J. Cent. South Univ. (2021) 28: 728−746
DOI: https://doi.org/10.1007/s11771-021-4641-x

Springer

A survey of deep learning-based visual question answering

HUANG Tong-yuan(黄同愿)[1, 2], YANG Yu-ling(杨钰玲)[1], YANG Xue-jiao(杨雪姣)[2]

1. School of Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China;
2. School of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China

Abstract: With the warming up and continuous development of machine learning, especially deep learning, the research on visual question answering field has made significant progress, with important theoretical research significance and practical application value. Therefore, it is necessary to summarize the current research and provide some reference for researchers in this field. This article conducted a detailed and in-depth analysis and summarized of relevant research and typical methods of visual question answering field. First, relevant background knowledge about VQA(Visual Question Answering) was introduced. Secondly, the issues and challenges of visual question answering were discussed, and at the same time, some promising discussion on the particular methodologies was given. Thirdly, the key sub-problems affecting visual question answering were summarized and analyzed. Then, the current commonly used data sets and evaluation indicators were summarized. Next, in view of the popular algorithms and models in VQA research, comparison of the algorithms and models was summarized and listed. Finally, the future development trend and conclusion of visual question answering were prospected.

Key words: computer vision; natural language processing; visual question answering; deep learning; attention mechanism

Cite this article as: HUANG Tong-yuan, YANG Yu-ling, YANG Xue-jiao. A survey of deep learning-based visual question answering [J]. Journal of Central South University, 2021, 28(3): 728−746. DOI: https://doi.org/10.1007/s11771-021-4641-x.

1 Introduction

Significant progress has been made in computer vision and natural language processing in recent years, but the joint task involving both still faces enormous challenges. In 2014, visual question answering (VQA) was presented as an emerging study combining these two tasks to generate answers from a given picture and a question about it, as shown in Figure 1.

VQA can be combined with image captioning (IC), visual problem generation(VQG) and visual dialogue(VD) to create an intelligent proxy that can perform human tasks in the real world and communicate with humans through language. It can also be applied to many specific areas, such as helping intelligence analysts, and visually impaired people to obtain image information from the web or life, and using image retrieval without image tags.

A large amount of researches have been accumulated in the field of visual question-and-answer, and the answer with the highest repetition rate for random guessing is its baseline method [1], which is often used to determine the quality of data sets and the minimum standards that other algorithms should meet. AGRAWAL et al [2] introduced the VQA task, combined pictures and

Foundation item: Project(61702063) supported by the National Natural Science Foundation of China
Received date: 2020-04-18; **Accepted date:** 2020-06-10
Corresponding author: HUANG Tong-yuan, Associate Professor; Tel: +86-13983955665; E-mail: tyroneh@cqut.edu.cn; ORCID: https://orcid.org/0000-0002-3155-6728

Figure 1 An example of visual question answering: (a) Image 1; (b) Image 2 (**Q:** How is the ground? **Q:** Where is the bird standing on? dry A: Sand B: Parking meter(correct))

problem features into a single vector, and classified them by nonlinear method. Then, MALINOWSKI et al [3] used the Bayesian algorithm for VQA tasks for the first time, first using semantic segmentation method to identify objects and positions in the picture, and then training Bayesian algorithm to simulate the spatial relationship between objects, to calculate the probability of each answer. To further improve the performance of the model, SHIH et al [4] introduced visual attention to highlight the most relevant image areas of the answer. To further increase the generalization capability of the model, MA et al [5] used memory enhancement networks to improve the memory ability of the model for the uncommon question-and-answer pair. In order to narrow the gap between machines and humans, FUKUI et al [6] introduced attention mechanisms and multi-modal joint embedding in the same framework. GORDON et al [7] introduced an interactive question-and-answer system, based on the provided scenes and questions, intelligent agents through the understanding of the visual scene for autonomous navigation, interaction with the real environment to obtain answers.

At present, there are few reviews of visual question answering tasks, and a small number of classical models are mainly studied in Refs. [8−10], which does not involve the research results of the past year or two. YU et al [11] mainly summarized the sub-issues of VQA mission, but still without integrity. Therefore, it is necessary to carry out a more detailed summary of the research in recent years to provide researchers in this field with a more comprehensive and holistic understanding.

