurMic-VQA, Visual Genome,VQA-CP V1, VQA-CP V2, VQA-Sports, VQA-Food, FM-IQA, Diagrams, FigureQA, ICQA, SHAPES, TDIUC, QPR, QE
Medical	VQA-MED-2018, VQA-RAD, VQA-RADPH, VQA-MED-2019, PubMedQA, RadVisDial, PathVQA, KGQA, MedQuAD, MedQA, Head-QA, ClinicalKBQA, Med-VQA, EndoVis-18-VQA, Cholec80-VQA,PSI-AVA-VQA, DrugEHRQA, MeQSum, HealthCareMagic, VQA-Med-2020, SLAKE, SLAKE-EN, , CliCR, BioASQ, MedVidCL, MedVidQA, VQA-Med-2021, CovidQA, IMAGE-CLEF 2019, BioTABQA
Video	Youtube-QA, TGIF-QA, TVQA, MovieQA, YouTube2TextQA, MSVD-QA,MSRVTT-QA, MedVidQA, VideoQA, CLEVRER, FILL IN THE BLANKS, Pororo-QA, TVQA+, TVQA, ActivityNet-QA, LifeQA, DramaQA, Social-IQ, MarioQA, EgoVQA, PlotGraphs, Tutorial-VQA, KnowIT-VQA, DME, NExT-QA, SVQA, CRIPP-VQA, SUTD-TrafficQA, How2QA, iVQA, WebVIdVQA3M, EMQA
Zero Shot Learning	ZSL-VQA, ZS-F-VQA
3D	Hypersim-VQA, ThreeDWorld-VQA, CLEVR3D, FE-3DGQA, ScanQA, TransVQA
Remote Sensing	RSVQA, RSIVQA, HRVQA

Table 3. General Domain specific question answering datasets

Domain	Dataset
Cross Lingual	PAXQA, GEN-TYDIQA
Aesthetic	AesVQA
Acoustic	CLEAR2
Chart	OpenCQA, ChartQA, FigureQA, DVQA,PlotQA
Map	MapQa
Numerical	ASDiv-a, DROP, TATQA, FinQA
Object counting	Tally-QA
Multilingual	xGQA, TyDi QA, XOR-TyDi QA, MCVQA, MuCo-VQA, ChAII, EVJVQA
Document	DocVQA, AQuAMuSe, VQAonBD 2023, FUNSD-QA, InfographicsVQA, PubVQA, SlideVQA, AWS Documentation

2.9 Evaluation Metrics

Evaluation metrics measure VQA system performance, aiding in comparing systems, providing feedback for researchers, and identifying strengths and weaknesses. They help researchers improve their systems and identify strengths and weaknesses. There are many evaluation metrics used for VQA. The most used metrics for VQA are:

2.9.1 Accuracy

This is the most basic metric, and it simply measures the percentage of questions that the model answers correctly. For example, if a model is given with 100 questions and the model answers 90 of them correctly, then its accuracy is 90%.

2.9.2 BLEU[135]

BLEU (Bilingual Evaluation Understudy) is a metric used to measure the similarity between two text strings, particularly in VQA models. It measures the overlap between the generated answer and a reference answer, considering word order. A higher BLEU score indicates greater similarity between the two strings.

Table 4. Text-based question answering datasets

Domain	Dataset
Text	SQuAD, WikiQA, DCQA, Natural Questions, CNN and Daily Mill, MCTest, WikiQA, TweetQA, CREPE, ReasonChainQA, INFOSEEK, ControversialQA, ASQASCDE, TAT-QA, NarrativeQA, Simple-SQUAD, MS Marco, QAMPARI, QASPER, QUASAR, NewsQA, BoolQ, SearchQA, TIFA v1.0, NarrativeQA, TellMeWhy, TriviaQA, ELI5, GooAQ, ANSQ, FinQA, SQuAD 2.0, SQUAD 1.1, SQuAD Open, TREC, AQAD, Simple Questions, IQA, QASC, CSQA, CS1QA, CodeQA, YH, QALD, IQUAD v1, Web Questions, WebQuestionsSP, WebQuestionsSP-tiny, GraphQuestions, WPQA, InsuranceQA, QAit, Quasar-T, SmartTV/Remote, S10QA, ArchivalQA, OpenBookQA, HotpotQA
MCQ	MCTEST, CHILDREN'S BOOK TEST, BOOK TEST, RACE, ARC
Knowledge	OKVQA, KRVQR, FVQA, ConceptNet, FVQA 2.0, KVQA, Free917, QALD, K-EQA, S3VQA, KQA Pro, CFQ, TIMEQUESTIONS, FORECASTTKGQUESTIONS, NEWSKVQA, A-OKVQA, GraphQuestions, LC-QUAD 1.0, LC-QUAD 2.0, GrailQA, TempQA-WD, OKVQAS3, DBLP-QUAD
Reasoning	VCR, LogiQA, PIQA, Mathematics, Commonsense QA, CommonsenseQA 2.0, SocialIQA, Abductive NLI, QASC, SWAG, Physical IQA, CosmosQA, CICERO v1-v2, CODAH, COPA, R-VQA, Winogrande, UnifiedQAv2
Multi-hop	WebQA, ComplexWebQuestions, HotpotQA, ConditionalQA, WikiHop, MetaQA, MuMuQA
Financial	FinQA
Conversational	CONVMIX, Spoken-CoQA, QuAC, CoQA, CONVREF, SQA, CONVQUESTIONS, CSQA, Topi-OCQA, Verbal-ConvQuestions

2.9.3 METEOR[136]

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is a metric that measures the similarity between two text strings, similar to BLEU but more robust to small text changes. It considers word order, synonyms, and related words, and a higher METEOR score indicates a higher degree of similarity between the two strings.

