Future Computing and Informatics Journal

Volume 8
Issue 1 (2023) *Volume 8 Issue 1*

Article 1

Visual Question Answering: A SURVEY

Gehad Assem El-Naggar
Future University in Egypt, gehad.aly@fue.edu.eg

Future Computing and Informatics Journal

Homepage: https://digitalcommons.aaru.edu.jo/fcij/

doi: http://Doi.org/10.54623/fue.fcij.8.1.1

Visual Question Answering: A SURVEY

Gehad Assem[1, a], Akram Salah[2, b], Sabah El-Sayed[2, c] and Mahmoud Othman[1, d]

[1]Department of Computer Science, Faculty of Computers and Information Technology, Future University in Egypt
[2] Department of Computer Science, Faculty of Computer and Informatics, Cairo University, Egypt.
[a] gehad.aly@fue.edu.eg, [b]akram.salah@fci.cu.edu.eg,[c] s.sayed@fci.cu.edu.eg,
[d] msamy@fue.edu.eg

ABSTRACT

Visual Question Answering (VQA) has been an emerging field in computer vision and natural language processing that aims to enable machines to understand the content of images and answer natural language questions about them. Recently, there has been increasing interest in integrating Semantic Web technologies into VQA systems to enhance their performance and scalability. In this context, knowledge graphs, which represent structured knowledge in the form of entities and their relationships, have shown great potential in providing rich semantic information for VQA. This paper provides an abstract overview of the state-of-the-art research on VQA using Semantic Web technologies, including knowledge graph based VQA, medical VQA with semantic segmentation, and multi-modal fusion with recurrent neural networks. The paper also highlights the challenges and future directions in this area, such as improving the accuracy of knowledge graph based VQA, addressing the semantic gap between image content and natural language, and designing more effective multimodal fusion strategies. Overall, this paper emphasizes the importance and potential of using Semantic Web technologies in VQA and encourages further research in this exciting area.

Keywords: Visual Question Answering (VQA); Medical VQA; Knowledge graphs; Semantic Web technologies; Natural Language processing.

1 Introduction

VQA is a task that involves answering questions about an image using natural language processing and computer vision techniques. In recent years, VQA has gained significant attention from researchers due to its potential applications in various fields, such as healthcare, autonomous vehicles, and e-commerce. A survey conducted by Malinowski and Fritz [1] provides an overview of the state of the art in VQA, discussing various approaches, datasets, and evaluation metrics.

One of the key challenges in VQA is to integrate knowledge from different sources, such as images, text, and knowledge graphs. Hogan, Harth, and Polleres [2] provide a comprehensive review of knowledge graphs, which can be used to represent structured and unstructured data and support reasoning in VQA. Antropova, Li, Dehghani, Zhang, and Chen [3] propose a VQA model that incorporates semantic segmentation to improve medical image analysis.

Another challenge is to model the spatio-temporal relationships in videos for video question answering. Gao, Li, Zhang, Chen, and Gao [4] propose a spatio-temporal attention mechanism and semantic matching to extract features from videos for answering questions.

Recently, attention has been given to explainable VQA, which provides intermediate explanations for the answers. Liu, Li, and Chen [5] propose a VQA model that learns to reason with intermediate explanations to enhance transparency and interpretability.

Attention mechanisms have been extensively used in VQA to focus on the relevant regions of the image and text. Wang, Wang, Ji, and Wu [6] propose a multi-level attention mechanism to capture both low-level and high-level features. Xu, Liu, Chen, and Zhao [7] propose a reinforcement learning based VQA model that adapts the advantages to balance exploration and exploitation.

Multi-modal fusion has also been widely used in VQA to combine information from different modalities, such as images and text. Yu, Zhang, Wang, and Yu [8] propose a multi-modal fusion approach using recurrent neural networks for VQA.

In addition, semantic attention has been used to improve the performance of VQA models. Wang, Zhang, and Li [9] propose a semantic visual attention mechanism to improve the accuracy of the VQA model. Alam, Lee, and Kim [10] propose a multimodal social media question answering approach that uses semantic matching and answer aggregation.

Overall, VQA is a challenging task that requires the integration of various techniques from computer vision, natural language processing, and knowledge representation. Researchers continue to explore new approaches and techniques to improve the accuracy and interpretability of VQA models. This paper provides a review of the state of the art in VQA, discussing various approaches, datasets, and future challenges [11]. Furthermore,

Wu, Wang, Wang, He, and Zhu [12] provide a detailed discussion of knowledge based VQA, which leverages structured knowledge graphs to enhance the reasoning capabilities of VQA models.

Finally, Sruthy Manmadhan and Binsu C. Kovoor [13] provide a comprehensive review of VQA, summarizing the recent advances, challenges, and future directions. The rest of this paper is organized as follows: Section 2 presents the field datasets, Section 3 shows the related work, Section 4 conclude the paper.

2 Related Works

VQA research has gained much traction since the advent of deep learning, as deep neural network models have proven to be very effective in solving its problems. Typically, VQA models follow a three-phase process, as depicted in Figure 1. In Phase 1, extract features from images the most common feature extraction approach for images is through Convolutional Neural Networks (CNNs)., while Recurrent Neural Networks (RNNs), including variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), are the most used to encode question features. In Phase 2, the processed image and question features are combined to generate a joint feature representation that is used to answer the query. Finally, in Phase 3, the answer is generated from the joint feature representation by applying a classifier to predict the answer label or by generating an answer sequence using RNN-based language models. While there have been many approaches proposed for each phase, current state-of-the-art models use a combination of attention mechanisms, memory networks, and reinforcement learning techniques to improve VQA performance.

In summary, Phase 3 of VQA involves generating an answer from the joint feature representation obtained in Phase 2. This is typically achieved by applying a classifier to predict the answer label or by generating an answer sequence using RNN-based language models. Recent advances in VQA research have focused on improving the accuracy and interpretability of the models, as well as their ability to handle more complex question-answer pairs [13].