This paper analyzed and summarized the four aspects of the visual question answering task. Section 1 made an introduction about the background of VQA. Section 2 discussed the problems and challenges facing VQA missions. Section 3 detailed the overall framework of VQA and the six sub- modules. Section 4 summarized the data sets currently available and the evaluation indicators. Section 5 summarized relevant algorithm performance comparison and listed them in tables. Section 6 combined the current situation of VQA research and looked forward to the future research trend. Section 7 made a summary.

2 Problems and challenges

The ideal visual question answering system should be similar to human vision and human dialogue. Humans can recognize objects in images, understand the spatial position of objects, infer relationships between them, and so on, and ask any questions about images through natural language. Thus, the VQA task goes far beyond the scope of image understanding and visual question answering system, and the main problems and challenges are summarized in Figure 2.

From the task scale, the visual question answering system is a comprehensive research task, and each sub-task has a profound impact on the final effect, which needs to be studied in depth. From the overall structure, the current model structure is relatively simple, most lying from local problems, and the design of reasonable and effective model framework research is very little.

2.1 Related research areas

In addition to the above questions and challenges, visual question answering can improve algorithm performance by combining with other

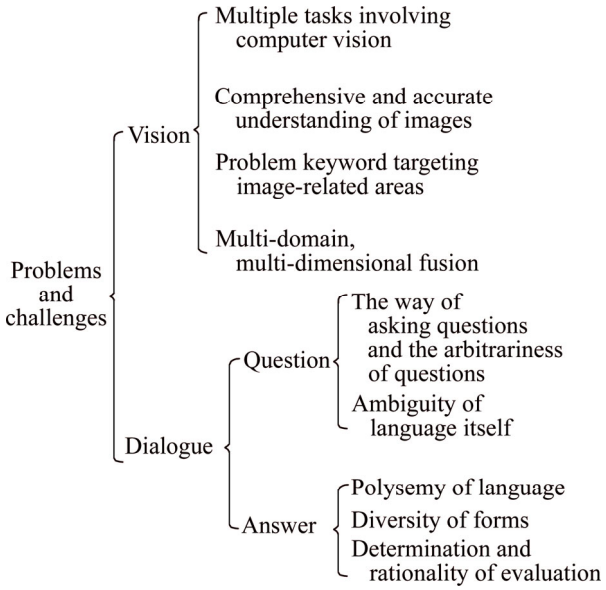

Figure 2 Problems and challenges

cross-cutting studies. Below will describe the role of image captioning, visual problem generation, and visual dialogue tasks in visual question answering.

2.1.1 Image caption

Image captions generate a description of a natural language based on the properties and object relationships of the image. Its universal framework is based on the Encoder-Decoder structure. Encoder encodes the target detected in the image into a vector, using the features of the last convolution layer or the fully connected layer as the image feature. Decoder maps the feature vector to the text. The processing of the image captions is shown in Figure 3.

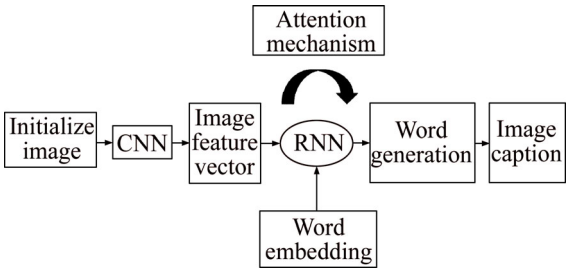

Figure 3 Flowchart of image captioning

The difference between image caption and visual question answering is that the former only needs to generate a general description of the image, while the latter needs to focus on different local areas of the image according to the problem, and complex problems require some intellectual reasoning. For image caption, the degree of understanding of images is arbitrary, while the degree of understanding of visual questions answering is specified by the nature of the questions asked. The correlation between the two reflected in the output of the image caption can be used as the input of the question answering system, providing a wealth of knowledge for the follow-up tasks. WU et al [12] used the long short term memory (LSTM) [13] network to generate image descriptions to serve as input to visual questions answering. It has made an important contribution to the follow-up development of LSTM and accelerated the progress of VQA. In this case, the problem and visual representation are input into the decoder, which is usually an LSTM network, and then trained to produce the corresponding answer. However, LSTM treats the problem as a series of words, which can not reflect the real complexity of language structure.