2.9.4 CIDEr[137]

CIDEr (Consensus-based Image Description Evaluation) is a metric that measures the n-gram overlap between two text strings, focusing on word ordering. It is calculated by identifying n-grams in both generated and reference answers, and the score is determined by calculating the sum of their frequencies. A higher CIDEr score indicates greater overlap between the two strings.

2.9.5 SPICE[138]

SPICE (Semantic Propositional Image Caption Evaluation) is a metric that assesses the semantic similarity between two text strings, ensuring it is robust to changes in text meaning. It calculates the score by extracting propositions from the generated answer and reference answer, and a higher score indicates a higher degree of semantic similarity between the two strings.

2.9.6 ROUGE[139]

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to assess the similarity between two text strings. Common metrics include ROUGE-L, ROUGE-N, and ROUGE-W, which measure fluency, overlap, and informativeness of the generated answer. ROUGE-L measures the longest common subsequence, ROUGE-N measures the number of n-grams in both answers, and ROUGE-W measures word overlap.

2.9.7 BERT Score

The BERT Score is a metric used to assess the similarity between two text strings, based on the BERT model, trained on a large dataset of text and code. It measures the similarity between word embeddings in the generated answer and the reference answer, with a higher score indicating greater similarity.

3 Discussion

3.1 Visual Feature Extraction

VQA is a challenging task requiring understanding textual questions and visual content. Traditional CNNs are widely used for feature extraction. Recent deep learning advancements, like CapsNet and Vision Transformers, have the potential to make

significant changes in visual feature extraction. From our literature review, we can say that although CapsNet can efficiently extract image features, there are not many works that use it as a feature extractor. Vision transformers (ViTs) are only being utilized in a few works. Therefore, we suggest an alternative of using these models as feature extractors.

3.1.1 Capsule Networks for Visual Feature Extraction

CapsNet is a powerful visual feature extraction tool that captures hierarchical features and spatial relationships through dynamic routing. It can be used for feature extraction in VQA. CapsNet is more robust to noise and deformations than pooling layers, which can lose information about spatial relationships. It can learn abstract features using a routing algorithm to combine lower-level capsules into higher-level ones. Capsule networks use a mechanism called ”routing by agreement” to learn relationships between different parts of an object, filtering out noise and identifying the most important features. To use capsule networks for VQA, a hybrid architecture combining capsule networks with other neural networks or transformers can be used. Another option is a single capsule network that can process both images and questions. Capsule networks can learn relationships between visual features and semantic meaning of questions, are robust to noise and occlusion, handle viewpoint changes, and answer complex scenes.

3.1.2 Vision Transformers for Visual Feature Extraction

ViTs are neural networks that are being explored for VQA tasks. They differ from traditional CNNs wherein they do not utilize convolutions to extract features from images but instead use self-attention. ViTs are effective in VQA tasks as they can learn long-range dependencies between pixels, which are necessary for understanding the relationships between objects in an image. They are capable of learning global features of an image, which could be helpful for answering questions that require a holistic understanding of the image.

The main benefits of using ViTs for VQA tasks are, they are significant and have improved accuracy and efficiency. There are many challenges to overcome before ViTs can be widely used for VQA tasks. Then challenges to be considered are, large amount of data is required to train, the computational cost of training and deployment. Research directions for improving the future scope of using ViTs for feature extraction in VQA tasks include developing more efficient and scalable training methods, improving the interpretability of ViTs, and developing ViTs for other tasks related to visual understanding, such as image captioning.

3.1.3 Integrating CapsNet and ViT for Visual Question Answering specific tasks

Capsule networks and transformers are neural networks which can be used for feature identification. Capsule networks represent objects and their parts as vector-based entities, while transformers focus on relevant parts of the input sequence. High-dimensional coincidence filtering and agreement can combine the strengths of both networks, enabling feature identification in multiple capsules. Research on combining capsule networks and transformers is still in its early stages, yet it holds a promise to build powerful and accurate models. Capsule architectures inspired by ViTs can be developed using various fusion strategies. These include parallel fusion, sequential fusion, attention fusion, and multi-modal stacking. These strategies can extract a wide range of features, but careful selection is needed to avoid duplication.

Apart from fusion techniques, we can also integrate ViT and Capsnet by replacing self-attention layers with capsule attention layers, adding a capsule layer after ViT, or a hybrid approach, which balances local and global relationship learning but requires more computational resources. Integrating CapsNets into ViTs for image feature extraction in VQA requires careful consideration of capsule parameters, routing algorithm, training data, feature fusion at multiple levels, joint training, progressive feature integration, and using capsules as attention mechanisms. The combined strengths of CapsNet and ViT can lead to better VQA performance especially in visual feature representation.

CapsNets are better at preserving positional information than ViTs. This is because CapsNets use capsules, which are vector representations of objects that encode both the identity and the orientation of the object. ViTs, on the other hand, use attention mechanisms, which is less effective at preserving positional information.

4 Conclusion

This review article explores the potential of multimodal deep learning for VQA, with a specific focus on image feature extraction. It emphasizes the importance of comprehensive image representation in VQA and the limitations of existing models that primarily focus on textual content of questions. Recent advances in image feature extraction techniques, such as Capsule Networks (CapsNets) and Vision Transformers (ViTs), are reviewed as promising alternatives to traditional Convolutional Neural Networks (CNNs) for VQA.