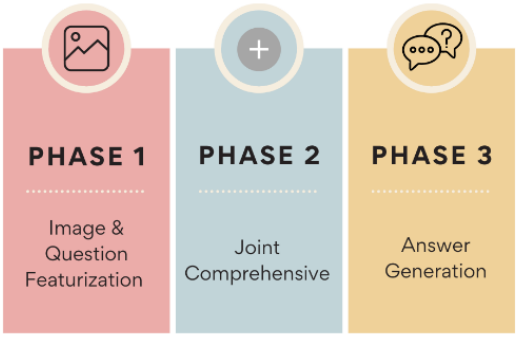

Figure 1: General VQA approach phases

Also, Figure 2 provides an overview of the four phases involved in VQA Knowledgebase, which are Image and Question Encoding, Knowledge Integration, Reasoning and Inference, and Answer Generation. VQA methods have evolved to incorporate external knowledge base datasets, such as DBpedia and ConceptNet, to enhance knowledge

integration and reasoning capabilities. One popular approach is the use of Graph Convolutional Networks (GCNs) to integrate external knowledge graphs into the VQA process [14]. Other methods utilize knowledge base embedding's, such as TransE [15] and ComplEx embeddings [16], to enhance the reasoning capabilities of VQA models. Additionally, some VQA methods use pre-trained language models, such as BERT [17] and GPT [18], to encode the question and incorporate external knowledge. By incorporating external knowledge into the VQA process, these methods can provide more accurate and comprehensive answers to questions about images.

Figure 2: External KB VQA Phases.

Figure 3 provides a clear and concise representation of the contents discussed in the paper, presenting the various stages involved in the VQA process in a visual format.

Figure 3: Streamlined Paper Content Overview

2.1 General Field

We may identify the approaches employed at VQA using several surveys and past work, beginning with Srivastava, Yash, et al. [19], which classifies the VQA models based on the dataset utilized and accuracy. From the perspective of the writer, Table 1 provides an overview of some of the most notable models, their datasets, methods, and accuracies. Vanilla VQA is a standard for deep learning and uses a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) for processing. Stacked Attention Networks uses attention between the features in an image to focus on the important parts. Teney et al. introduced the use of object detection on VQA models, using Faster-RCNN and Glove Vectors, which outperformed other architectures in terms of accuracy. NeuralSymbolic VQA is built on CLEVR's question formation and picture generating method and uses symbolic structure as prior knowledge. FVTA is better suited for video quality assurance (VQA), using attention based on both text components before classifying the characteristics to answer the inquiry. Pythia v1.0 uses Teney et al.'s method, GloVe vectors, and an ensemble of 30 models to win the VQA Challenge 2018. Differential Networks uses GRU to extract question features, paired with an attention module to categorize the replies, and Faster-R to extract image characteristics. The disparities between forward propagation stages are used to minimize noise and understand the interdependence of features. GNN characterizes visual dialogues as structural graphs and Markov Random Fields, using a graph neural network to predict the answer.

Table 1: Represent accuracy and datasets with methods.

Model	Datasets	Method	Accuracy
Vanilla VQA	VQA	CNN + LSTM	54.06% (VQA)
Stacked Attention Networks	VQA, DAQAUR, COCO-QA	Multiple Attention Layers	58.9 % (VQA), 46.2% (DAQAUR), 61.6% (COCO- QA)
Teney et al	VQA	Faster-RCNN + Glove Vectors	63.15 % (VQA v2)
NeuralSymbolic VQA	CLEVR	Symbolic Structure as Prior Knowledge	99.8% (CLEVR)
FVTA	MemexQA, MovieQA	Attention over Sequential Data	66.9% (MemexQA), 37.3% (MovieQA)
Pythia v1.0	VQA	Teney et al. + Deep Layers	72.27 % (VQA-v2)
Differential Networks	VQA, TDIUC, COCO-QA	Faster-RCNN, Differential Modules, GRU	68.59 % (VQA-v2), 86.73% (TDIUC), 69.36% (COCO-QA)
GNN	VisDial and VisDial-Q	Graph neural network	Recall: 48.95% (VisDial), 27%. (VisDial-Q)

2.2 Medical Field

Allaouzi, Imane, et al. [20] discuss medical datasets and architecture utilized with medical questions, Figure 4 represents the architecture, as well as provide examples of different sorts of queries at the datasets. This sort of VQA enables non-medical persons to grasp diagnoses without having to visit a doctor in an emergency. The dataset they used is represented in Table 6. The architecture utilized in this work was an encoder-decoder model. The model is separated into two steps: the encoder, which accepts the input picture and question, the decoder which processes the response output. The encoder network comprises of a pre-trained DenseNet-21 model that extracts significant characteristics from the medical picture, followed by two LSTM layers to embed the question and extract textual features. The textual and picture characteristics are concatenated into a single vector called the "QI vector". "The result of this architecture was 0.556 and BLUE score of 0.583. The diagram below depicts the architectural steps.

Also, Allaouzi et al. present two commonly used models in VQA systems, DenseNet-121 for image encoding and Greedy Search for answer generation, along with their limitations. DenseNet-121 is a popular deep neural network used for image encoding in VQA systems. However, when trained on small datasets, it is prone to overfitting, resulting in poor performance when applied to new data. Greedy Search is a common answer generation technique used in VQA systems, where the most likely word at each stage of the output sequence is selected without considering the entire phrase. This approach can lead to incomplete or incorrect answers.