Image caption can also be directly used as inputs for the visual question answering system, JAIN et al [14] encoded images, subtitles, historical question-and-answer pairs, questions, and answers separately, using the stitched vector as input to the model.

At present, ZHOU et al [15] have proposed a new video description model that can use these boundary box annotations to achieve the most advanced performance in video description, video paragraph description and image description. In this work, they explicitly link sentences to the evidence in the video. To some extent, it provides a new way to solve the defects of LSTM.

2.1.2 Visual problem generation

Visual problem generation is to generate various types of questions for a given image, dynamically determined at running time, which does not need to fully understand the image and does not limit the range of correct answers, generally generates open questions, even questions that humans cannot answer.

Visual question answering and visual problem generation are mutually reinforcing relationships. HEDI et al [16] improved VQA and VQG, proposing a Tucker decomposition based on multi-modal tension, effective parametrization of the bilinear relationship between image and text expression, and designed a low-level decomposition based on matrix to clearly constrain the level of interaction. LI et al [17] conducted joint training with VQA and VQG to provide real answers as a

hint. LIU et al [18] put forward IVQA (inverse visual question answering), which took VQG as a multi-modal dynamic reasoning process, guided by partially generated questions and answers, and gradually adjusts the focus. Although there have been a lot of researches on neural problem generation, how to generate high-quality VQA from unstructured text is still a major challenge. Most of the existing neural problem generation methods try to solve an answer aware problem, in which an answer block and the surrounding paragraphs are the input of the model, and the output is the problem to be generated. They describe the task as a sequence to sequence problem and design various encoders, decoders and input features to improve the quality of the generated problem. However, the answer based question generation model is far from enough because the questions generated from an article are essentially one to many.

2.1.3 Visual dialogue

Visual dialogue is a meaningful dialogue between intelligent agents and humans that use natural language to observe visual content. Given an image, a historical conversation, and a question about the image, the smart agent must place the question in the image and the historical conversation, then infer the background, and answer it accurately.

Visual dialogue is the product of the further development of visual question answering tasks. According to the form of question answering, the visual dialogue is divided into one-way and bidirectional visual dialogue, and the former, is similar to the visual question answering, which uses the image and historical question-and-answer pair as input to generate the answer to the current moment. VRIES et al [19] introduced an interactive game that attempts to narrow the range of candidates by asking questions, obtaining ingesting answers from information given by users, and thus targeting the objects that users were interested in and understanding the interests of users. Although the current visual dialog model has achieved impressive performance, it is difficult for the model to give an accurate answer when the problem involves a specific area of the image or a more detailed dialogue segment. The reason is that single-step analysis needs too much information at the same time. When a specific position is needed or the problem, image and dialogue history need to be repeatedly understood, the single-step understanding is greatly limited. Therefore, multi-step reasoning is very necessary from coarse-grained to fine-grained.

3 Visual question answering

A large number of visual question answering studies were analyzed, with almost all models using CNN (convolutional neural network) to extract image features, RNN (recurrent neural network) extracting text semantic features and then merging features. Complex models introduce attention mechanisms to obtain better regional or textual attention, and a knowledge base to provide a more comprehensive source of knowledge. The fusion feature is eventually fed into the classifier or generator. The entire workflow is roughly summarized as shown in Figure 4.

