Artificial Intelligence Review (2022) 55:6277–6300
https://doi.org/10.1007/s10462-022-10151-2

An analysis of graph convolutional networks and recent datasets for visual question answering

Abdulganiyu Abdu Yusuf[1,4] · **Feng Chong**[1,2] · **Mao Xianling**[1,3]

Published online: 9 April 2022

Abstract
Graph neural network is a deep learning approach widely applied on structural and non-structural scenarios due to its substantial performance and interpretability recently. In a non-structural scenario, textual and visual research topics like visual question answering (VQA) are important, which need graph reasoning models. VQA aims to build a system that can answer related questions about given images as well as understand the underlying semantic meaning behind the image. The critical issues in VQA are to effectively extract the visual and textual features and subject both features into a common space. These issues have a great impact in handling goal-driven, reasoning, and scene classification subtasks. In the same vein, it is difficult to compare models' performance because most existing datasets do not group instances into meaningful categories. With the recent advances in graph-based models, lots of efforts have been devoted to solving the problems mentioned above. This study focuses on graph convolutional networks (GCN) studies and recent datasets for visual question answering tasks. Specifically, we reviewed current related studies on GCN for the VQA task. Also, 18 common and recent datasets for VQA are well studied, though not all of them are discussed at the same level of detail. A critical review of GCN, datasets and VQA challenges is further highlighted. Finally, this study will help researchers to choose a suitable dataset for a particular VQA subtask, identify VQA challenges, the pros and cons of its approaches, and improve more on GCN for the VQA.

Keywords Computer vision · NLP · VQA · GCN · Datasets

✉ Abdulganiyu Abdu Yusuf
abdulg720@gmail.com

Feng Chong
fengchong@bit.edu.cn

Mao Xianling
maoxl@bit.edu.cn

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

[2] South-East Information Technology Institute of Beijing Institute of Technology, Beijing, China

[3] Beijing Engineering Research Centre of High Volume Language Information Processing and Cloud Computing Application, Beijing, China

[4] National Biotechnology Development Agency, Abuja, Nigeria

1 Introduction

Graph neural network as deep learning model play a vital role in many research fields including natural language processing (NLP) (Yao et al. 2019), computer vision (CV) (Teney et al. 2017), and natural sciences (Do et al. 2019). There are several existing studies on deep learning methods that utilized graph data. The recent success recorded in image, text, and video processing largely depends on the increasing size of training data, graphics processor units (GPUs) enhancements, and typical representation in euclidean space. The goal of the VQA task is to answer natural language questions based on the corresponding information present in a given image. Since the inception of the VQA task in the year 2014, there exist several improvements and active research on the topic to date. These range from the benchmarking datasets (Antol et al. 2015; Gurari et al. 2018; Gupta et al. 2021), techniques (Cho et al.2014; Ren et al. 2016; Narasimhan et al. 2018), to evaluation metrics (Wu and Palmer 1994). The generic process of the VQA system is given in Fig. 1. It includes inputs, feature extraction, fusion, and answer prediction modules. The inputs to VQA systems are image and textual questions. Firstly, the features of both inputs are extracted using computer vision and natural language processing methods, to generate image and textual representation. Different varieties of convolutional neural network (CNN) architectures (He et al. 2016; Krizhevsky et al. 2012; Simonyan and Zisserman 2014; Anderson et al. 2018) can be used to extract semantic features of the input image.

The input question can be extracted using recurrent neural networks (RNNs) (Cho et al. 2014) and long short term memory (LSTM) (Sundermeyer et al. 2012). Secondly, the extracted image and question features are further fused via operations like direct concatenation, element-wise product, element-wise sum, bilinear pooling, and bayesian models to generate output vectors. Finally, the output vectors are passed through VQA prediction or generation model to infer answers according to input image and question.

The VQA models are commonly designed based on joint embedding and attention mechanism approaches. Yet, these approaches still face the challenge of detecting objects that are present in an image and justifying the relationships among the objects. For instance, an image may contain multiple objects and every word in a question has its own importance. Consequently, they still have issues with answering counting and

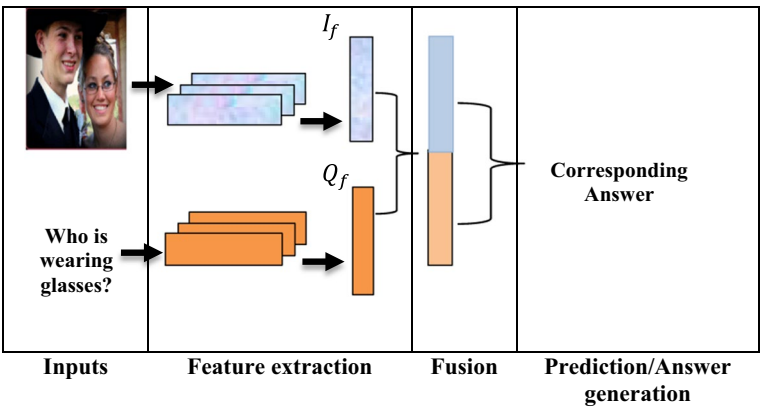


Fig. 1 Generic VQA process. Where I_f and Q_f represents the extracted image and question features.

reasoning questions. Therefore, a better image representation and question refinement are needed to improve the model quality thereby answering scene classification, goal-driven, and reasoning subtasks. Recently, graph-based approaches (Yao et al. 2019; Hu et al. 2020; Zhu et al. 2021) have been proposed to address the aforementioned challenges. They can capture the node features and encode the neighborhood properties among nodes provided in the graphs. This is essential for integrating semantic and spatial information needed for VQA and other vision-to-language tasks.

There are significant number of review papers on GCN, most of which discuss taxonomy, models, and applications in various domains (Wu et al. 2020a, b; Zhang et al. 2019b) for future research. This paper reviewed the most recent GCN work and datasets for the VQA task. The statistics of the recent VQA datasets are compared and categorized according to various features. We further provide a critical review for GCN and datasets. Finally, the VQA challenges were briefed.

The contributions are summarized as follows:

- To the best of the authors' knowledge, this study is among the first surveys that focus on GCN for the VQA task.
- We reviewed the most recent VQA benchmark datasets from the year 2016 to 2020. Detailed comparisons of the datasets are made based on their uniqueness. Also, key features of the datasets are categorized according to the number of images, number of questions, image type, annotation type, answer type, average number of questions per image, training, validation, testing percentage, and subtask.
- This study will help young researchers in the field to identify still existing VQA challenges and approaches with their strengths and weaknesses.

The rest of the study is organized as follows: In Sect. 2, we discussed related studies on VQA and GCN. Section 3 highlighted recent VQA datasets that can be considered for rigorous benchmarking. These datasets are standard and contain challenging problems that can be applied in a real-world scenario. A critical review of the techniques, datasets and VQA challenges is given in Sect. 4. The study is concluded in Sect. 6 with remarks for further studies.

2 Related studies

2.1 Visual question answering

Since the emergence of the VQA dataset towards the end of 2014, there have been several active research works on the topic to date. According to Singh et al. (2019a), VQA methods are categorized into joint embedding methods, attention mechanisms, and compositional models. This study categorized VQA approaches into joint embedding, attention mechanism, and compositional, external knowledge, and graph-based approaches.