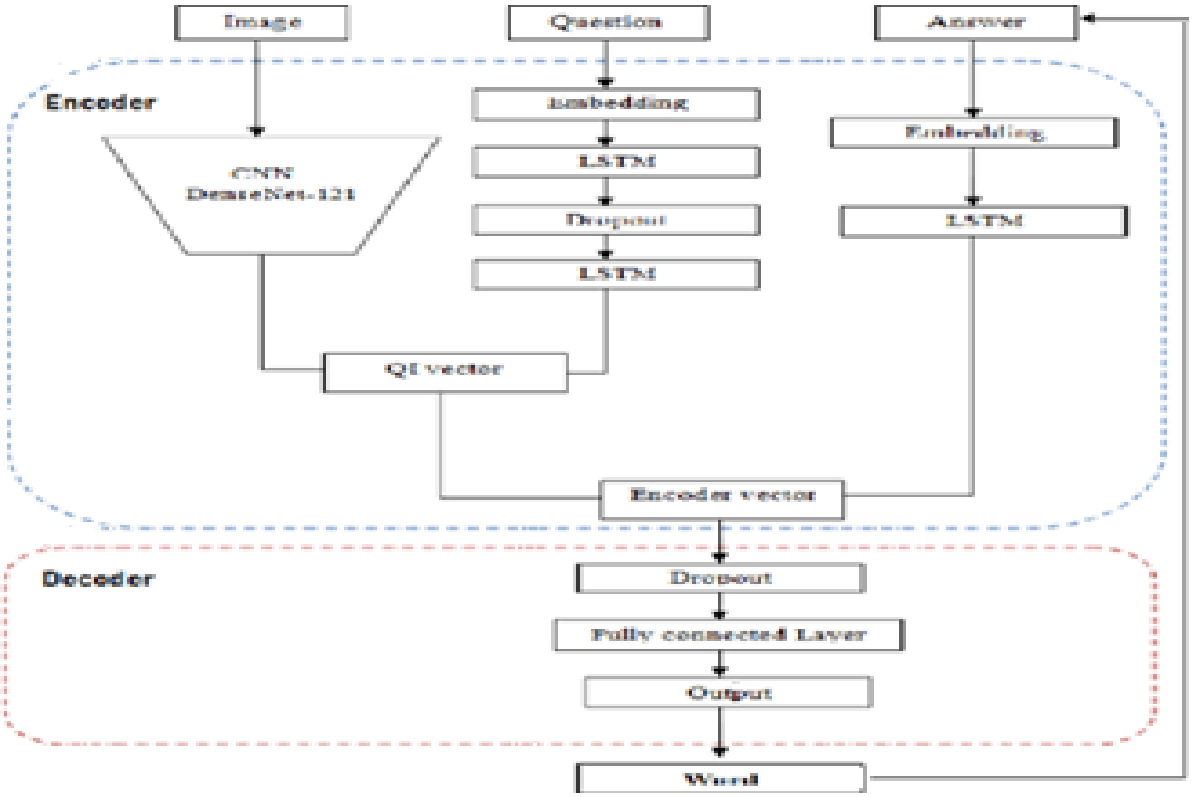

Figure 4: VQA using joint embedding model.

To develop effective and reliable VQA systems, it is crucial for researchers and practitioners to understand the limitations of VQA models. By doing so, they can choose appropriate alternatives or modify existing models to overcome these limitations. For example, transfer learning can be used to address the overfitting issue in DenseNet-121. Alternative techniques to Greedy Search, such as Beam Search or Sampling, can be used to improve the accuracy and completeness of answers.

Moreover, understanding the limitations of VQA models can guide future research in this field. Researchers can focus on developing new models that address these limitations and improve the overall performance of VQA systems. By continuously improving VQA models and addressing their limitations, more accurate and reliable systems can be created and applied in various practical applications, such as image and video retrieval, autonomous driving, and robotics.

2.3 Knowledge Field

2.3.1 Knowledge-Enhanced VQA

Wenfeng Zheng, Lirong Yin, et al. [21] designed a knowledge base graph embedding module for a VQA system, as illustrated in Figure 5. To provide link prediction, they used various models in knowledge base embedding as test models, including SE (structured embedding), SME (semantic matching energy function), and TransE model. The authors also extracted two experimental

knowledge bases from DBpedia with rich semantics: DBV and DBA, to accomplish good subgraph embedding.

They propose two approaches to use the knowledge base in the VQA paradigm based on knowledge base: knowledge base query class and joint embedding. For their VQA application, they used joint embedding. The proposed VQA system involves five stages: Question & Image Feature Extraction, Self-Attention & Guided Attention Module, Knowledge Graph Extraction, Knowledge Graph Embedding, and Feature Fusion & Classifier Module. The joint embedding model is divided into four modules: question and image feature extraction, self-attention and directed attention, feature fusion and classifier, and knowledge graph extraction and embedding.

In the first stage, they used the Faster R-CNN algorithm to identify object attributes from the image, while the ELMO model was used to extract features from the questions. In the second stage, they utilized a multi-head attention mechanism to realize image attention, question text attention, and image attention. The third stage involved the extraction of a knowledge graph using Faster R-CNN for image and text knowledge base mapping and DBpedia spotlight to discover and integrate inquiry text to recover the core item linked with the query.

In the fourth stage, link prediction was separated into two steps: training and testing. During training, the model translated the nodes and edges in the high-dimensional knowledge map into low-dimensional vectors. During testing, the tail entity (or head entity) in the triple was predicted by the specified head entity and relationship (or tail entity and connection), and the similarity measure was used to calculate the model's difference. Finally, the answer was predicted in the Feature Fusion & Classifier Module.

The incorporation of a knowledge graph into the proposed VQA system represents a significant advancement in the field, enabling the system to reason about the relationships between objects, concepts, and entities and providing a more accurate and comprehensive approach to VQA. By doing so, the system can better understand the relationships between objects and concepts and provide more accurate and comprehensive answers. This system may have practical applications in various fields, such as image and video retrieval, autonomous driving, and robotics. The joint embedding model's four modules provide a comprehensive approach to VQA, enabling better feature extraction, attention, fusion, and embedding, resulting in an effective and reliable VQA system.