3.1 Convolutional neural network

The convolutional neural network was first proposed by FUKUSHIMA [20] in 1980, by overlaying the network layer to extract the semantic

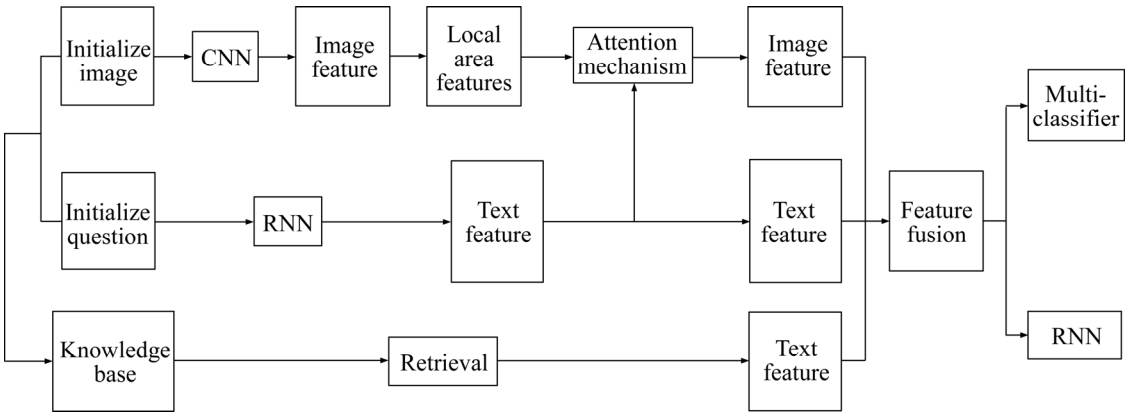

Figure 4 Flowchart of visual question answer system

features of different levels of the image, so as to complete the follow-up tasks such as image classification, object detection, behavior recognition and image segmentation. Therefore, CNN is the basis of image understanding and application, the extracted image features will directly affect the performance of the subsequent high-level tasks. In recent years, a variety of deep learning models based on CNN have emerged, reaching model convergence through continuous iteration of "input-forward propagation-output-loss calculation-reverse propagation", the classic CNN model summarized in Table 1.

The primary goal of visual question answering is to extract the semantic information related to the problem from the image, which ranges from the detection of small details to the abstract understanding of the whole image. Based on the scope of image feature extraction, the visual question and answer model is classified into two categories: the global feature of the image extracted by CNN and the image area feature extracted from the region proposal network (RPN). The network model that uses global and local features for this area is summarized in Table 2.

3.2 Recurrent neural network

There are many sequences of data (video/text/voice, etc.) in life that are relevant in timing, i.e., the output of a certain moment is related to the output of the current input and the previous moments. Recurrent neural networks, which are designed for sequence data and can handle fixed or variable length data, are widely used in areas such as machine translation, text processing and speech recognition, and are increasingly popular in other fields.

Recurrent neural networks are divided into simple recurrent neural networks and complex recurrent neural networks. The former was first proposed by ELMAN [39] in 1990, by entering each word in the text in turn, using the hidden layer output corresponding to the last word as semantic information for the entire text, and retaining the information above. Complex recurrent neural networks explore many ways to extract text features, including the BOW (Bag Of Word) model, the LSTM encoder, the GRU (Gate Recurrent Unit) and Skip-Thought Vectors [40], etc.

Compared with the traditional BOW model, RNN model can capture word sequences and reduce the size of parameters by sharing parameters. Early RNN can only remember the contents of finite time units, and the improved version can better capture long-distance information, such as LSTM, GRU, etc. In addition to their ability of long-term memory,

Table 1 Classic convolutional neural network model

Literature	Model name	Main contribution	Years	Characteristics
[21]	LeNet5	A complete network structure (convolution/pooling/total) is proposed for the first time	1994	Basic network
[22]	AlexNet	The nonlinear activation function ReLU and Dropout are introduced to prevent over fitting, data enhancement and LRN normalization layer, expand network model and accelerate calculation with multiple GPU	2012	Basic network
[23]	VGGNet	All convolution layers use 3×3 convolution kernel to obtain larger receptive field and characterize complex features	2014	3×3 convolution kernel
[24]	NiN	A 1 × 1 convolution kernel is used to extract more joint features; The average pool layer replaces the full connection layer	2013	1×1 convolution kernel; No full connection layer
[25]	GoogLeNet (Inception V1)	For the first time, the Inception architecture is proposed, using 1×1 convolution to reduce the number of features before parallel modules	2015	Inception
[26] [27]	Inception V2 Inception V3	V2 calculates the mean and standard deviation on the output, and V3 splits the two-dimensional convolution into two smaller one-dimensional convolutions to deal with more and richer spatial features	2015	Bulk naturalization; Asymmetric convolution
[28]	ResNet	Residual learning framework, reduce the burden of network training, make the network layer deeper, can achieve a variety of parallel modules or continuous modules	2015	Jump connection
[29]	DenseNet	Any two layers are directly connected, and the input of each layer is a combination of all the previous layers' outputs	2017	Dense connection
[30]	SENet	The fusion between feature channels is carried out by the method of feature relocation	2018	Channel attention mechanism