Key directions for future research in VQA include developing more effective image feature extraction techniques that capture fine-grained details and long-range dependencies in images. It emphasizes the need to design multimodal fusion mechanisms that leverage the complementary strengths of visual and textual modalities to generate more precise and context-aware answers. Addressing challenges related to multimodal bias and data scarcity, developing VQA systems that can generalize to real-world scenarios, and handling complex questions that require reasoning and common-sense knowledge is another essential aspect.

The review underscores the critical role of images in multimodal understanding, offering contextual cues, resolving language ambiguities, and serving as essential references. It also highlights the persistent imbalance between the modalities and the influence of data biases in decision-making. The taxonomy introduced in this review categorizes various VQA techniques, applications, and challenges, providing a comprehensive overview of state-of-the-art multimodal deep learning approaches for VQA.

In the pursuit of enhancing the accuracy and context-awareness of answers to questions about images, the exploration of image feature extraction within the realm of multimodal deep learning research remains an exciting and evolving frontier. As technology continues to advance, the fusion of visual and textual information promises to provide more precise and accurate responses in VQA.

References

1) Cao J, Gan Z, Cheng Y, Yu L, Chen YCC, Liu J. Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models. *Computer Vision – ECCV 2020*. 2020;p. 565–580. Available from: https://doi.org/10.48550/arXiv.2005.07310.
2) Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A Simple and Performant Baseline for Vision and Language. 2019. Available from: http://arxiv.org/abs/1908.03557.
3) Liu L, Su X, Guo H, Zhu D. A Transformer-based Medical Visual Question Answering Model. *2022 26th International Conference on Pattern Recognition (ICPR)*. 2022;2022:1712–1720. Available from: https://doi.org/10.1109/ICPR56361.2022.9956469.
4) Yamada M, Amario D, Takemoto V, Boix K, Sasaki X, Sasaki T. Transformer Module Networks for Systematic Generalization in Visual Question Answering. 2022. Available from: http://arxiv.org/abs/2201.11316.
5) Sikarwar A, Kreiman G. On the Efficacy of Co-Attention Transformer Layers in Visual Question Answering. 2022. Available from: http://arxiv.org/abs/2201.03965.
6) Yu Z, Jin Z, Yu J, Xu M, Wang H, Fan J. Bilaterally Slimmable Transformer for Elastic and Efficient Visual Question Answering. *IEEE Transactions on Multimedia*. 2023;p. 1–15. Available from: https://doi.org/10.48550/arXiv.2203.12814.
7) Heo YJJ, Kim ESS, Choi WS, Zhang BTT. Hypergraph Transformer: Weakly-Supervised Multi-hop Reasoning for Knowledge-based Visual Question Answering. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022. Available from: https://doi.org/10.48550/arXiv.2204.10448.
8) Seenivasan L, Islam M, Krishna AK, Ren H. Surgical-VQA: Visual Question Answering in Surgical Scenes Using Transformer. *Lecture Notes in Computer Science*. 2022;p. 33–43. Available from: https://doi.org/10.48550/arXiv.2206.11053.
9) Siebert T, Clasen KN, Ravanbakhsh M, Demir B. Multi-modal fusion transformer for visual question answering in remote sensing. *Image and Signal Processing for Remote Sensing XXVIII*. 2022. Available from: https://doi.org/10.48550/arXiv.2210.04510.
10) Bazi Y, Rahhal MMA, Mekhalfi ML, Zuair MAA, Melgani F. Bi-Modal Transformer-Based Approach for Visual Question Answering in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*. 2022;60:1–11. Available from: https://doi.org/10.1109/TGRS.2022.3192460.
11) Zhang H, Wu W. CAT: Re-Conv Attention in Transformer for Visual Question Answering. *2022 26th International Conference on Pattern Recognition (ICPR)*. 2022;2022:1471–1478. Available from: https://doi.org/10.1109/ICPR56361.2022.9956247.
12) Ding H, Li LE, Hu Z, Xu Y, Hakkani-Tur D, Du Z, et al. Semantic Aligned Multi-modal Transformer for Vision-LanguageUnderstanding: A Preliminary Study on Visual QA. *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*. 2021. Available from: https://assets.amazon.science/4f/aa/de1faccc4ff9866c76b82e57a29b/multimodal-reltag-final-34.pdf.
13) Khan AU, Mazaheri A, Da VLN, Shah M, Mmft-Bert. Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering. 2020. Available from: http://arxiv.org/abs/2010.14095.
14) Ye S, Kong W, Yao C, Ren J, Jiang X. Video Question Answering Using Clip-Guided Visual-Text Attention. *2023 IEEE International Conference on Image Processing (ICIP)*. 2023. Available from: https://doi.org/10.48550/arXiv.2303.03131.
15) Mishra A, Anand A, Guha P. Dual Attention and Question Categorization-Based Visual Question Answering. *IEEE Transactions on Artificial Intelligence*. 2023;4(1):81–91. Available from: https://doi.org/10.1109/TAI.2022.3160418.
16) Pan H, He S, Zhang K, Qu B, Chen C, Shi K. AMAM: An Attention-based Multimodal Alignment Model for Medical Visual Question Answering. *Knowledge-Based Systems*. 2022;255:109763. Available from: https://doi.org/10.1016/j.knosys.2022.109763.
17) Song J, Zeng P, Gao L, Shen HT. From Pixels to Objects: Cubic Visual Attention for Visual Question Answering. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 2018. Available from: https://doi.org/10.48550/arXiv.2206.01923.
18) Huang X, Gong H. A Dual-Attention Learning Network with Word and Sentence Embedding for Medical Visual Question Answering. *IEEE Transactions on Medical Imaging*. 2023;p. 1–1. Available from: https://doi.org/10.48550/arXiv.2210.00220.
19) Xia Q, Yu C, Hou Y, Peng P, Zheng Z, Chen W. Multi-Modal Alignment of Visual Question Answering Based on Multi-Hop Attention Mechanism. *Electronics*. 2022;11(11):1778. Available from: https://doi.org/10.3390/electronics11111778.
20) Yang Z, Wu L, Wen P, Chen P. Visual Question Answering reasoning with external knowledge based on bimodal graph neural network. *Electronic Research Archive*. 2023;31(4):1948–1965. Available from: https://doi.org/10.3934/era.2023100.
21) Hu X, Gu L, Kobayashi K, Chen AQ, Lu Q, Lu Z, et al. Interpretable Medical Image Visual Question Answering via Multi-Modal Relationship Graph Learning. 2023. Available from: http://arxiv.org/abs/2302.09636.
22) Jiang L, Meng Z. Knowledge-Based Visual Question Answering Using Multi-Modal Semantic Graph. *Electronics*. 2023;12(6):1390. Available from: https://doi.org/10.3390/electronics12061390.