In joint embedding approaches, text features can be extracted using bag-of-words (BOW) or long short term memory (LSTM) (Ben-Younes et al. 2017; Fukui et al. 2016; Shih et al. 2016). Then, CNN can be used to extract image features. The respective features are further combined into common feature space using either concatenation or element-wise multiplication methods. Finally, the combined feature vector is passed into a

classifier to predict an answer to the input question. The joint embedding methods focus on the entire region of the image, which poses a challenge of understanding question specific semantic information of an image. Some of the early research in VQA concentrate on joint embedding approaches and they are considered as common practice in vision and language research communities (Ben-Younes et al. 2017; Fukui et al. 2016; Shih et al. 2016). Although these approaches are mostly used for open-end questions and multiple choice answers, yet they can only generate answers that are observed during training.

The attention mechanism based approaches improve further by considering part of the input space (Wang et al. 2018a; Shah et al. 2019). They extract the global features of an image to answer a question focusing on only a specific region of the image that might confuse the VQA system. For instance, given the question 'what size is the book', the salient part of the image contains the book. Also, 'size' and 'book' are the most relevant words in the sentence. The attention mechanism is guided by an algorithm that represents feature vectors corresponding to each region in an image at a more local level which are then ranked based on their similarity with the features of the question asked (Kallooriyakath et al. 2020). The global image features (like the last hidden layer of a CNN) and global text features (bag-of-words) may not be capable of addressing region specific questions. More promising results were obtained on the attention-based VQA methods using benchmark datasets. But, they still have issues with complex questions that contain reasoning and counting.

In compositional models, questions involve a series of reasoning steps to infer proper answers. For instance, a question like 'what is beside the cup' entails finding the cup and naming the object beside it. These approaches are mostly restricted to visual reasoning only. In these approaches, two compositional systems have been proposed in an attempt to solve VQA in a series of sub-steps. The first framework is Neural Module Network (NMN), and the second structure is Recurrent Answering Units (RAU). The NMN structure utilizes external question parsers to find the subtask in the question, whereas RAU is prepared end-to-end, and subtask can be implicitly learned (Andreas et al. 2016; Noh et al. 2016).

In more recent scenarios, the aforementioned VQA approaches underperform on more complex questions like 'why is the baby smiling' and 'can we get water here'. These types of questions require common sense reasoning and knowledge about the spatial relationship among objects. Consequently, there is a need for robust VQA systems capable of solving more natural questions. Wu et al. (2016) proposed an external knowledge base VQA approach. This approach is useful when some common sense or additional background knowledge is required to answer questions correctly. The advantage of this method is that it answer more general questions and understand the reasoning behind how it arrives at the answer by viewing supporting facts generated in the process. But, lack of more precise semantic or visual attention applied to the question and image before querying the knowledge bases is a disadvantage.

Recently, most external knowledge or attention-based VQA approaches do not adequately examine or utilize how the words in the question interact with each other. Hence, there are still issues in answering complex or reasoning questions. The evolution of graph-based approaches (Teney et al. 2017; Kipf and Welling 2017; Narasimhan et al. 2018; Zhu et al. 2021) create new opportunities to overcome still existing challenges in VQA, with better performance on standard datasets. Table 1 summarizes the VQA approaches with respect to representative works as well as pros and cons.

Table 1 Visual question answering approaches

Approaches	Some representative works	Pros	Cons
Joint embedding	Multi-modal Compact Bilinear (MCB) (Fukui et al. 2016)	Can be used for open-end questions and multiple choice answers	Generated answers are restricted to words observed in the course of training
	Multi-modal Low-rank Bilinear pooling (MLB) (Kim et al. 2017)	Compared to MCB, MLB has less computational complexity when used with a neural network of few parameters	Focus on the global features of the image
			MCB can be computationally expensive
Attention mechanism	Stacked Attention Network (Yang et al. 2016)	Focus on the salient region of the image to produce a more accurate prediction	Have issues with natural questions that contain reasoning and counting
	Focused Dynamic Attention (Ilievski et al. 2016)		
	Object-difference attention (Wu et al. 2018)	Filter noise	
Compositional models	Dynamic Memory Networks (DMN) (Kumar et al. 2016)	Can answer reasoning questions	Mainly restricted to visual reasoning only
	Neural Module Networks (NMN) (Andreas et al. 2016)		
	Recurrent Answering Units (RAU) (Noh et al. 2016)		
External knowledge	Wu et al. (2016)	It answers more general questions and understands the reasoning behind how to get the final answer	Lack of more precise semantic or visual attention applied to the question and image before querying the knowledgebase is a disadvantage
Graph-based	Graph Neural Network (GNN) (Teney et al. 2017)	Can extract the relationship between the objects and words within the question	It is difficult to model compared to CNN and RNN networks
	Graph Convolutional Network (GCN) (Narasimhan et al. 2018)		

2.2 Graph neural networks for visual question answering

The graph neural networks receive attention significantly due to the ability to capture spatial and semantic relationships. In contrast to other VQA methods, GNN uses deep neural networks and dimensionality reduction techniques to encode the structural information about a graph. Moreover, the encoded information is used to carry out clustering or classification tasks. From the literature, GNN can be grouped into graph auto-encoders, recurrent graph neural networks, spatial–temporal graph neural networks, and convolutional graph neural networks (Wu et al. 2020b). This study focuses on GCN as it is a quite new and powerful network architecture for generalizing CNN to graph structured data. The application of GCN in various domains is given in Table 2.

Graph convolutional network was introduced by Kipf and Welling (2017) as an accessible approach for semi-supervised learning to embed hidden features of graph nodes and edges. It is newly applied for textual and visual understanding. For example, it is used to build zero-shot recognition (Wang et al. 2018b), to learn binary codes for image retrieval (Zhou et al. 2020), and for scene graph generation to capture contextual relations between objects (Yang et al. 2018). In the VQA task, Teney et al. (2017) proposed a graph-based approach that used graph neural networks to combine abstract images and graph

Table 2 Applications of GCN in various domains (Asif et al. 2021)

Domain	Application	Some representative work
Computer vision	Visual question answering	Narasimhan et al. (2018)
	Action recognition	Guo et al. (2018)
Natural language processing	Machine translation	Zhang et al. (2018a)
	Event detection	Nguyen et al. (2018)
	Question answering	
	Relation extraction	
Science (Biology and Chemistry)	Retro synthesis prediction	Zitnik et al. (2018)
	Modeling polypharmacy side effects	Dai et al. (2019)
	Exploiting relationships in molecules	
Recommendation system	Collaborative filtering	Chen et al. (2020)
	Recommender systems	Zhang et al. (2019a)
	Matrix completion	
Social network	Rumor detection	Bian et al. (2020)
	Fine-grained event categorization	Wu et al. (2020a)
	Social spammer detection	
Security analysis	Malware detection	Wang et al. (2020)
	Data management and anomaly detection	Pei et al. (2020)
	Effective vulnerability identification	
Traffic network	Traffic flow forecasting	Guo et al. (2019)
	Passenger demand modelling	Wang et al. (2019)
	High-resolution routing	
Knowledge-based	Knowledge graph alignment	Schlichtkrull et al. (2018)
	Modeling relation data	Wang et al. (2018c)
	Logic reasoning	

representations of questions. The model achieved significant results compared to state-of-the-art, indicating the capability of the graph-based approach for VQA. Nevertheless, their method is not easily applicable to real-world images in a situation where scene graph representation is not known earlier.

Narasimhan et al. (2018) introduced factual visual question answering using GCN. They reported that a large set of curated facts are given to output with two likely answers through relation. For a given pair of question-image, deep network techniques were employed to continually reduce the large set of facts until one of the two final entities remains a fact and is generated as an answer. However, there is a need to incorporate steps into a unified framework since the fact retrieval component is not trainable end-to-end.