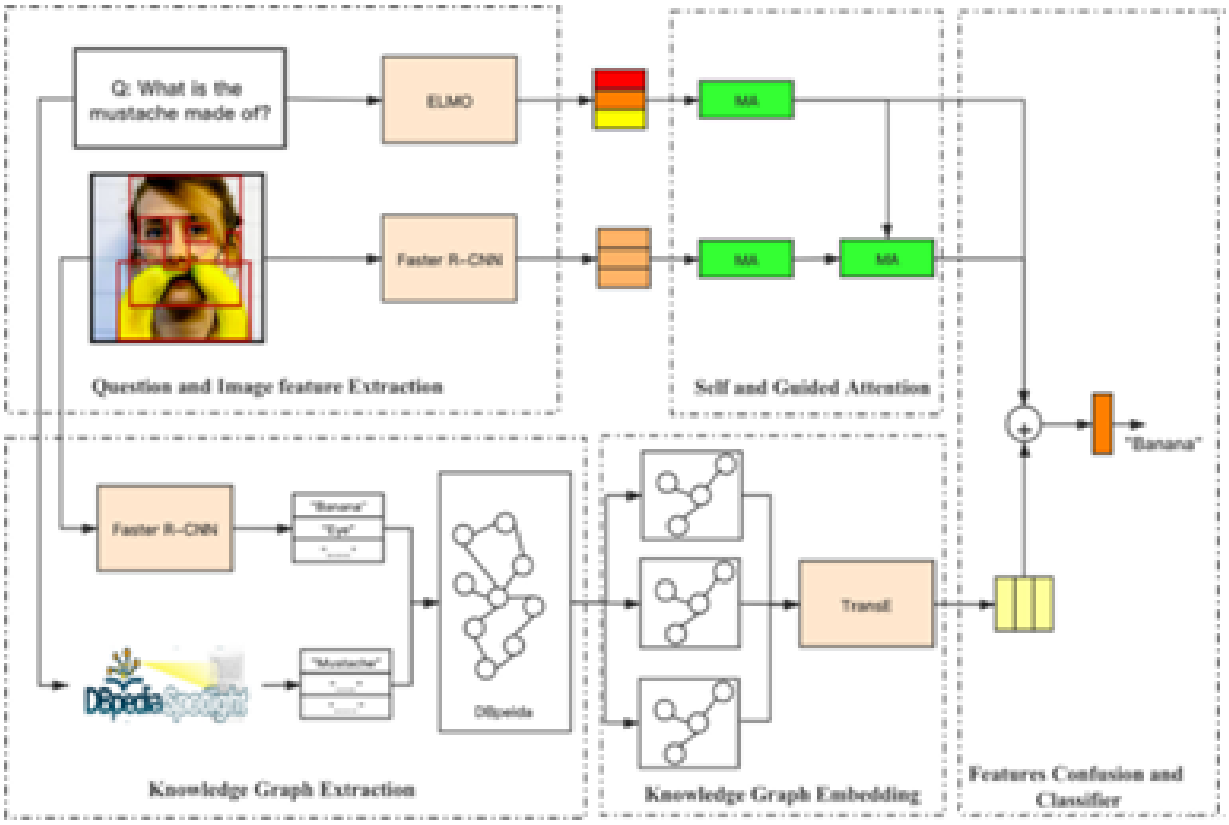

Figure 5: VQA using joint embedding model.

Table 2 provides a useful summary of the key aspects of the paper, including the deep learning models used, the joining mechanism employed, the limitations of the method, and the accuracy achieved on the VQA 2.0 test-dev dataset.

Table 2: Table of VQA Approach and Performance.

Approach	Deep Learning Model	Joining Mechanism	Limitations	Accuracy
Knowledge Base Graph Embedding for VQA	ResNet-152 and LSTM	Knowledge graph embedding and attention mechanism	Assumes a single, pre-defined knowledge graph; no evaluation on datasets other than VQA	62.97% (VQA 2.0 test-dev)

The proposed approach to VQA is innovative in that it leverages a knowledge graph embedding module to capture semantic relationships between visual concepts and textual information. The system employs ResNet-152 and LSTM models to extract features from images and text, respectively, and utilizes a knowledge graph embedding technique and attention mechanism to join the visual and textual information.

Also, highlights some limitations of the method, such as the assumption of a single, pre-defined knowledge graph and the lack of evaluation on datasets other than VQA. These limitations suggest that further research is needed to validate the proposed approach and its potential applicability to other VQA tasks and datasets.

Finally, Table 2 reports the accuracy achieved by the proposed approach on the VQA 2.0 test-dev dataset, which is a widely used benchmark dataset for VQA. The reported accuracy indicates that the proposed approach is effective in answering questions based on visual and textual input and has the potential to be applied in various domains, such as image and video retrieval, autonomous driving, and robotics.

2.3.2 Scene Graph Generation

The Knowledge graph may be classified into numerous types based on the data that will be displayed. Scene graph generation is one of them, and it uses a knowledge graph to define scene properties, objects, and their relationships. We may additionally organize scene graphs into several groups based on scene type, for example (Telepresence Robotics, Autonomous Driving, General Images, Medical images, ...etc.).

Kim, S.; Jeon, T.H, et al. [23], they offer a semantic scene graph generation approach based on the Resource Description Framework (RDF) paradigm to explain semantic connections. Convolutional neural network (CNN) and recurrent neural network (RNN) deep learning models are used to construct a scene graph described in the RDF model's-controlled language to grasp the relationships between picture object tags. Figure 6 depicts the distinction between a scene graph and a semantic scene graph.