23) Qian Y, Hu Y, Wang R, Feng F, Wang X. Question-Driven Graph Fusion Network for Visual Question Answering. *2022 IEEE International Conference on Multimedia and Expo (ICME)*. 2022. Available from: https://doi.org/10.48550/arXiv.2204.00975.
24) Li M, Moens MFF. Dynamic Key-Value Memory Enhanced Multi-Step Graph Reasoning for Knowledge-Based Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*;36(10):10983–10992. Available from: https://doi.org/10.48550/arXiv.2203.02985.
25) Li H, Li X, Karimi B, Chen J, Sun M. Joint Learning of Object Graph and Relation Graph for Visual Question Answering. *2022 IEEE International Conference on Multimedia and Expo (ICME)*. 2022. Available from: https://doi.org/10.48550/arXiv.2205.04188.
26) Zhu Z. From Shallow to Deep: Compositional Reasoning over Graphs for Visual Question Answering. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022. Available from: https://doi.org/10.48550/arXiv.2206.12533.
27) Wang Y, Yasunaga M, Ren H, Wada S, Leskovec J, Vqa-Gnn. Reasoning with Multimodal Semantic Graph for Visual Question Answering. 2022. Available from: http://arxiv.org/abs/2205.11501.
28) Salaberria A, Azkune G, De Lacalle OL, Soroa A, Agirre E. Image captioning for effective use of language models in knowledge-based visual question answering. *Expert Systems with Applications*. 2023;212:118669. Available from: https://doi.org/10.48550/arXiv.2109.08029.
29) Zhu H, Togo R, Ogawa T, Haseyama M. Interpretable Visual Question Answering Referring to Outside Knowledge. *2023 IEEE International Conference on Image Processing (ICIP)*. 2023. Available from: https://doi.org/10.48550/arXiv.2303.04388.
30) Lerner P, Ferret O, Guinaudeau C. Multimodal Inverse Cloze Task for Knowledge-Based Visual Question Answering. *Lecture Notes in Computer Science*. 2023;p. 569–587. Available from: https://doi.org/10.48550/arXiv.2301.04366.
31) Shao Z, Yu Z, Wang M, Yu J. Prompting Large Language Models with Answer Heuristics for Knowledge-Based Visual Question Answering. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. Available from: https://doi.org/10.48550/arXiv.2303.01903.
32) Liu Y, Li G, Lin L. Causality-aware Visual Scene Discovery for Cross-Modal Question Reasoning. 2023. Available from: http://arxiv.org/abs/2304.08083.
33) Li Z, Guo Y, Wang K, Wei Y, Nie L, Kankanhalli M. Joint Answering and Explanation for Visual Commonsense Reasoning. *IEEE Transactions on Image Processing*. 2023;32:3836–3846. Available from: https://doi.org/10.48550/arXiv.2202.12626.
34) Li H, Huang J, Jin P, Song GP, Wu Q, Chen J. Weakly-Supervised 3D Spatial Reasoning for Text-Based Visual Question Answering. *IEEE Transactions on Image Processing*. 2023;32:3367–3382. Available from: https://doi.org/10.1109/TIP.2023.3276570.
35) Shen X, Han D, Chen C, Luo G, Wu Z. An effective spatial relational reasoning networks for visual question answering. *PLOS ONE*. 2022;17(11):e0277693–e0277693. Available from: https://doi.org/10.1371/journal.pone.0277693.
36) Wu Y, Ma Y, Wan S. Multi-scale relation reasoning for multi-modal Visual Question Answering. *Signal Processing: Image Communication*. 2021;96:116319. Available from: https://doi.org/10.1016/j.image.2021.116319.
37) Han Y, Yin J, Wu J, Wei Y, Nie L. Semantic-Aware Modular Capsule Routing for Visual Question Answering. *IEEE Transactions on Image Processing*. 2023;32:5537–5549. Available from: https://doi.org/10.48550/arXiv.2207.10404.
38) Cao Q, Liang X, Wang K, Lin L. Linguistically Driven Graph Capsule Network for Visual Question Reasoning. 2020. Available from: http://arxiv.org/abs/2003.10065.
39) Tian W, Li H, Zhao ZQ. Dual Capsule Attention Mask Network with Mutual Learning for Visual Question Answering. 2022. Available from: https://aclanthology.org/2022.coling-1.500.