To mitigate counting problems in VQA, an explicit and implicit graph structure need to be captured to model the interaction between objects. To further the research, high-level semantic details like visual relationships and attributes were studied (Singh et al. 2019a; Shah et al. 2019) to make the model more robust and understandable. Norcliffe-Brown et al. (2018) used graph convolutions to learn image representations and capture interactions of a particular question. The accuracy of up to 66.18% was obtained on the VQA dataset. This shows that the model performance extremely depends on the quality of the detector, resulting in duplicates or missing objects. However, to eliminate duplicate object detections, Zhang et al. (2018b) used to model interactions amongst images via implicit and explicit graph structures focusing on counting tasks. Zhang et al. (2018b) computed a graph based on the outside product of the attention weights of proposed features. The computed graph is transformed explicitly to improve the counting ability of their baseline models'. In the same vein, Trott et al. (2018) employ an iterative method that relies on the similarities of the images to enhance the models' counting capabilities and interpretability.

Another development was made by Cadene et al. (2019) to further model spatial and semantic relations within all pairs of regions for relation reasoning. Similarly, a knowledge-enabled VQA model that can read and reason was also proposed by Singh et al. (2019a). They combined visual cues, textual cues, and rich knowledge bases and performed reasoning using a novel gated graphical neural network formulation. Although the result obtained outperformed traditional VQA models, but are restricted to questions that can be answered from one-hop reasoning on the knowledge graph. In general, knowledge-enabled VQA models mutually embed entire information without fine-tuned selection, which introduces unpredicted noises for reasoning the correct answer (Yu et al. 2020).

In addition, Yang et al. (2019) proposed relational reasoning in visual question answering by utilizing prior visual relationship learning. The proposed scene graph convolution operation used information from objects and relationships to update each node in hidden states. A fine-grained question representation (FQ-GCN) is proposed by Hu et al. (2020) using GCN. They constructed an object relation graph and fine-grained question features that are used to find the relations between objects in an image, pruning unconnected edges between object nodes. Graph convolutional network is used to collect all neighborhood details of each object in an object graph. It was found that there is a need to improve similar relations between words in a particular question and entities in the image to build a more accurate graph network.

Kim et al. (2020) employed hypergraph attention networks (HAN) for textual and visual learning to convert the low-level multi-modal inputs into symbolic graph forms that incorporate the multiple symbolic graphs with the co-attention maps. On scene graph generation, it was found that authors could not achieve better improvement on the GQA dataset when compared to some other similar studies. Xu et al. (2020) proposed an answer centric visual question generation (VQG) method named radial-GCN, in which image analysis

emphasizes the most relevant image regions only. Then, a novel sparse graph of the radial structure is naturally built to obtain the relations between likely answer regions and other regions. The graphic attention is later used to direct the convolutional propagation to more relevant nodes for final question generation. Their method needs to accommodate more challenging and detailed questions. Also, they can explore more evaluation metrics to evaluate the strength of the generated questions for the VQG task.

Gao et al. (2020a, b) introduced a multi-modal graph neural network (MM-GNN) for VQA on-scene texts. The MM-GNN represents the image with multi-modal contents as a configuration of three graphs, representing each graph as one modality. Furthermore, the designed multi-modal aggregators in MM-GNN uses multi-modal contents to get a refined representation of objects in the image, mainly for unknown or rare words. Zhu et al. (2021) proposed a GCN method that can directly model objects with their relationships without any earlier knowledge or pre-training, and the question-adaptive object relation graph can be learned end-to-end. A description of some GCN studies for the VQA task is given in Table 3.

3 Some recent VQA datasets

There are numerous datasets available to researchers for the VQA task since their inception in the year 2014. These datasets contain at least a triplet of an image, a question, and a corresponding answer. They are curated by researchers for a specific need and are varied in size. Some of the early research reported that VQA datasets are like DAQUAR (Malinowski et al. 2015), COCO-QA (Ren et al. 2015), and VQA (Antol et al. 2015) contains a lot of useful information for benchmarking that is publically available online. This section explores the most recent datasets for VQA. The key features of the datasets are discussed and compared based on their uniqueness and similarities. These datasets are prominent and contain several variations of subtasks to test the efficiency of the VQA techniques. Several characteristics such as subtask to handle, annotation type, and image types, among others, were captured in the study. They are also free online and can be used to perform rigorous benchmarking for VQA tasks. Some of the selected datasets are briefly explained below.

3.1 Visual question answering version 2.0 (VQA 2.0)

Visual question Answering version 2.0 is fully released in 2017, containing open-ended and multiple choice questions about given images. These questions require an understanding of vision, language, and common sense knowledge to produce a correct answer (Goyal et al. 2017). Recently, it is regarded as one of the most commonly used benchmark datasets for evaluating VQA methods that contain 265,016 images (COCO and abstract scenes), 1,105,904 questions, and 11,059,040 answers (10 ground-truth answers per question). In VQA 2.0 dataset, each sample of the image contains questions answered by ten independent annotators. The images in VQA 2.0 are divided into training (40.1%), validation (19.4%), and test sets (40.5%). Compared to other previous versions, VQA 2.0 has a large number of images for training, validation, and testing.

Table 3 Comparison of GCN studies for VQA task

Authors	Techniques	Datasets	Results (%)	Settings	Remarks
Teney et al. (2017)	GNN	Abstract scenes	74.37	Multiple choice	The method is not easily applicable to real-world images
			70.42	Open-ended	
Narasimhan et al. (2018)	GCN	FVQA	72.97	Open-ended	The fact retrieval component is not trainable end-to-end
Norcliffe-Brown et al. (2018)	GCN	VQA 2.0	66.18	Open-ended	The models' performance largely depends on the quality of the detector, resulting in duplicates or missing objects
Singh et al. (2019a)	GNN	Text-KVQA	71.30	Open-ended	The approach is restricted to questions that can be answered from one-hop reasoning on the knowledge graph
Yang et al. (2019)	SceneGCN	GQA	54.56	Open-ended	The model trained and evaluated on VQA 2.0 gained less improvement compared to the one on GQA
		VQA 2.0	67.14		
Hu et al. (2020)	FQ-GCN	VQA 2.0	65.49	Open-ended	There is a need to improve similar relations between words in a particular question and entities in the image to build a more accurate graph network
Kim et al. (2020)	GNN + HAN	GQA	61.88	Open-ended	On scene graph generation, authors could not achieve better improvement on the GQA dataset when compared to some other similar studies
		VQA 2.0	65.05		
Xu et al. (2020)	Radial-GCN	VQA 2.0	57.20	Multiple choice	Their method needs to accommodate more challenging and detailed questions. Also, they can explore more evaluation metrics to evaluate the strength of the generated questions for the VQG task
		Visual 7w	52.70		
Gao et al. (2020b)	MM-GNN	TextVQA,	31.10	Open-ended	The result obtained show that the new representation of image and message passing schemes significantly improve the VQA performance on both datasets
		ST-VQA	16.00		
Zhu et al. (2021)	GCN	VQA 2.0	66.67	Open-ended	There is a need to explore more on the use of detailed relation types for a particular question reasoning

3.2 Fact-based visual question answering (FVQA)

Fact-Based Visual Question Answering (FVQA) contains 2190 images, 5286 questions, and 4126 unique facts equivalent to the questions (Wang et al. 2018a). The knowledge base consists of 193,449 facts, which were captured extracting top visual concepts for all images in the dataset and querying for those concepts in the knowledge bases, such as WebChild (Tandon et al. 2014), ConceptNet (Speer et al. 2017), and DBPedia (Auer et al. 2007). The three types of visual concepts extracted in FVQA are action, scene, and object. The answers in this dataset are limited to image concepts and knowledge bases, which exclude Yes/No questions.