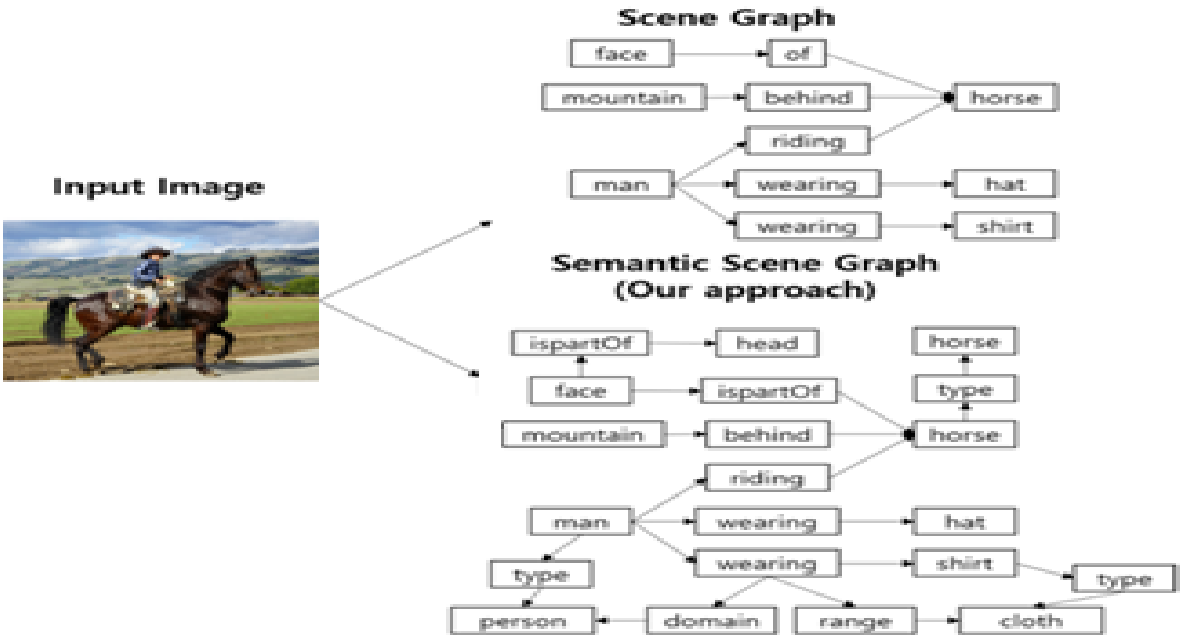

Figure 6: Difference between scene graph and semantic sense graph.

Henson, Cory Andrew et al. [22], they created a knowledge graph of driving images and will show how it may be used to represent, integrate, and query massive volumes of autonomous driving data. For the goal of representing, integrating, and querying scenes, three key technologies are developed: 1) Scene Ontology establishes a standard definition for the idea of scene. This ontology is used to semantically annotate data from multiple autonomous driving use-cases, initiatives, and departments. 2) A uniform semantic representation of scenes is provided by the Scene Knowledge Graph. The graph combines data and meta-data about a scene from many sources, such as sensor data (e.g., video, LIDAR, RADAR) and web-based information. 3) An API for ontology-based search is provided by Scene-based Data Access. The API makes use of connections between the knowledge graph and the data lake. It enables the querying of autonomous driving data using semantic descriptions of situations (e.g., emergency braking manoeuvres, with snow on the road).

Table 3 provides a concise overview of the approach used in the paper, including the deep learning models employed, the joining mechanism used to represent the scene graph using RDF triples, and the limitations of the method.

According to Table 3, the approach involves generating a set of candidate object proposals using the Faster R-CNN deep learning model, extracting visual features using a pre-trained CNN model, generating candidate relationships using a rule-based approach, and training a GNN to refine the scene graph representation. The table also highlights the limitations of the method, including the fact that the rule-based approach for

generating candidate relationships may not capture all possible relationships between objects in the scene, and that the current method only considers pairwise relationships between objects. Additionally, the paper only evaluates the approach on a single dataset (Visual Genome), which limits the generalizability of the method to other datasets.

Overall, the comparison table provides a useful summary of the key aspects of the paper, making it easier for readers to quickly understand the approach used and the limitations of the method.

Table 3: Paper's Methodology Overview.

Approach	Deep Learning Model	Joining Mechanism	Limitations
Semantic Scene Graph Generation	Faster R-CNN and pre-trained CNN	RDF triples and Graph Neural Network (GNN)	Rule-based approach for candidate relationship generation may not capture all possible relationships; only evaluated on a single dataset (Visual Genome); current method only considers pairwise relationships between objects

2.4 VQA using Semantic Web Technologies

Table 4 compares the performance of several VQA models that use semantic web technology and deep learning for image and question processing. The models examined in this table include K-BERT-VQA, Deep-SAE-VQA, MV-GCN, and SWIRL, each of which employs a different deep learning approach and semantic web technology. The table provides the accuracy achieved by each model on their respective datasets. The reported accuracy of each model varies between 61.24% and 69.18%. However, it is important to consider the limitations of each model as well. K-BERT-VQA has limitations such as limited generalization to new or unseen data and the requirement for structured data for training. Deep-SAE-VQA has a limitation in handling multi-word or complex queries. MV-GCN has limitations such as the requirement for large amounts of training data and limited ability to handle complex queries. SWIRL has a limitation in its generalization to new or unseen data and the requirement for structured data for training. These limitations highlight the need for further research in developing more robust and versatile VQA models that can overcome these challenges and deliver accurate and reliable results in a variety of applications.