40) Zhou Y, Ji R, Su JR, Sun X, Chen W. Dynamic Capsule Attention for Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*;33(01):9324–9331. Available from: https://doi.org/10.1609/aaai.v33i01.33019324.
41) Khan AU, Kuehne H, Gan C, Lobo NDV, Shah M. Weakly Supervised Grounding for VQA in Vision-Language Transformers. *Lecture Notes in Computer Science*. 2022;p. 652–670. Available from: https://doi.org/10.48550/arXiv.2207.02334.
42) Bai L, Islam M, Seenivasan L, Ren H. Surgical-VQLA:Transformer with Gated Vision-Language Embedding for Visual Question Localized-Answering in Robotic Surgery. *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2023;p. 6859–6865. Available from: https://doi.org/10.48550/arXiv.2305.11692.
43) Seenivasan L, Islam M, Kannan G, Ren H. SurgicalGPT: End-to-End Language-Vision GPT for Visual Question Answering in Surgery. *Lecture Notes in Computer Science*. 2023;p. 281–290. Available from: https://doi.org/10.48550/arXiv.2304.09974.
44) Liu Y, Li G, Lin L. Causality-aware Visual Scene Discovery for Cross-Modal Question Reasoning. 2023. Available from: http://arxiv.org/abs/2304.08083.
45) Gupta D, Attal K, Demner-Fushman D. A dataset for medical instructional video classification and question answering. *Scientific Data*. 2023;10(1). Available from: https://doi.org/10.1038/s41597-023-02036-y.
46) Bazi Y, Rahhal MMA, Bashmal L, Zuair M. Vision–Language Model for Visual Question Answering in Medical Imagery. *Bioengineering*. 2023;10(3):380–380. Available from: https://doi.org/10.3390/bioengineering10030380.
47) Biten AF, Litman R, Xie Y, Appalaraju S, Manmatha R. LaTr: Layout-Aware Transformer for Scene-Text VQA. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022;2022:16527–16564. Available from: https://doi.org/10.48550/arXiv.2112.12494.
48) Tiong AMH, Li J, Li B, Savarese S, Hoi SCH. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training. *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022. Available from: https://doi.org/10.48550/arXiv.2210.08773.
49) Yuan Z, Mou L, Xiong Z, Zhu XX. Change Detection Meets Visual Question Answering. *IEEE Transactions on Geoscience and Remote Sensing*. 2022;60:1–13. Available from: https://doi.org/10.48550/arXiv.2112.06343.
50) Le T, Nguyen HT, Nguyen L, and. Vision And Text Transformer For Predicting Answerability On Visual Question Answering. *Proceedings - International Conference on Image Processing*. 2021;p. 934–942. Available from: https://doi.org/10.1109/ICIP42928.2021.9506796.
51) Yan M, Xu H, Li C, Tian J, Bi B, Wang W, et al. Achieving Human Parity on Visual Question Answering. *ACM Transactions on Information Systems*. 2023;41(3):1–40. Available from: https://doi.org/10.48550/arXiv.2111.08896.
52) Li P, Liu G, Tan L, Liao J, Zhong S. Self-Supervised Vision-Language Pretraining for Medial Visual Question Answering. *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. 2023. Available from: https://doi.org/10.48550/arXiv.2211.13594.
53) He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. Available from: https://doi.org/10.1109/CVPR.2016.90.
54) Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017;39(6):1137–1149. Available from: https://doi.org/10.48550/arXiv.1506.01497.
55) Shuang K, Guo J, Wang Z. Comprehensive-perception dynamic reasoning for visual question answering. *Pattern Recognition*. 2022;131:108878. Available from: https://doi.org/10.1016/j.patcog.2022.108878.
56) Zhang S, Chen M, Chen J, Zou F, Li YFF, Lu P. Multimodal feature-wise co-attention method for visual question answering. *Information Fusion*. 2021;73:1–10. Available from: https://doi.org/10.1016/j.inffus.2021.02.022.