Table 2 Visual question answering algorithm based on convolutional neural network

Feature extraction	Literature	Category
CNN (VGGNet/ResNet/ GoogleNet, etc.)	[18−19, 31−34]	Global features
RPN (Edge Box/Selective Search/RPN)	[8, 35−38]	Regional characteristics

they can also alleviate problems such as the disappearance of gradients. At present, using RNN to extract text semantic information is the main method of VQA task.

3.3 Feature fusion

Feature fusion is an essential part of cross-cutting research, and the mapping of different feature expressions to uniform feature space through feature fusion is to explore the better joint embedding expression of "image problem pair".

A two-branch neural network is proposed in Ref. [41], consisting of CNN image encoder and an LSTM problem encoder, and the image encoding and problem encoding are fused before being transferred to the decoder. For better feature expression, LIN et al [42] changed CNN's full-connection layer to a bilinear layer. Using its ideas, FUKUI et al [6] introduced the MCB (multi-modal compact bilinear) to more compactly fuse images and text features through external products. Because the outer product produces high dimensional features, GAO et al [43] further compressed the fusion features, after compression still tended to high dimensions. As a result, KIM et al [44] proposed MLB (multimodal low-rank bilinear), which used Hadamard for feature fusion, but with a slower convergence. YU et al [45] put forward MFB (multi-modal factorized bilinear), which used matrix decomposition technique to combine features, which not only reduced the amount of parameters, but also improved the convergence speed. In Refs. [46, 47], External knowledge and multi-modality residual learning were explored to improve the effectiveness of joint embedded expression.

At present, the characteristic fusion method of series, corresponding elements added, point multiplication and ex-product is used. But most methods interact only through simple matrix operations, and do not achieve intensive interaction of images and problem characteristics. In response to this problem, SHRESTHA et al [48] proposed a new VQA algorithm that can focus on the VQA data set under natural image understanding while taking into account the synthetic data set of test reasoning. In order to better capture the high-level interaction between the language and visual fields, thereby improving the answer performance of visual questions, GAO et al [49] proposed a new method for dynamic fusion of multi-modality features and information flow in and between modules.

3.4 Attention mechanism

Humans focus on specific areas of the image when answering questions, so the researchers also hope that the visual question answering model will focus on the "relevant areas" of the image or the "keywords" of the text. Attention mechanisms focuses on the most important image areas and problem words and plays an important role in all areas of artificial intelligence. According to the differences in concerns, there are four main parts, as shown in Figure 5.

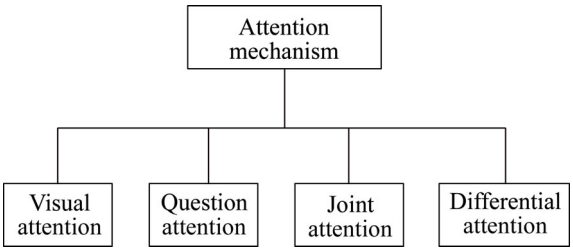

Figure 5 Attention mechanism

Visual attention focuses attention to a specific area of the image based on a given problem, with the area of the image and the relevance of the problem weighting the area. KAZEMI et al [50] learned the expression of characteristics of a particular area based on a given problem combined with spatial attention. YANG et al [51] introduced SANs (stacked attention networks), using the problem features to search the image area, in turn to generate multiple attention maps, and finally select the most relevant area. However, some questions involved multiple areas of the image, and ZHU et al [52] proposed a structured attention model that encodes cross-regional relationships, modeling visual attention as multiple distributions in a conditional random field, with the aim of correctly answering questions involving complex regional relationships.