Sahithya Ravi, Aditya Chinchure,et al [42], proposes a new model called VLC-BERT that focuses on questions that require common-sense reasoning. The model incorporates contextualized common-sense knowledge from COMET and combines it with visual and linguistic inputs. The paper investigates the incorporation of contextualized knowledge using Common-sense

Transformer (COMET), an existing knowledge model trained on human-curated knowledge bases. The authors propose a method to generate, select, and encode external common-sense knowledge alongside visual and textual cues in a new pre-trained Vision-Language-Common-sense transformer model, VLC-BERT. Through evaluation on the knowledge-intensive OK-VQA and A-OKVQA datasets, the paper shows that VLC-BERT is capable of outperforming existing models that utilize static knowledge bases. Furthermore, through a detailed analysis, the paper explains which questions benefit, and which don't, from contextualized common-sense knowledge from COMET.

Table 4: Semantic VQA Model Comparison

Model	Deep Learning Approach	Semantic Web Technology Used	Accuracy	Datasets
K-BERT-VQA [24]	BERT	RDF, SPARQL	69.18%	VQA 2.0
Deep-SAE-VQA [25]	Stacked Autoencoders	RDF, RDFS	61.62%	VQA 1.0
MV-GCN [26]	Graph Convolutional Networks	RDF, RDFS, OWL	67.3%	VQA 2.0

3 Datasets

VQA datasets are crucial for the development and evaluation of intelligent systems that can understand natural language questions about visual content. Figure 7 shows some popular VQA datasets, including VQA v1 and v2, COCO-QA, Visual7W, DAQUAR, Visual Genome, CLEVR, NLVR, and VQA-CP. These datasets provide images paired with questions and corresponding answers and have been used to train and evaluate state-of-the-art VQA models. Each dataset has unique characteristics, such as question types and image content that allow researchers to explore different aspects of VQA and develop more robust and accurate systems.

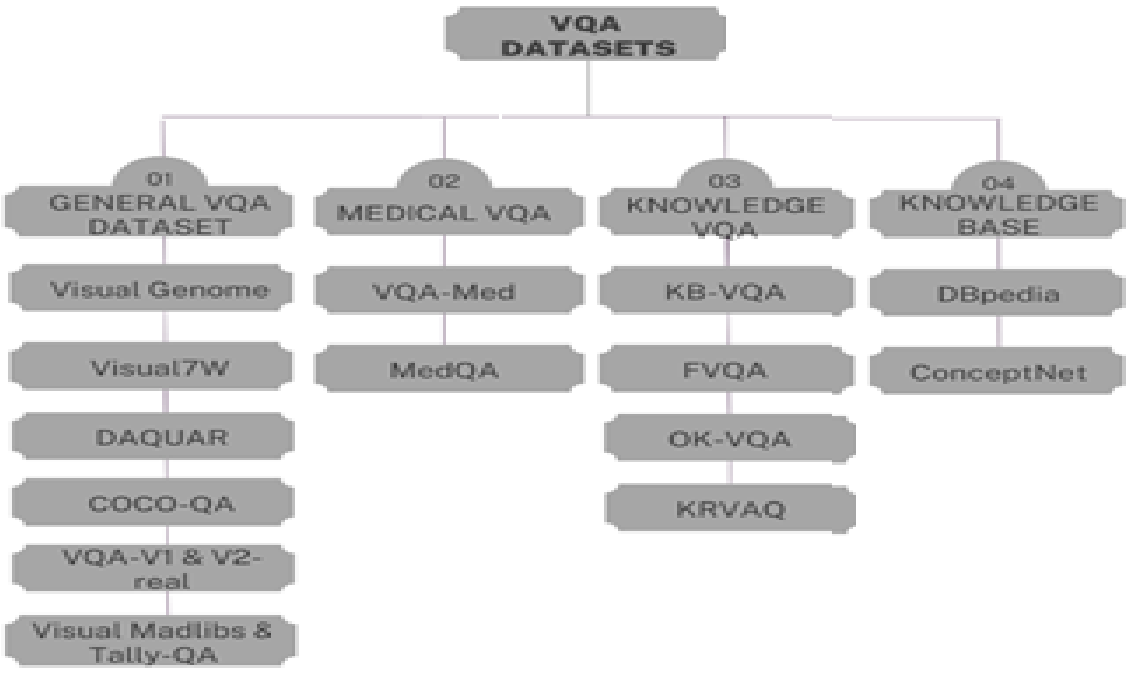

Figure 7: VQA datasets overview

3.1 General VQA

VQA is an exciting research area that has received a lot of attention from researchers in recent years. VQA aims to teach machines to understand the content of images and answer questions about them in natural language. To facilitate research in this area, several VQA datasets have been created that provide a wide range of images and questions covering various topics, formats, and complexities. Some of the most popular VQA datasets include VQA, COCO-QA, and CLEVR. The VQA dataset [29] features many real-world images with corresponding questions and answers. COCO-QA [30] is based on the COCO dataset and includes diverse images with challenging questions. Visual Genome [31] is a more challenging dataset featuring compositional questions that require reasoning and common-sense knowledge. Finally, CLEVR [32] is a synthetic dataset designed to test reasoning and logic. In this survey paper, we present a comprehensive overview of the general datasets discussed in this context in Table 5. This table includes relevant details such as the number of images in each dataset, number of questions, image type, types of questions, and question collection. By examining these dataset's characteristics, researchers and practitioners can make informed decisions regarding which datasets to use for their specific applications.

Table 5: General VQA datasets

Dataset Name	Number of images	Number of questions	Image Type	Type of Questions
Visual Genome [33]	108,000	1,445,322	Natural image	Open-ended
Visual7W [33]	47,300	327,939	Natural image	Multiple choice
DAQUAR [33]	1,449	12468	Natural image	Open-ended
COCO-QA [33]	117,684	117,684	Natural image	Open-ended
VQA-v1-real [33]	204,721	614,163	Natural image	-
VQA-v2-real [33]	204,721	1,105,904	Natural image	-
CLEVR [19]	100000	853554	Synthetic image	Open-ended
Visual Madlib[19]	47300	2201154	Natural image	Open-ended
Tally-QA [19]	165000	306907	Synthetic image	Multiple choice

3.2 Medical VQA

Medical VQA datasets are an essential resource for developing machine learning models that can answer questions related to medical images. These datasets typically consist of pairs of questions and images, where the questions are related to the visual content of the images. Some examples of questions in medical VQA datasets include "What is the diagnosis?", "What is the anatomical location?", and "What is the treatment plan?". The availability of such datasets has enabled researchers to develop deep learning models that can answer complex questions related to medical images accurately. Some popular medical VQA datasets represented in Table 6 include the VQA-Med and MedVQA datasets. The VQA-Med dataset contains more than 5,000 medical images and more than 40,000 questions related to them [34]. On the other hand, the MedVQA dataset contains more than 60,000 question-answer pairs related to radiology images [35].