3.3 Graph question answering (GQA)

GQA is one of the recent VQA datasets that is designed to leverage semantic representations of both scene and questions (Hudson et al. 2019). The dataset contains about 20 million structured representations of questions with over 110 thousand real-world images. The questions in GQA involve multiple reasoning skills, spatial understanding, and multistep inference. Each image is linked with a scene graph of objects, attributes, and predicates. A total number of 1740 objects, 620 attributes, and 330 predicate labels are defined as a semantic ontology for GQA. Each image contains 16.4 distinct objects, and each object has 0.54 attributes with 3.08 relationships on average. The dataset is divided roughly into training (87%), validation (12%), and test sets (1%).

3.4 TextVQA

TextVQA is also a new VQA dataset released in the year 2019 (Singh et al. 2019b). The dataset needs a model to read and reason about the text in an image to infer answers based on related questions. This contains 45,336 real-world questions, 28,408 images from Open-Images and 453,360 ground-truth answers. The images in TextVQA are divided into training (77.3%), validation (11.1%), and test sets (11.6%). Each question–answer pairs contain a list of tokens generated by an Object Character Recognition (OCR) model. The OCR approach is called Look, Read, Reason, and Answer (LoRRA).

3.5 Visual question answering 360° (VQA 360°)

Chou et al. (2020) introduced a VQA 360 image that captures the entire visual content around the optical center of a camera, needing more reasoning and spatial understanding. The dataset contains 1490 images and 16,945 question-answer pairs, which are split into training (50%), validation (10%), and for test sets (40%). Each image consists of 11 questions divided into five different question types: counting, exist, property, scene, and spatial. All the images in VQA 360° are stored in equirectangular format and resized to 1024 by 512. Yet, there is still about a 25% performance gap between humans and machines. The gap is higher mainly on counting, property, and spatial types, signifying the directions to improve algorithms to match human inference abilities.

3.6 VizWiz

VizWiz is the first goal-driven VQA dataset that originates from a real-world setting (Gurari et al. 2018). It consists of over 31,000 visual questions originating from blind people who each took a picture using a mobile phone and recorded a spoken question about it, together with ten crowd sourced answers per visual question. The dataset is divided into training (65%), validation (10%), and test sets (35%). VizWiz differs from the many existing VQA datasets (Gurari et al. 2018) because:

- Images are captured by blind photographers and so are often poor quality,
- Questions are spoken, and so are more conversational, and
- Usually, visual questions cannot be answered.

For these reasons, existing algorithms find it difficult to evaluate the VizWiz dataset.

The figure below gave the trend of some common and recent VQA datasets with the corresponding year of creation. Subsequently, a comparison of the datasets is given in one representation (Fig. 2).

The visual question answering 360° is the first VQA dataset that captures the entire visual contents of the image in 360°. It can be used for VQA research that contains reasoning, spatial understanding, and counting tasks. VQA 2.0 is an open-ended dataset for common sense reasoning. It is large and commonly used to perform VQA benchmarking because it has a large number of images for testing, validation, and training as compared to most VQA datasets. In another development, compositional language and elementary visual reasoning (CLEVR) (Johnson et al. 2017) and GQA were introduced to robustly generate answers to questions by performing explicit multistep reasoning on an image. The questions in these datasets demand complex and compositional reasoning aiming at mitigating answer biases and identifying the reasoning ability of VQA models.

Visual7W dataset is used for multiple choice VQA tasks (Zhu et al. 2016). In contrast to Visual7w and CLEVR, OK-VQA (Marino et al. 2019) emphasizes questions that cannot be answered by the corresponding information in the image and require external knowledge to be answered. Similar to FVQA, it requires outside knowledge, but FVQA uses the knowledge triplet to annotate questions from a fixed knowledge base. At the same time, OK-VQA extracts relevant facts from the databases, internet, and some other knowledge sources that were not used to create the questions. Optical character recognition (OCR-VQA) contains images of book covers used for reading text in the image (Mishra et al. 2019). Apart from visual question answering, OCR-VQA can also be used for document image analysis.

Both TextVQA and scene text visual question answering (ST-VQA) datasets are conceptually similar (Biten et al. 2019), which involve reading and reasoning about scene text.

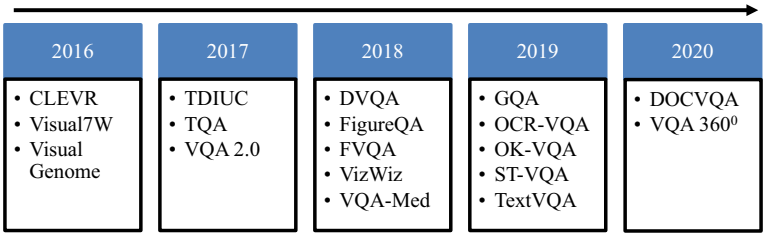


Fig. 2 Some recent VQA datasets

In ST-VQA, images were obtained from multiple sources, unlike from a single source in TextVQA. Scene text visual question answering dataset focus directly on answered questions using part of the image text, while in TextVQA, any question that requires reading the image text is allowed. Task Driven Image Understanding Challenge (TDIUC) dataset (Kafle and Kanan 2017) addresses VQA limitations such as difficulty in the evaluation process and unbalanced questions type.

Both FigureQA (Kahou et al. 2018) and data visualizations via question answering (DVQA) (Kafle et al. 2018) are commonly used VQA datasets on figures and graphical plots. In contrast, DVQA is developed with FigureQA, but it contains three types of questions while FigureQA contains 15 question types. VizWiz is the first goal-oriented VQA dataset originated from blind people (Gurari et al. 2018), consisting of daily visual questions from blind people in quest of corresponding answers. However, it suffers from limitations like most questions cannot be answered, poor questions recording, and image quality. Textbook Question Answering (TQA) (Kembhavi et al. 2017) is also similar to DVQA but, it does not contain questions specific to bar charts and uses a multiple choice scheme to reduce the ranking problem.

VQA-Med is the first VQA dataset in the medical field, consisting of medical images with clinical related question–answer pairs (Gupta et al. 2021). Compared to other VQA datasets, VQA-Med is very small for evaluating VQA models. Compared to other VQA datasets like ST-VQA, FigureQA, TQA, and TextVQA, DocVQA (Mathew et al. 2021) covers several document types as well as different textual, graphical, and structural features. The characteristics of the most recent datasets are summarized in Table 4.

The table characterized the VQA datasets based on the type of image, answer format, the average length of questions, average length of answers, annotation type, and specific tasks to handle for VQA. Firstly, the images are classified into natural (real), and synthetic (clip art or cartoon). The majority of the VQA benchmark datasets like Visual7w, FVQA, Visual genome (Krishna et al. 2017), and VQA 360° uses natural images. The second characteristic is the answer format which is classified into multiple choice and open-ended settings. The answers in an open-ended setting are more difficult to evaluate because they are in free form. In multiple choice setting, a limited number of candidate answers are provided as options to select a single correct answer. The results obtained from multiple choice settings are more interpretable and easy to evaluate than these of open-ended setting. Although the result obtained on both settings differs, datasets like VQA 2.0 and VQA-Med evaluate answers in either of the settings.