YIN et al [53] further considered the use of images to guide the attention of key words.

ILIEVSKI et al [54] took the detected object region as the candidate region of the problem, and then selected the most relevant region according to the problem. XU et al [55] treated the process of selecting an image area as a single "jump" that captures fine-grained information about the problem through multistage jumps. In order to narrow the gap between image features and problem features, YU et al [56] put forward a multilevel focus on the network, using the attention mechanism for semantic attention, thereby reducing the semantic gap. JANG et al [57] paid attention to a single sequence.

Joint attention is to learn the strong relevance of "image problem pair". LU et al [58] used joint attention to process the most relevant image regions and text words before joint embedding of image problem features. LIANG et al [59] proposed FVTA (focal visual text attention) network. In the sequence data, the hierarchical structure is used to dynamically determine which picture to be paid attention to and when to answer questions. NGUYEN et al [60] considered the dense symmetric interaction of image problems, so that the problem words and image regions were focused on each other to enhance the effect of the model.

Differential attention is a kind of differential attention with foreground and background, which is more in line with human's attention habit. For example, PATRO et al [61] learnt the differences between data by collecting positive and negative samples, so as to achieve different concerns. Experiments show that this helps to improve the accuracy of answers. JAIN et al [14] maximized the use of option information by iterating input history options.

Attention mechanism can effectively improve the performance of the model. At present, many researches only focus on the image region of the whole problem or only consider the limited correlation between the image region and the problem words. Although joint attention takes into account the common concern for problem words and images, it focuses on the whole image. Therefore, in order to obtain the complex relationship of image problem pairs, attention mechanism can be extended to deal with any interaction between any image regions and any problem words. In contrast, CADENE et al [62] abandoned the classical attention framework and used vector representation to simulate the semantic interaction between visual content and problems in each region.

XU et al [63] proposed two kinds of attention: soft mechanism and hard mechanism. LUONG et al [64] proposed two different types of attention, global and local. Global attention is actually soft attention; local attention is a compromise between hard attention and soft attention. According to the calculation area of attention, it can be divided into the following categories:

1) Soft attention, which is a common attention method, calculates the weight probability of all keys, and each key has a corresponding weight, which is a global calculation method (also called global attention). This method is more rational. It refers to the contents of all keys and then weights them. But the amount of calculation may be larger.

2) Hard attention. In this way, a key can be accurately located directly, and the rest of the keys are ignored. The probability of this key is 1, and the probability of other keys is 0. Therefore, this kind of alignment requires very high requirements, one-step in place, if not correctly aligned, will bring great impact. On the other hand, because it is not differentiable, it is generally necessary to use reinforcement learning method for training.

3) Local attention is actually a compromise between the above two methods. It calculates a window area. First use hard mode to locate a certain place. Take this point as the center, and then you can get a window area. In this small area, use the soft method to calculate the attention.

3.5 Knowledge and reasoning

According to the difficulty of the answer generation, visual question answering can be divided into three levels:

1) Easy. The answer is directly obtained from the results of image recognition.

2) Middle. The answer is too small or indistinguishable objects in the image, which needs the support of facts.

3) Hard. The answer is not in the image, which needs to be inferred according to the image content. For the "hard" problem, it has gone beyond the scope of image understanding, which may involve common sense, specific topics and encyclopedia knowledge. Therefore, knowledge base and

combinational reasoning are introduced into visual question answering.