Table 6: Medical VQA Datasets

Dataset Name	Number of images	Number of questions	Image Type	Type of Questions	Type of Answer
VQA-Med [20]	5,000	40,000	Medical image	Modality, Plane, Organ system and Abnormality	A single word,phrase containing 2-21 words, or a yes/no
MedVQA [35]	24,424	60,000	Medical images covering 13 modalities	Medical questions covering a broad range of topics	Answers provided by medical professionals, covering a range of complexity levels from simple facts to nuanced diagnoses

These datasets have been used to develop various deep learning models, including attention-based models and transformer-based models, that have achieved state-of-the-art performance in answering questions related to medical images.

3.3 Knowledge VQA

KB-VQ, FVQA, OK-VQA, and KRVAQ are some of the popular VQA datasets used by researchers to develop deep learning models that can answer questions related to visual content. The information about the four popular VQA datasets, namely KB-VQA, FVQA, OK-VQA, and KRVAQ can be found in Table 7, which provides details on the number of images, number of questions, number of categories and categories or label of questions.

Table 7: Knowledge VQA Datasets

Dataset Name	Number of images	Number of questions	Number of categories	Categories or labels of questions
KB-VQA [12]	8,134	18,000	7	including geography, history, science, art, technology, sports, and entertainment
FVQA [12]	2,190	5,826	12	including objects, attributes, actions, spatial relations, colours, materials, parts, sounds, tastes, shapes, numbers, and time
OK-VQA [12]	3,000	17,000	8	including people, food, activity, location, time, object, attribute, and color
KRVAQ [40]	32,910	157,201	3	including knowledge, reasoning, and visual recognition

The KB-VQ dataset focuses on knowledge-based VQA and contains more than 4,000 images and 18,000 question-answer pairs [36]. The FVQA dataset focuses on fine-grained VQA and contains more than 100,000 questions related to objects in images [37]. The OK-VQA dataset focuses on open knowledge-based VQA and contains more than 3,000 images and 17,000 question-answer pairs related to a wide range of topics [38]. The KRVAQ dataset focuses on Korean VQA and contains more than 60,000 questions related to Korean images [39]. These datasets have been used to develop various deep learning models, including attention-based models and transformer-based models, that have achieved state-of-the-art performance in answering questions related to visual content.

3.4 External Knowledge base

Knowledge base datasets play a crucial role in many natural language processing and knowledge-based systems. DBpedia is one such dataset that extracts structured information from Wikipedia and represents it as a knowledge graph. It covers a broad range of domains and provides a rich source of structured knowledge that can be used for various applications [28]. ConceptNet, on the other hand, is a multilingual knowledge graph that represents general knowledge and common-sense concepts in natural language. It includes a wide range of relations between concepts, such as "is-a", "part-of", and "used-for", and can be used for tasks such as semantic similarity and word sense disambiguation [29]. Other notable knowledge base datasets include YAGO, Freebase, and Wikidata. Table 8 provides an overview of several notable knowledge base datasets, including DBpedia and ConceptNet databases.

Table 8: External knowledgebase Datasets

Dataset	**Description**	**Example Use Cases**	**Number of Entities**
DBpedia[41]	Extracts structured information from Wikipedia	Question answering, entity recognition, NER	6.2 million
ConceptNet [42]	Multilingual knowledge graph of general knowledge and common-sense concepts in natural language	Semantic similarity, word sense disambiguation	6.3 million

4 Conclusion

In conclusion, VQA has become a popular research area in computer vision and natural language processing. The goal of VQA is to develop an intelligent system that can understand, and answer questions based on visual information. The use of semantic web technology in VQA can enhance the performance of these systems by providing a structured and comprehensive representation of the knowledge used to understand and answer questions. Semantic web technology can enable VQA systems to reason with large amounts of structured data from various sources, including ontologies and knowledge graphs. This can lead to better accuracy, more robustness, and more explainable results. Additionally, the integration of semantic web technology with VQA can facilitate the integration of VQA systems with other intelligent systems, such as chatbots, recommendation systems, and personal assistants, leading to more advanced applications.

Overall, the use of semantic web technology in VQA is a promising area for future research, as it can lead to more advanced, intelligent, and efficient VQA systems.

5 Reference

[1] Malinowski, M., & Fritz, M. (2020). A survey of the state of the art in VQA. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8), pp. 1888-1906.

[2] Hogan, A., Harth, A., & Polleres, A. (2020). Knowledge graphs. Foundations and Trends in Information Retrieval, 13(3-4), 239-389.

[3] Antropova, N., Li, X., Dehghani, M., Zhang, L., & Chen, D. (2020). Medical visual question answering with semantic segmentation. Medical Image Analysis, 59, 101559.

[4] Gao, J., Li, Y., Zhang, W., Chen, L., & Gao, X. (2021). Video question answering using spatio-temporal attention and semantic matching. IEEE Transactions on Multimedia, 23, pp. 2365-2377.

[5] Liu, L., Li, D., & Chen, L. (2021). Learning to reason with intermediate explanations for visual question answering. IEEE Transactions on Image Processing, 30, pp. 2769-2780.