Thirdly, the datasets can also be categorized according to the length of the input questions and the length of the output answers. The dataset like GQA and FigureQA provides more lengthy answers, which are more acceptable in human form. Fourthly, the annotation procedure of any VQA dataset is collected in either manual or automatic form. For example, annotations provided include image caption, candidate answers in multiple choice settings, and bounding box image regions to support answers. The dataset such as VizWiz, DVQA, and ST-VQA provides annotation in both forms. Lastly, the datasets are curated for evaluating specific aspects of VQA tasks. Some datasets like FVQA, GQA, and CLEVR require more reasoning ability to infer the right answers. Conversely, ST-VQA, TextVQA, and DocVQA are meant for scene classification. In "Appendix 2", the pair samples of VQA datasets, consisting of images, questions, and corresponding answers are depicted. As well, the link to the dataset can be found in "Appendix 1".

Table 4 Characteristics of some recent VQA datasets

Dataset	Number of		Type of		Average length of		Quest. Cat	N	(Tr, V, Te)%	Annotation	Subtask
	Images	Questions	Image	Answer	Answer	Question					
Visual7w	47,300	327,939	Natural	MC	2.0	6.9	7	7	50, 20, 30	Manual	Scene classification
CLEVR	70,000	999,968	Synthetic	OE	1.1	4	5	–	70, 15, 15	Automatic	Visual reasoning
Visual Genome	108,000	1,773,258	Natural	OE	4	17	6	6	80, 10, 10	Manual	Scene classification Reasoning
TDIUC	167,437	1,654,167	Natural	OE	–	3	12	3	70, 15,15	Both	Scene classification Reasoning
TQA	3455	26,260	Natural	MC	4.5	8	7	5	58, 20, 20	Manual	Textual and Visual reasoning
VQA 2.0	265,016	1,105,904	Natural, Synthetic	OE/ MC	1.2	6.1	–	3	40, 19, 41	Manual	Common sense reasoning
FigureQA	100,000	1,300,000	Synthetic	MC	1	10	15	5	62, 19, 19	Manual	Visual reasoning
FVQA	2190	5826	Natural	OE	1.2	9.5	32	3	51, –, 49	Automatic	Knowledge based reasoning
VizWiz	3000	31,173	Natural	OE	1.66	2	8	10	65, 10, 25	Both	Goal-driven
DVQA	300,000	3,487,194	Synthetic	OE	3	9.5	3	2	–	Both	Scene classification Reasoning
VQA-Med	4200	15,292	Natural	OE/MC	–	7	11	4	–	Automatic	Goal-driven
OCR-VQA	207,572	1,002,146	Natural	OE	3.31	4.83	3	6	80, 10, 10	Manual	Scene classification
OK-VQA	14,031	14,055	Natural	OE	1.3	8.1	10	5	–	Manual	Knowledge based reasoning
ST-VQA	23,038	31,791	Natural	OE	1.5	5.6	5	–	83, 4, 13	Both	Scene classification
GQA	113,018	22,669,678	Natural	OE	1.1	10	10	4	70, 10, 10	Manual	Visual Reasoning
TextVQA	28,408	45,336	Natural	OE	1.7	7.18	15	2	76, 11, 13	Manual	Textual Reasoning
VQA 360°	1490	16, 945	Natural	OE	2	–	5	11	50, 10, 40	Manual	Scene classification Reasoning
DocVQA	12,767	50,000	Natural	MC	2.17	8.12	9	3	80, 10, 10	Automatic	Scene Text classification

N Average number of questions per image, *Tr* Training, *V* Validation, *Te* Test sets, *MC* multiple choice; *OE* Open-ended-: not clearly given

4 Critical review and VQA challenges

4.1 Critical review

4.1.1 Datasets

There are several benchmarking datasets available to facilitate VQA research. The images present in these datasets are either natural or synthetic, and the answers are either in open-ended or multiple choice settings. In a multiple choice setting, questions contain a single correct answer, which is straightforward. There can be multiple correct answers in open-ended questions because questions are in synonyms and phrases form. A natural image VQA dataset is more applicable to real-world scenarios, some of them with manually labeled ground truth. Although manual annotation is labor intensive, especially for large VQA datasets, a well-trained manual annotator is better in terms of precision and quality of the training set. From the literature, only a few VQA datasets like VizWiz and VQA-med are goal-oriented. These types of datasets help models to extend to real-world applications. In summary, a benchmark VQA dataset must have the following criteria (Manmadhan and Kovoor, 2020):

- It should be sufficient to capture question and image concepts in natural settings.
- Should have a fair evaluation scheme to validate different VQA models.
- It should be minimally biased.

4.1.2 Techniques

In recent years, there has been an increase in research on VQA tasks, spanning language and vision. Most of the research is designed based on CNN, RNN, and LSTM. Although significant progress has been recorded using these methods on many standard datasets, they mostly underperform in capturing the semantic details of an image. In addition, they do not resolve all the problems especially handling reasoning questions. To further on that, few studies focus on higher level image representation, which is capable of capturing semantic and spatial relationships. Lately, graph networks have received a lot of consideration because they have high expressive power on structural feature learning. The results obtained show a competitive or even better performance. However, GCN are faced with the following challenges: the existing models may generate features that can be redundant and not optimal for a specific question. Also, GCN fails to model actual semantic relationships in specified questions, which leads to noisy information (Zhu et al. 2021).

In general, scene classification, reasoning, and goal-driven subtasks are the recent and most common VQA problems being researched. In scene classification subtask, scene text are usually assigned to images that are highly related in terms of visual appearance like product type, plant species and license plate number. It deals with questions like ‘what is in the image’, ‘is it raining’, and ‘what is the color of the train’. The goal-driven subtask require VQA system to answer questions about images which has real time applications for visually impaired persons and digital personal assistant. Yet, most current VQA datasets collect questions in a non-goal oriented setting (the questions are less conversational) making it not suitable for training data in real time application. The reasoning subtask involves more complex questions to information that are not present in the image. Therefore, it is

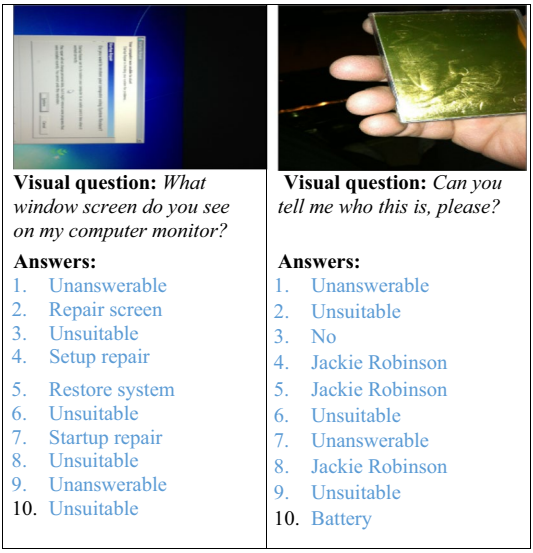


Fig. 3 Sample of images obtained from VizWiz dataset

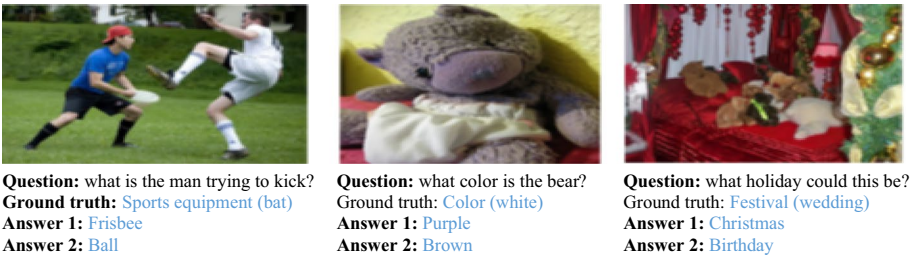


Fig. 4 Questions with different but similar answers (Gao et al. 2020a)

centered on knowledge based, common sense, and visual concepts. In this subtask, questions like 'why is she smiling' and 'is it rainy season' can be posed which needs ability to reason about the given images.