57) Guo D, Xu C, Tao D. Bilinear Graph Networks for Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems*. 2023;34(2):1023–1034. Available from: https://doi.org/10.48550/arXiv.1907.09815.
58) Liu Y, Guo Y, Yin J, Song X, Liu W, Nie L, et al. Answer Questions with Right Image Regions: A Visual Attention Regularization Approach. *ACM Transactions on Multimedia Computing, Communications, and Applications*. 2022;18(4):1–18. Available from: https://doi.org/10.48550/arXiv.2102.01916.
59) Sharma H, Jalal AS. A framework for visual question answering with the integration of scene-text using PHOCs and fisher vectors. *Expert Systems with Applications*. 2022;190:116159. Available from: https://doi.org/10.1016/j.eswa.2021.116159.
60) Zhang W, Yu J, Zhao W, Ran C. DMRFNet: Deep Multimodal Reasoning and Fusion for Visual Question Answering and explanation generation. *Information Fusion*. 2021;72:70–79. Available from: https://doi.org/10.1016/j.inffus.2021.02.006.
61) Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018;p. 6077–6086. Available from: https://doi.org/10.48550/arXiv.1707.07998.
62) Whitehead S, Wu H, Ji H, Feris R, Saenko K. Separating Skills and Concepts for Novel Visual Question Answering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021. Available from: https://doi.org/10.48550/arXiv.2107.09106.
63) Yang X, Gao C, Zhang H, Cai J. Auto-Parsing Network for Image Captioning and Visual Question Answering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021;p. 2177–2187. Available from: https://doi.org/10.48550/arXiv.2108.10568.
64) Banerjee P, Gokhale T, Yang Y, Baral C. Weakly Supervised Relative Spatial Reasoning for Visual Question Answering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. Available from: https://doi.org/10.48550/arXiv.2109.01934.
65) Gamage B, Hong LC. Improved RAMEN: Towards Domain Generalization for Visual Question Answering. 2021. Available from: http://arxiv.org/abs/2109.02370.
66) Chen L, Zheng Y, Niu Y, Zhang H, Xiao J. Counterfactual Samples Synthesizing and Training for Robust Visual Question Answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023;p. 1–16. Available from: https://doi.org/10.48550/arXiv.2003.06576.
67) Nguyen BX, Do T, Tran H, Tjiputra E, Tran QD, Nguyen AX. Coarse-to-Fine Reasoning for Visual Question Answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2022. Available from: https://doi.org/10.48550/arXiv.2110.02526.
68) Cao J, Qin X, Zhao S, Shen J. Bilateral Cross-Modality Graph Matching Attention for Feature Fusion in Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems*. 2022;p. 1–12. Available from: https://doi.org/10.48550/arXiv.2112.07270Focustolearnmore.
69) Ben-Younes H, Cadene R, Thome N, Cord M. BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*;33(01):8102–8109. Available from: https://doi.org/10.48550/arXiv.1902.00038.
70) Guo W, Zhang Y, Yang J, Yuan X. Re-Attention for Visual Question Answering. *IEEE Transactions on Image Processing*. 2021;30:6730–6743. Available from: https://doi.org/10.1109/TIP.2021.3097180.
71) Liu F, Xu G, Wu Q, Du Q, Jia W, Tan M. Cascade Reasoning Network for Text-based Visual Question Answering. *Proceedings of the 28th ACM International Conference on Multimedia*. 2020;p. 4060–4069. Available from: https://tanmingkui.github.io/files/publications/Cascade.pdf.
72) Dancette C, Cadene R, Teney D, Cord M. Beyond Question-Based Biases: Assessing Multimodal Shortcut Learning in Visual Question Answering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. Available from: https://doi.org/10.48550/arXiv.2104.03149.
73) Yang C, Wu W, Wang Y, Zhou H. Multi-Modality Global Fusion Attention Network for Visual Question Answering. *Electronics*. 2020;9(11):1882–1882. Available from: https://doi.org/10.3390/electronics9111882.
74) Jiang L, Meng Z. Knowledge-Based Visual Question Answering Using Multi-Modal Semantic Graph. *Electronics*. 2023;12(6):1390–1390. Available from: https://doi.org/10.3390/electronics12061390.
75) Park S, Hwang S, Hong J, Byun H. Fair-VQA: Fairness-Aware Visual Question Answering Through Sensitive Attribute Prediction. *IEEE Access*. 2020;8:215091–215099. Available from: https://doi.org/10.1109/ACCESS.2020.3041503.
76) Xiong P, You Q, Yu P, Liu Z, Wu Y, Sa-Vqa. Structured Alignment of Visual and Semantic Representations for Visual Question Answering. 2022. Available from: http://arxiv.org/abs/2201.10654.
77) Xiong P, Shen Y, Jin H. MGA-VQA: Multi-Granularity Alignment for Visual Question Answering. 2022. Available from: http://arxiv.org/abs/2201.10656.
78) Ding Y, Yu J, Liu B, Hu Y, Cui M, Wu Q. MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. Available from: https://doi.org/10.48550/arXiv.2203.09138.
79) Wang R, Qian Y, Feng F, Wang X, Jiang H. Co-VQA : Answering by Interactive Sub Question Sequence. *Findings of the Association for Computational Linguistics: ACL 2022*. 2022. Available from: https://doi.org/10.48550/arXiv.2204.00879.
80) Gupta V, Li Z, Kortylewski A, Zhang C, Li Y, Yuille A. SwapMix: Diagnosing and Regularizing the Over-Reliance on Visual Context in Visual Question Answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. Available from: https://doi.org/10.48550/arXiv.2204.02285.
81) Chae J, Kim J. Uncertainty-based Visual Question Answering: Estimating Semantic Inconsistency between Image and Knowledge Base. *2022 International Joint Conference on Neural Networks (IJCNN)*. 2022. Available from: https://doi.org/10.48550/arXiv.2207.13242.
82) Nooralahzadeh F, Sennrich R. Improving the Cross-Lingual Generalisation in Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022;37(11):13419–13427. Available from: https://doi.org/10.48550/arXiv.2209.02982.
83) Wang J, Zhao Z, W J. Frame-Subtitle Self-Supervision for Multi-Modal Video Question Answering. 2022. Available from: http://arxiv.org/abs/2209.03609.
84) Reich D, Putze F, Schultz T. Visually Grounded VQA by Lattice-based Retrieval. 2022. Available from: http://arxiv.org/abs/2211.08086.
85) Si Q, Liu Y, Lin Z, Fu P, Wang W. Compressing And Debiasing Vision-Language Pre-Trained Models for Visual Question Answering. 2022. Available from: http://arxiv.org/abs/2210.14558.
86) Cao F, Luo S, Nunez F, Wen Z, Poon J, Han SC. SceneGATE: Scene-Graph Based Co-Attention Networks for Text Visual Question Answering. *Robotics*. 2022;12(4):114. Available from: https://doi.org/10.48550/arXiv.2212.08283.
87) Hu X, Gu L, Kobayashi K, Chen AQ, Lu Q, Lu Z, et al. Interpretable Medical Image Visual Question Answering via Multi-Modal Relationship Graph Learning. 2023. Available from: http://arxiv.org/abs/2302.09636.
88) Peng M, Wang C, Gao Y, Shi Y, Zhou XD. Temporal Pyramid Transformer with Multimodal Interaction for Video Question Answering. 2021. Available from: http://arxiv.org/abs/2109.04735.
89) Raj H, Dadhania J, Bhardwaj A, P K. Multi-Image Visual Question Answering. 2021. Available from: http://arxiv.org/abs/2112.13706.
90) Gupta P, Gupta M. NewsKVQA: Knowledge-Aware News Video Question Answering. *Advances in Knowledge Discovery and Data Mining*. 2022;p. 3–15. Available from: https://doi.org/10.48550/arXiv.2202.04015.