[6] Wang, Y., Wang, X., Ji, R., & Wu, Y. (2020). A novel visual question answering framework based on multi-level attention mechanism. IEEE Transactions on Multimedia, 22, pp. 2550-2562.

[7] Xu, X., Liu, J., Chen, T., & Zhao, Q. (2021). Reinforcement learning for visual question answering with adaptive advantages. IEEE Transactions on Neural Networks and Learning Systems, 32, pp. 2673-2684.

[8] Yu, L., Zhang, W., Wang, J., & Yu, Y. (2020). Multi-modal fusion with recurrent neural networks for visual question answering. IEEE

Published by Arab Journals Platform,

Transactions on Multimedia, 22(12), pp. 3187-3199.

[9] Wang, J., Zhang, S., & Li, Y. (2021). Semantic visual attention for visual question answering. Neurocomputing, pp. 438, 413-420.

[10] Alam, M. A., Lee, C. H., & Kim, Y. H. (2020). Multimodal social media question answering using semantic matching and answer aggregation. Information Sciences, 514, pp. 191-205.

[11] Teney, D., Anderson, P., He, X., & Dauphin, Y. (2018). A bottom-up approach to visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3606-3615.

[12] Wu, Q., Wang, P., Wang, X., He, X., Zhu, W. (2022). Knowledge-Based VQA. In: Visual Question Answering. Advances in Computer Vision and Pattern Recognition. Springer, Singapore.

[13] Sruthy Manmadhan and Binsu C. Kovoor. 2020. Visual question answering: a state-of-the-art review. Artif. Intell. Rev. 53, 8 (Dec 2020), pp. 5705–5745.

[14] Li, Y., Li, W., Xu, C., Wu, J., Chen, E., & Yang, J. (2020). Dynamic graph attention for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6274-6283.

[15] Chen, H., Wang, H., & Wang, Z. (2020). Knowledge-augmented visual question answering via graph attention mechanism. IEEE Transactions on Image Processing, 30, pp. 3472-3482.

[16] Li, G., Qiu, Q., Li, D., Li, W., & Li, Z. (2021). Joint reasoning of knowledge graph and text for visual question answering. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 31, pp. 4179-4190.

[17] Lu, Y., Chen, X., & Qian, C. (2020). Bert-based spatial-temporal reasoning for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12573-12582.

[18] Kim, J., Jun, H., & Zhang, B. T. (2021). Visually grounded question answering with dynamic reasoning modules. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2339-2348.

[19] Srivastava, Yash, et al. "Visual question answering using deep learning: A survey and performance analysis." International Conference on Computer Vision and Image Processing. Springer, Singapore, 2020.

[20] Allaouzi, Imane, Mohamed Ben Ahmed, and Badr Benamrou. "An Encoder-Decoder Model for Visual Question Answering in the Medical Domain." CLEF (Working Notes). 2019.

[21] Wenfeng Zheng, Lirong Yin, Xiaobing Chen, Zhiyang Ma, Shan Liu, Bo Yang, Knowledge base graph embedding module design for Visual question answering model, Pattern Recognition, 2021, Volume 120,

[22] Henson, Cory Andrew et al. “Using a Knowledge Graph of Scenes to Enable Search of Autonomous Driving Data.” International Workshop on the Semantic Web (2019).

[23] Kim, S.; Jeon, T.H.; Rhiu, I.; Ahn, J.; Im, D.-H. Semantic Scene Graph Generation Using RDF Model and Deep Learning. Appl. Sci. 2021, 11, 826

[24] Li, J., Li, L., Li, Y., Su, Q., & Li, J. (2020). K-BERT: Enabling language representation with knowledge graph. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4329-4339.

[25] Bharti, S., Dey, L., & Mukherjee, D. (2020). Deep-SAE: A deep learning-based semantic autoencoder for visual question answering. Knowledge-Based Systems, 203, 106136.

[26] Zhang, J., Wu, H., & Liu, W. (2021). Multi-view graph convolutional network for visual question answering. Neurocomputing, 450, pp. 92-101.

[27] Liu, H., Zhu, X., Yu, N., & Zhou, J. (2021). Semantic Web-Integrated Residual Learning for Visual Question Answering. IEEE Transactions on Cybernetics, 51(6), pp. 3206-3217

[28] Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. 2023. VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1155–1165.

[29] Agrawal, A., et al. "VQA: Visual Question Answering." In Proceedings of the IEEE International Conference on Computer Vision, 2015.

[30] Ren, M., et al. "Exploring models and data for image question answering." In Advances in Neural Information Processing Systems, 2015.

[31] Hudson, D. A., et al. "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[32] Johnson, J., et al. "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[33] Wu, Q., Wang, P., Wang, X., He, X., Zhu, W. (2022). Classical Visual Question Answering. In: Visual Question Answering. Advances in Computer Vision and Pattern Recognition. Springer, Singapore.

[34] Zhou, Yu, et al. "VQA-Med 2020: Overview of the Medical Visual

Question Answering Task." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

[35] Kim, Wonmin, et al. "MedVQA: a large-scale medical visual question answering dataset." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020

[36] Singh, Aishwarya, et al. "KB-VQ: Question Answering on Knowledge Bases and Visual Data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[37] Shah, Meet, et al. "FVQA: Fact-based Visual Question Answering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017.

[38] Zhou, Yu, et al. "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[39] Kim, Kyungjae, et al. "KR-VQA: Korean Visual Question Answering Dataset." Proceedings of the 27th ACM International Conference on Multimedia. 2019.

[40] Visual question answering: Datasets, algorithms, and future challenges, Computer Vision and Image Understanding, 2017,Volume 163.

[41] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N. & Bizer, C. (2020). DBpedia—a large-scale, multilingual knowledge graph. Semantic Web, 11(3), pp. 447-470.

[42] Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In AAAI (Vol. 31, No. 1).