4.2 Visual question answering challenges

The studies on VQA continue to attain new feat in terms of accuracy, efficiency, reduced complexity, and computational time. Yet, there are many research challenges due to its broad research area and varied application in a real-world scenario. The VQA challenges identified in this study are given below.

4.2.1 Image quality and size

The current VQA techniques can reasonably identify objects and object detection questions when the input image and size are visible to some extent. But, if the image quality is too low or the objects present in the image are too small, then the existing VQA

Question: Which object in this image is used for protecting the head?

Ground truth: helmet

Question: which part of the machine in the image can be used for typing?

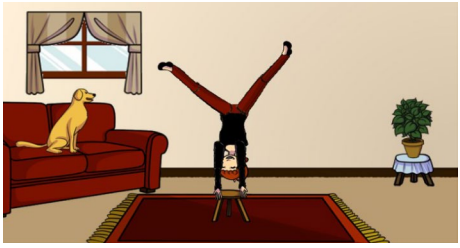
Ground truth: keyboard

Question: Where can you find the right object on the table shown in the image?

Ground truth: wedding

Fig. 5 Sample of images obtained from FVQA dataset

Fig. 6 Sample of an image from VQA dataset



Question: Why would you say this woman is strong?

Ground truth: yes

system could not present the correct answer. For instance, the VizWiz dataset (Gurari et al. 2018) is challenging and can be deployed to real-world applications. Figure 3 shows a sample of images obtained from VizWiz where some questions are not answerable due to poor image quality.

4.2.2 Inconsistent answers

In this situation, VQA techniques are faced difficulties with questions containing similar answers. For example, the questions can be answered with different words which are related to ground-truth answers but not precisely the same. In Fig. 4, different answers were obtained using different techniques.

4.2.3 Reasoning type of questions

The existing VQA systems are challenging to evaluate common sense and knowledge base reasoning questions due to their level of complexity. These involve questions like 'why is the boy running', 'is this a mosque or a church', 'what is the purpose of umbrella', and 'what is between the dog and the door'. To evaluate these types of questions, the VQA system requires additional knowledge from external sources and background knowledge to support facts. Hence, reasoning questions cannot be evaluated

Question: What number is on the train?

Ground truth: 7907

Fig. 7 Sample of an image obtained from VQA dataset

when the background information of the image is not given. The VQA tasks in Fig. 5 depend on external knowledge to some extent.

4.2.4 Multiple words in answers

In the VQA task, questions that require answers in phrases or sentences pose challenges during evaluation. From the plethora of VQA datasets, only a few techniques can generate descriptive or explanatory answers. Most of the answers are generated in single words or phrases. Yet, the few techniques cannot infer answers in full sentences or proper human acceptable form. An instance of this challenge is depicted in Fig. 6, where the ground truth answer is 'yes', which is not correct for the given question. The human annotator answers ('yes', 'can lift up', 'on arms', 'headstand', 'handstand', 'can stand on her head', 'she is standing upside down on stool') are not reliable. Therefore, such kinds of questions need wordy explanation.

4.2.5 Answering number specific question

Most of the recent VQA techniques are unable to answer question-related numbers displayed in the image. In contrast to the counting problem, an input image may contain various objects, including numbers. How to identify these numbers correctly is an issue to the techniques. The counting problem requires techniques to count entire objects or part of the objects present in the image. An example of number specific question challenge is shown in Fig. 7 below.

5 Conclusion and further studies

Recently, numerous research works have been reported on VQA to bridge the modality gap between textual and visual data by capturing spatial and semantics relationships among objects.

This study reviewed graph convolution networks and the most recent datasets for visual question answering. Firstly, a brief description of VQA approaches and their generic

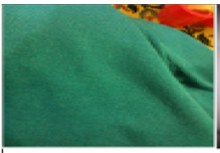
evaluation process were highlighted. Secondly, GCN is a quite new and powerful network architecture for generalizing CNN to graph structured data. Related works, discussion, and real-time application of GCN in various domains were given. To further the study, benchmarking datasets are required to train and test the performance of VQA system. A short description and comparison of a few selected datasets were also given. A critical review of the related datasets, approaches and VQA challenges are further provided.

In future study, we will examine how to improve the performance of VQA results using graph-based approach. Graph convolutional networks are capable of handling high-level semantic information and visual relationship, making the VQA process more interpretable and robust. The study will focus on solving Sect. 4.2 (4.2.2) and (4.2.3) challenges using GCN to handle reasoning and goal-driven subtasks. Finally, the important roles of hyperparameters in the performances of convolutional networks have been shown in some image analysis applications (Goceri 2019). Therefore, a comparative performance evaluation of GCN with different hyperparameters can also serve as future research.

Appendix 1

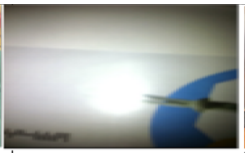
Datasets	Link
VISUAL7W	www.ai.stanford.edu/~yukez/visual7w/
CLEVR	www.cs.stanford.edu/people/jcjohns/clevr
VISUAL GENOME	https://visualgenome.org
TDIUC	www.goo.gl/Ng9ix4
TQA	www.textbookqa.org
VQA 2.0	www.visualqa.org
FIGUREQA	www.datasets.maluuba.com/FigureQA
FVQA	www.metatext.io/datasets/fact-based-visual-question-answering-(fvqa)
VIZWIZ	www.vizwiz.org/tasks-and-datasets/vqa/
DVQA	www.textbookqa.org
VQA-MED	www.imageclef.org/2021/medical/vqa
OCR-VQA	www.ocr-vqa.github.io/
OK-VQA	www.okvqa.allenai.org
ST-VQA	www.rrc.cvc.uab.es/?ch=11
GQA	www.cs.stanford.edu/people/dorarad/gqa//about/html
TEXTVQA	www.textvqa.org/
VQA 3600	www.aliensunmin.github.io/project/360-VQA/
DOCVQA	www.docvqa.org

Appendix 2



Q: what color is this?

A: green



Q: what is this mail for?

A: unanswerable

VIZWIZ

Q: where is the child sitting?

A: fridge

Q: is the umbrella upside down?

A: yes

VQA v2.0

Q: which brand are the crayons?

A: crayola

Q: what is the license number?

A: cu58 ckk

TEXTVQA

Q: Where can i find the bed?

A: at your left side

Q: What room is depicted in the image?

A: hallway

VQA 360^0

Q: what century is this?

A: 20th

Q: where is this monument located?

A: washington dc

OK-VQA

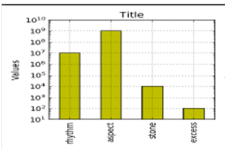
Q: what is the price of bananas per kg?

A: $11.98

Q: What does the red sign say?

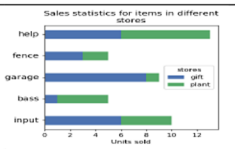
A: Stop

ST-VQA



Q: What is the value of the largest bar?

A: 10^9



Q: Does the chart contain stacked bars?

A: yes

DVQA

Q: what is the price of bananas per kg?