At present, a lot of knowledge bases have been built, as listed in Table 3. In Ref. [72], images and problems were encoded into discrete vectors, and knowledge base is retrieved by vectors. WANG et al [73] proposed "Ahab" method to infer image content, build knowledge map by searching relevant knowledge, and learn the mapping of image problem to knowledge, but only for the problem of manual template analysis. SHEN et al [74] further learned the mapping of image problem to knowledge by introducing LSTM and data-driven, but the process of knowledge search is uncertain and knowledge fuzzy is introduced. ZHOU et al [75] used memory networks to alleviate this uncertainty. The modular network framework is adopted in joint reasoning, and the structural layout of the problem is assembled into predefined sub-tasks. A group of neural modules are designed to solve specific sub-tasks respectively. In Ref. [76], the problem and image were analyzed into tree or graph structure, and the local characteristics of nodes are combined to generate answers. SHIN et al [77] combined question answer pairs into a statement based on rules, which was used for later knowledge reasoning. For the visual context structure, only a few research works constitute the visual structure of the image. Traditional statistical methods can not put the object of each image as a whole in the context, so as to infer the way specific to content/task. TANG et al [78] then proposed a dynamic visual content and problem tree context model, VCTREE. Compared with the former, their tree structure is strengthened, so they do not need the correct marked data (ground truth). When human beings learn a task, we learn it constantly when we perform it. In machine learning, the trained model is frozen in the reasoning process. WORTSMAN et al [79] proposed a partial meta reinforcement learning method for adaptive visual navigation (SAVN), which can learn to adapt to the new environment in the process of training and reasoning even without any explicit supervision. It is proved that the interaction loss of acquisition is better than the manual loss. In contrast, NOH et al [80] used language knowledge resources such as structured vocabulary database (such as WordNet) and visual description to discover unsupervised tasks, constructed task conditional visual classifier to realize transfer learning, realized visual question answering, and solved the out of vocabulary answer problems in visual question answering tasks.

Table 3 Typical data sets for knowledge base

Literature	Data set	Data presentation
[65]	Dbpedia	Large knowledge base, unstructured data, about 4.58 million, including people, places, films, species, diseases, etc
[66]	Freebase	Structured data, 61 domains, 765 types, 230 thousand topics
[67]	ConcepNet	Common sense knowledge base, triple structure, 28 million relationship descriptions, 600000 Chinese data
[68]	KB	Structured data, diversified relationships, 500 million variables, millions of parameters
[69]	Webchild	Large common sense knowledge base, triple structure, 19 range sets of different relationships, more than 4 million fine-grained relationships
[70]	Knowledge Vault	Web data, 1.6 billion facts, 271 million "trusted facts"
[71]	NEIL	1152 object categories, 1034 scene categories, 87 attributes, more than 1700 relationships, more than 400000 visual instances

For the complex face distribution problem, WANG et al [81] proposed a face clustering method based on linkage by using graph convolution network. It has strong robustness in audio-visual face clustering by constructing case perspective sub graph (IPS) which describes the given node context. For the visually impaired users, in order to help them answer the questions about texts related to reading and reasoning in daily images, SINGH et al [82] proposed that inductive bias and special components (such as OCR for optical character recognition) should be put into the model, giving them the different skills that they need for VQA (e.g., reading, reasoning).

Data sets for the real world usually have long tailed skew distributions [13, 24], that is to say, a few classes (also known as the head class) occupy most of the data, while most classes (also known as the tail class) have few sample problems. ZHOU et al [83] proposed a unified two branch network (BBN) model, which takes both representation learning and classifier learning into account, so as to greatly improve long tail recognition. At the same time, a new cumulative learning strategy is proposed to adjust bilateral learning, which is combined with the training of BBN model.

3.6 Memory enhancement network

Cognitive research shows that when people answer questions, they will compare new stimuli with examples stored in memory, and synthesize some examples to generate answers [84]. A memory enhancement network is proposed to store useful historical information.

SUKHBAATAR et al [85] improved the memory network in an end-to-end learning way. In the training stage, no external knowledge was introduced, but better practicability was achieved. In order to store more important information, KUMAR et al [86] combined attention mechanism with memory network, which allowed memory network to selectively focus on specific input. Memory enhancement network adopts dynamic storage mode. XIONG et al [87] realized the dynamic storage of features in the attention gating loop unit. The previous research pays more attention to the head of the problem, but the tail of the problem is the key of the whole problem, at the same time, memory enhancement network was used to selectively focus on the tail of the problem, and LSTM was used to control the reading and writing of external memory. In Ref. [88], a neural network is proposed to enhance external memory, which can generate scarce data.