91) Piergiovanni A, Li W, Kuo W, Saffar M, Bertsch F, Angelova A. Answer-Me: Multi-Task Open-Vocabulary Visual Question Answering. 2022. Available from: http://arxiv.org/abs/2205.00949.
92) Xu Z, Zhong W, Su Q, Ou Z, Zhang F. Modeling Semantic Composition with Syntactic Hypergraph for Video Question Answering. 2022. Available from: http://arxiv.org/abs/2205.06530.
93) Chang S, Palzer D, Li J, Fosler-Lussier E, N X. MapQA: A Dataset for Question Answering on Choropleth Maps. 2022. Available from: http://arxiv.org/abs/2211.08545.
94) Tang X, Zhang W, Yu Y, Turner K, Derr T, Wang M, et al. Interpretable Visual Understanding with Cognitive Attention Network. *Lecture Notes in Computer Science*. 2021;p. 555–568. Available from: https://doi.org/10.48550/arXiv.2108.02924.
95) Wu T, Garcia N, Otani M, Chu C, Nakashima Y, Takemura H. Transferring Domain-Agnostic Knowledge in Video Question Answering. 2021. Available from: https://www.bmvc2021-virtualconference.com/assets/papers/1187.pdf.
96) Jiang J, Liu Z, Zheng N. LiVLR: A Lightweight Visual-Linguistic Reasoning Framework for Video Question Answering. *IEEE Transactions on Multimedia*. 2023;25:5002–5013. Available from: https://doi.org/10.1109/TMM.2022.3185900.
97) Wang J, Bao BKK, Xu C. DualVGR: A Dual-Visual Graph Reasoning Unit for Video Question Answering. *IEEE Transactions on Multimedia*. 2022;24:3369–3380. Available from: https://doi.org/10.1109/TMM.2021.3097171.
98) Gao L, Chen T, Li X, Zeng P, Zhao L, Li YF. Generalized pyramid co-attention with learnable aggregation net for video question answering. *Pattern Recognition*. 2021;120:108145–108145. Available from: https://doi.org/10.1016/j.patcog.2021.108145.
99) Manmadhan S, Kovoor BC. Multi-Tier Attention Network using Term-weighted Question Features for Visual Question Answering. *Image and Vision Computing*. 2021;115:104291. Available from: https://doi.org/10.1016/j.imavis.2021.104291.
100) Chen Z, Chen J, Geng Y, Pan JZ, Yuan Z, Chen H. Zero-Shot Visual Question Answering Using Knowledge Graph. *The Semantic Web – ISWC 2021*. 2021;p. 146–162. Available from: https://doi.org/10.48550/arXiv.2107.05348.
101) Yu Z, Yu J, Xiang C, Fan J, Tao DJ. Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems*. 2018;29(12):5947–5959. Available from: https://doi.org/10.48550/arXiv.1708.03619.
102) Chappuis C, Lobry S, Kellenberger B, Saux B, Le, Tuia D. How to find a good image-text embedding for remote sensing visual question answering?. 2021. Available from: http://arxiv.org/abs/2109.11848.
103) Ilievski I, Feng J. Multimodal Learning and Reasoning for Visual Question Answering. . Available from: https://proceedings.neurips.cc/paper/2017/hash/f61d6947467ccd3aa5af24db320235dd-Abstract.html.
104) Lan Y, Guo Y, Chen Q, Lin S, Chen Y, Deng X. Visual question answering model for fruit tree disease decision-making based on multimodal deep learning. *Frontiers in Plant Science*. 2023;13. Available from: https://doi.org/10.3389/fpls.2022.1064399.
105) Gao L, Zeng P, Song J, Li YFF, Liu W, Mei T, et al. Structured Two-Stream Attention Network for Video Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022;33(01):6391–6398. Available from: https://dl.acm.org/doi/pdf/10.1609/aaai.v33i01.33016391#:~:text=First%2C%20we%20infer%20rich%20long,focuses%20on%20the%20relevant%20text.
106) Ramnath K, Hasegawa-Johnson M. Seeing is Knowing! Fact-based Visual Question Answering using Knowledge Graph Embeddings. 2020. Available from: http://arxiv.org/abs/2012.15484.
107) Sood E, Kögel F, Müller P, Thomas D, Bâce M, Bulling A. Multimodal Integration of Human-Like Attention in Visual Question Answering. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2023. Available from: https://doi.org/10.48550/arXiv.2109.13139.
108) Gouthaman KV, Mittal A. On the role of question encoder sequence model in robust visual question answering. *Pattern Recognit*. 2022;131. Available from: https://doi.org/10.1016/j.patcog.2022.108883.
109) Reich D, Putze F, Schultz T. Adventurer's Treasure Hunt: A Transparent System for Visually Grounded Compositional Visual Question Answering based on Scene Graphs. 2021. Available from: http://arxiv.org/abs/2106.14476.
110) Zhang P, Lan H. Multiple Context Learning Networks for Visual Question Answering. *Sci Program*. 2022. Available from: https://doi.org/10.1155/2022/4378553.
111) Liu H, Gong S, Ji Y, Yang JY, Xing T, Liu C. Multimodal Cross-guided Attention Networks for Visual Question Answering. *Advances in Intelligent Systems Research*. 2018. Available from: https://www.atlantis-press.com/article/25897541.pdf.
112) Cascante-Bonilla P, Wu H, Wang L, Feris R, Ordonez V. Sim VQA: Exploring Simulated Environments for Visual Question Answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. Available from: https://doi.org/10.1109/CVPR52688.2022.00500.
113) Winterbottom T, Xiao S, Mclean A, Moubayed NA. Bilinear pooling in video-QA: empirical challenges and motivational drift from neurological parallels. *PeerJ Computer Science*. 2022;8:e974–e974. Available from: https://doi.org/10.7717/peerj-cs.974.
114) Lee D, Cheon Y, Han WSS. Regularizing Attention Networks for Anomaly Detection in Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*;35(3):1845–1853. Available from: https://doi.org/10.48550/arXiv.2009.10054.
115) Seo A, Kang GCC, Park J, Zhang BTT. Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021. Available from: https://doi.org/10.48550/arXiv.2106.10446.
116) Changpinyo S, Kukliansy D, Szpektor I, Chen X, Ding N, Soricut R. All You May Need for VQA are Image Captions. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022. Available from: https://doi.org/10.48550/arXiv.2205.01883.
117) Li H, Huang J, Jin P, Song GP, Wu Q, Chen J. Weakly-Supervised 3D Spatial Reasoning for Text-Based Visual Question Answering. *IEEE Transactions on Image Processing*. 2023;32:3367–3382. Available from: https://doi.org/10.1109/TIP.2023.3276570.
118) Pan H, He S, Zhang K, Qu B, Chen C, Shi K. MuVAM: A Multi-View Attention-based Model for Medical Visual Question Answering. 2021. Available from: http://arxiv.org/abs/2107.03216.
119) Salaberria A, Azkune G, De Lacalle OL, Soroa A, Agirre E. Image captioning for effective use of language models in knowledge-based visual question answering. *Expert Systems with Applications*. 2023;212:118669–118669. Available from: https://doi.org/10.48550/arXiv.2109.08029.
120) Girshick R, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014. Available from: https://doi.org/10.48550/arXiv.1311.2524.
121) Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014. Available from: http://arxiv.org/abs/1409.1556.
122) Sarkar A, Rahnemoonfar M. VQA-Aid: Visual Question Answering for Post-Disaster Damage Assessment and Analysis. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. 2021. Available from: https://doi.org/10.48550/arXiv.2106.10548.