A: $11.98

Q: What does the red sign say?

A: Stop

ST-VQA

Q: What can the red object on the ground be used for?

A: firefighting

Q: Which object in this image is used for protecting the head?

A: helmet

FVQA

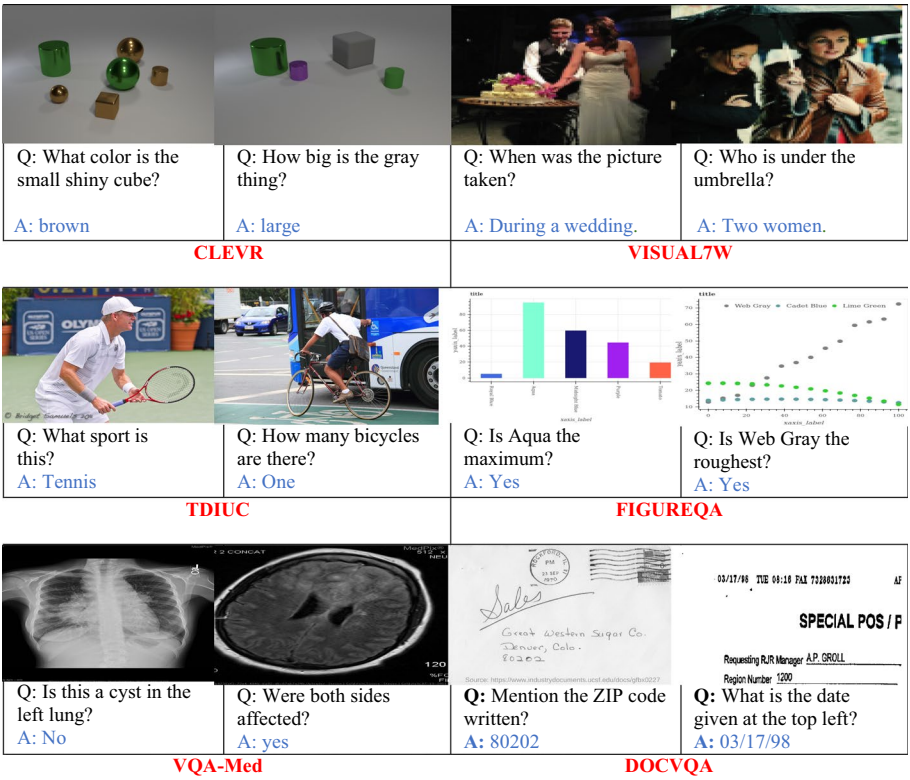
Q: is the small table both oval and wooden?

A: yes

Q: Is there a box inside the plastic bag?

A: no

GQA



Acknowledgements We thank the anonymous reviewers for their valuable comments and suggestions. This work is supported by the National Key R&D Program of China No. 2017YFB1002101 and the Joint Advanced Research Foundation of China Electronics Technology Group Corporation No. 6141B08010102.

References

Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6077–6086

Andreas J, Rohrbach M, Darrell T, Klein D (2016) Neural module networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 39–48

Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) Vqa: visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433

Asif NA, Sarker Y, Chakrabortty RK, Ryan MJ, Ahamed MH, Saha DK, Tasneem Z (2021) Graph neural network: a comprehensive review on non-euclidean space. IEEE Access

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) Dbpedia: a nucleus for a web of open data. The semantic web. Springer, Berlin, pp 722–735

Ben-Younes H, Cadene R, Cord M, Thome N (2017) Mutan: multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2612–2620

Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y, Huang J (2020) Rumor detection on social media with bi-directional graph convolutional networks. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, no 01, pp 549–556

Biten AF, Tito R, Mafla A, Gomez L, Rusinol M, Valveny E, Karatzas D (2019) Scene text visual question answering. In: Proceedings of the IEEE/CVF International conference on computer vision, pp 4291–4301

Cadene R, Ben-Younes H, Cord M, Thome N (2019) Murel: multimodal relational reasoning for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1989–1998

Chen L, Wu L, Hong R, Zhang K, Wang M (2020) Revisiting graph based collaborative filtering: a linear residual graph convolutional network approach. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, no 01, pp 27–34

Cho K, van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder–decoder approaches. In: Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation, pp 103–111

Chou SH, Chao WL, Lai WS, Sun M, Yang MH (2020) Visual question answering on 360deg images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1607–1616

Dai H, Li C, Coley CW, Dai B, Song L (2019) Retrosynthesis prediction with conditional graph logic network. In: Proceedings of the 33rd international conference on neural information processing systems, pp 8872–8882

Do K, Tran T, Venkatesh S (2019) Graph transformation policy network for chemical reaction prediction. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp 750–760

Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Conference on empirical methods in natural language processing, pp 457–468, ACL

Gao D, Wang R, Shan S, Chen X (2020b) Learning to recognize visual concepts for visual question answering with structural label space. IEEE J Sel Top Signal Process 14(3):494–505

Gao D, Li K, Wang R, Shan S, Chen X (2020a). Multi-modal graph neural network for joint reasoning on vision and scene text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12746–12756

Goceri E (2019) Analysis of deep networks with residual blocks and different activation functions: classification of skin diseases. In: 2019 ninth international conference on image processing theory, tools and applications (IPTA). IEEE, pp 1–6

Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) Making the v in vqa matter: elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 6904–6913

Guo M, Chou E, Huang DA, Song S, Yeung S, Fei-Fei L (2018) Neural graph matching networks for few-shot 3d action recognition. In: Proceedings of the European conference on computer vision (ECCV), pp 653–669

Guo S, Lin Y, Feng N, Song C, Wan H (2019) Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, no 01, pp 922–929

Gupta D, Suman S, Ekbal A (2021) Hierarchical deep multi-modal network for medical visual question answering. Expert Syst Appl 164:113993

Gurari D, Li Q, Stangl AJ, Guo A, Lin C, Grauman K, Bigham JP (2018) Vizwiz grand challenge: answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3608–3617

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Hu Z, Wei J, Huang Q, Liang H, Zhang X, Liu Q (2020) Graph convolutional network for visual question answering based on fine-grained question representation. In: 2020 IEEE fifth international conference on data science in cyberspace (DSC), pp 218–224, IEEE

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

Ilievski I, Yan S, Feng J (2016) A focused dynamic attention model for visual question answering. Preprint http://arxiv.org/abs/1604.01485

Johnson J, Hariharan B, Van Der Maaten L, Fei-Fei L, Lawrence Zitnick C, Girshick R (2017) Clevr: a diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2901–2910

Kafle K, Price B, Cohen S, Kanan C (2018) Dvqa: understanding data visualizations via question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5648–5656

Kafle K, Kanan C (2017) Visual question answering: datasets, algorithms, and future challenges. Comput vis Image Underst 163:3–20

Kahou SE, Michalski V, Atkinson A, Kádár Á, Trischler A, Bengio Y (2018) FigureQA: an annotated figure dataset for visual reasoning. ICLR 2018

Kallooriyakath LS, Jithin MV, Bindu PV, Adith PP (2020) Visual question answering: methodologies and challenges. In: 2020 international conference on smart technologies in computing, electrical and electronics (ICSTCEE). IEEE, pp 402–407

Kembhavi A, Seo M, Schwenk D, Choi J, Farhadi A, Hajishirzi H (2017) Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4999–5007

Kim J, On KW, Lim W, Kim J, Ha J, Zhang B (2017) Hadamard product for low-rank bilinear pooling. In: proceeding of international conference on learning representations