123) Tan M, Le QV, Efficientnet. Rethinking Model Scaling for Convolutional Neural Networks. 2019. Available from: http://arxiv.org/abs/1905.11946.
124) Silva JD, Martins B, Magalhães J. Contrastive training of a multimodal encoder for medical visual question answering. *Intelligent Systems with Applications*. 2023;18:200221. Available from: https://doi.org/10.1016/j.iswa.2023.200221.
125) Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. Available from: https://doi.org/10.48550/arXiv.1506.02640.
126) Gómez L, Biten AF, Tito R, Mafla A, Rusiñol M, Valveny E. Multimodal grid features and cell pointers for scene text visual question answering. *Pattern Recognition Letter*. 2021;150:242–251. Available from: https://doi.org/10.48550/arXiv.2006.00923.
127) Brock A, De S, Smith SL, Simonyan K. 2021. Available from: http://arxiv.org/abs/2102.06171.
128) Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y. 2022. Available from: http://arxiv.org/abs/2204.14198.
129) Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T. 2020. Available from: http://arxiv.org/abs/2010.11929.
130) Thai TM, Luu ST. Integrating Image Features with Convolutional Sequence-to-sequence Network for Multilingual Visual Question Answering. 2023. Available from: http://arxiv.org/abs/2303.12671.
131) Li LH, Zhang P, Zhang H, Yang J, Li CH, Zhong Y, et al. Grounded Language-Image Pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. Available from: https://doi.org/10.48550/arXiv.2112.03857.
132) Lin Y, Xie Y, Chen D, Xu Y, Zhu C, Yuan L. REVIVE: Regional Visual Representation Matters in Knowledge-Based Visual Question Answering. 2022. Available from: http://arxiv.org/abs/2206.01201.
133) Radford A, Kim JW, Hallacy C, Ramesh A, Gabriel G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. 2021. Available from: http://arxiv.org/abs/2103.00020.
134) Sabour S, Frosst N, Hinton GE. Dynamic Routing Between Capsules. 2017. Available from: http://arxiv.org/abs/1710.09829.
135) Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a Method for Automatic Evaluation of Machine Translation. 2002. Available from: https://aclanthology.org/P02-1040.
136) Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. 2005. Available from: https://aclanthology.org/W05-0909.
137) Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. Available from: https://doi.org/10.48550/arXiv.1411.5726.
138) Anderson P, Fernando B, Johnson M, Gould SM. SPICE: Semantic Propositional Image Caption Evaluation. *Computer Vision – ECCV 2016*. 2016;p. 382–398. Available from: https://doi.org/10.48550/arXiv.1607.08822.
139) Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. 2004. Available from: https://aclanthology.org/W04-1013.