Kim ES, Kang WY, On KW, Heo YJ, Zhang BT (2020) Hypergraph attention networks for multimodal learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14581–14590

Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks Preprint http://arxiv.org/abs/1609.02907. ICLR 2017

Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Fei-Fei L (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput vis 123(1):32–73

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105

Kumar A, Irsoy O, Ondruska P, Iyyer M, Bradbury J, Gulrajani I, Socher R (2016) Ask me anything: dynamic memory networks for natural language processing. In: International conference on machine learning. PMLR, pp 1378–1387

Malinowski M, Rohrbach M, Fritz M (2015) Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the IEEE international conference on computer vision, pp 1–9

Manmadhan S, Kovoor BC (2020) Visual question answering: a state-of-the-art review. Artif Intell Rev 53(8):5705–5745

Marino K, Rastegari M, Farhadi A, Mottaghi R (2019) Ok-vqa: a visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp 3195–3204

Mathew M, Karatzas D, Jawahar CV (2021) DocVQA: a dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2200–2209

Mishra A, Shekhar S, Singh AK, Chakraborty A (2019) OCR-VQA: visual question answering by reading text in images. In: 2019 international conference on document analysis and recognition (ICDAR), Sydney, NSW, pp 947–952. https://doi.org/10.1109/ICDAR.2019.00156

Narasimhan M, Lazebnik S, Schwing AG (2018) Out of the box: reasoning with graph convolution nets for factual visual question answering. In: Proceedings of the 32nd international conference on neural information processing systems, pp 2659–2670

Nguyen TH, Grishman R (2018) Graph convolutional networks with argument-aware pooling for event detection. In: Thirty-second AAAI conference on artificial intelligence.

Noh H, Han B (2016) Training recurrent answering units with joint loss minimization for vqa. Preprint http://arxiv.org/abs/1606.03647

Norcliffe-Brown W, Vafeias E, Parisot S (2018) Learning conditioned graph structures for interpretable visual question answering. In: Proceedings of the 32nd international conference on neural information processing systems, pp 8344–8353

Pei X, Yu L, Tian S (2020) AMalNet: a deep learning framework based on graph convolutional networks for malware detection. Comput Secur 93:101792

Ren S, He K, Girshick R, Sun J (2016) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149

Ren M, Kiros R, Zemel RS (2015) Exploring models and data for image question answering. In: Proceedings of the 28th international conference on neural information processing systems, vol 2, pp 2953–2961

Schlichtkrull M, Kipf TN, Bloem P, Van Den Berg R, Titov I, Welling M (2018) Modeling relational data with graph convolutional networks. In: European semantic web conference. Springer, Cham. pp 593–607

Shah S, Mishra A, Yadati N, Talukdar PP (2019) Kvqa: knowledge-aware visual question answering. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, no 01, pp 8876–8884

Shih KJ, Singh S, Hoiem D (2016) Where to look: focus regions for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4613–4621

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Preprint http://arxiv.org/abs/1409.1556

Singh AK, Mishra A, Shekhar S, Chakraborty A (2019a) From strings to things: knowledge-enabled VQA model that can read and reason. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4602–4612

Singh A, Natarajan V, Shah M, Jiang Y, Chen X, Batra D, Rohrbach M (2019b) Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8317–8326

Speer R, Chin J, Havasi C (2017) Conceptnet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the AAAI conference on artificial intelligence, vol 31, no 1

Sundermeyer M, Schlüter R, Ney H (2012) LSTM neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association

Tandon N, De Melo G, Suchanek F, Weikum G (2014) Webchild: harvesting and organizing commonsense knowledge from the web. In: Proceedings of the 7th ACM international conference on web search and data mining, pp 523–532

Teney D, Liu L, van Den Hengel A (2017) Graph-structured representations for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9

Trott A, Xiong C, Socher R (2018) Interpretable counting for visual question answering. In: International conference on learning representations

Wang Z, Luo N, Zhou P (2020) GuardHealth: Blockchain empowered secure data management and graph convolutional network enabled anomaly detection in smart healthcare. J Parallel Distrib Comput 142:1–12

Wang P, Wu Q, Shen C, Dick A, van den Hengel A (2018a) FVQA: fact-based visual question answering. IEEE Trans Pattern Anal Mach Intell 40(10):2413–2427

Wang X, Ye Y, Gupta A (2018b) Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6857–6866

Wang Z, Lv Q, Lan X, Zhang Y (2018c) Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proceedings of the 2018c conference on empirical methods in natural language processing, pp 349–357

Wang Y, Yin H, Chen H, Wo T, Xu J, Zheng K (2019) Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp 1227–1235

Wu Z, Palmer M (1994) Verbs semantics and lexical selection. In: Proceedings of 32nd annual meeting on association for computational linguistic, pp 133–138

Wu Q, Wang P, Shen C, Dick A, Van Den Hengel A (2016) Ask me anything: free-form visual question answering based on knowledge from external sources. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4622–4630

Wu C, Liu J, Wang X, Dong X (2018) Object-difference attention: a simple relational attention for visual question answering. In: Proceedings of the 26th ACM international conference on multimedia, pp 519–527

Wu Y, Lian D, Xu Y, Wu L, Chen E (2020a) Graph convolutional networks with markov random field reasoning for social spammer detection. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, no 01, pp 1054–1061

Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY (2020b) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32(1):4–24

Xu X, Wang T, Yang Y, Hanjalic A, Shen HT (2020) Radial graph convolutional network for visual question generation. IEEE Trans Neural Netw Learn Syst 32(4):1654–1667

Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 21–29

Yang J, Lu J, Lee S, Batra D, Parikh D (2018). Graph r-cnn for scene graph generation. In: Proceedings of the European conference on computer vision (ECCV), pp 670–685

Yang Z, Qin Z, Yu J, Hu Y (2019) Scene graph reasoning with prior visual relationship for visual question answering. Preprint http://arxiv.org/abs/1812.09681

Yao L, Mao C, Luo Y (2019) Graph convolutional networks for text classification. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, no 01, pp 7370–7377
Yu J, Zhu Z, Wang Y, Zhang W, Hu Y, Tan J (2020) Cross-modal knowledge reasoning for knowledge-based visual question answering. Pattern Recognit 108:107563
Zhang Y, Hare J, Prügel-Bennett A (2018a) Learning to count objects in natural images for visual question answering. In: International conference on learning representations.
Zhang Y, Qi P, Manning CD (2018b) Graph convolution over pruned dependency trees improves relation extraction. In: Proceedings of the 2018b conference on empirical methods in natural language processing, pp 2205–2215
Zhang J, Shi X, Zhao S, King I (2019a) STAR-GCN: stacked and reconstructed graph convolutional networks for recommender systems. In IJCAI
Zhang S, Tong H, Xu J, Maciejewski R (2019b) Graph convolutional networks: a comprehensive review. Comput Soc Netw 6(1):1–23
Zhou X, Shen F, Liu L, Liu W, Nie L, Yang Y, Shen HT (2020) Graph convolutional network hashing. IEEE Trans Cybern 50(4):1460–1472
Zhu Y, Groth O, Bernstein M, Fei-Fei L (2016) Visual7w: grounded question answering in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4995–5004
Zhu X, Mao Z, Chen Z, Li Y, Wang Z, Wang B (2021) Object-difference drived graph convolutional networks for visual question answering. Multimed Tools Appl 80(11):16247–16265
Zitnik M, Agrawal M, Leskovec J (2018) Modeling polypharmacy side effects with convolutional networks. Bioinformatics 34(13):i457–i466