Memory enhancement network can contain internal or external memory blocks or both, and can selectively focus on each training sample. It can improve the accuracy of visual question answering, predict the scarce answers in the training data set, and keep the relatively long-term memory of the scarce training samples, which is very important. There are some limitations in the previous input response fusion methods, which can not correctly represent the common vectors of image, history and questions. They focus on the short and safe answers and ignore the detailed information. GUO et al [89] developed a collaborative network to learn the representation of images, questions, answers and history in a single step. Expand the traditional one-stage solution to two-stage solution. In the first phase, candidate answers are roughly graded based on their relevance to images and question pairs. Then, in the second stage, through the cooperation with images and questions, the answers with high accuracy are reordered. The efficiency network proposed by them takes the discriminant visual dialog model to a new level in the visual dialog v1.0 data set. WANG et al [90] proposed two new visual language navigation methods, RCM (Reinforced Cross-Modal Matching) and SIL (Self-Supervised Imitation Learning), which combined reinforcement learning and self supervised imitation learning, to increase the effectiveness and efficiency of standard test scenarios and lifelong learning scenarios. In the field of vision and navigation, KE et al [91] proposed a forward perceptual search with a backtracking (fast) navigator, a general framework for motion decoding.

4 Data sets and evaluation indicators

4.1 Common data sets

A large-scale and diverse data set is the basis of learning visual question answering tasks; common data sets are summarized in Table 4. In the first 12 data sets, except DAQUAR, other data sets contain images from Microsoft COCO data set [1]. COCO data set contains 3.28 million images, 91 common object classes, more than 2 million tagged instances, and each picture has 5 subtitles on average. Visual Genome and Visual7W also contain images from Flickr100M. SYNTH-VQA is a composite cartoon image in VQA data set. The rest of the VQA data set will be referred to as COCO-VQA. The last nine data sets in Table 4 are widely used in the past two years. TDIUC (task driven image understanding challenge) attempts to solve the problem type deviation of the annotator by dividing the problem into 12 different types, so as to achieve detailed task driven assessment. It has metrics to evaluate generalization across problem types. Data set of CLEVR-Humans is abbreviated as CLEVR-H [47].

4.2 Evaluation indicators

The correctness of sentence syntax and semantics should be considered in evaluating text description, and the relevance between answers and questions should be considered in visual question answering. According to the answer forms, there are two types: binary/multiple choice and open-ended questions. The former selects the answer from the options, and the latter is a string answer, which is divided into words, phrases or sentences.

The accuracy is used to evaluate the binary/multiple choice problem. The accuracy

Table 4 Typical data sets for vision question answer

Literature	Data set	Image source	Image	Question	Question/ Each image	Number of problem categories	Average problem length	Average answer length	Question source	Evaluation index	Characteristics
[3]	DAQUAR	NYUDv2	1449	12468	8.6	4	11.5	1.2	Human	Accuracy WUPS	Small data, indoor scene; image noise, difficult to answer questions.
[92]	COCO-QA	COCO	117684	117684	1	4	8.6	1	Automatic	Accuracy WUPS	There are 4 types of questions, the answers are words, and there are many grammatical errors in the question answer pairs.
[93]	VQA-real	COCO	20421	614163	3	20+	6.2	1.1	Human	Accuracy human	Picture: question: answer = 1:3:10; bias; subjectivity.
[94]	VQA-abstract	Clipart	50000	150000	3	20+	6.2	1.1	Human	Accuracy	Indoor and outdoor real scene; the answer type is comprehensive and fuzzy.
[94]	Visual7W	COCO	47300	327939	6.9	7	6.9	1.1	Human	Accuracy	Visual genome subset; comprehensive problem types.
[67]	FVQA	COCO/ ImageNet	1906	4608	2.5	12	9.7	1.2	Human	Accuracy	Triple (question/answer, supporting fact, fact); hard question.
[66]	KB-VQA	COCO	700	2402	3.4	23	6.8	2	Human	Human	Hard questions; each question follows one of 23 predefined templates.
[95]	Visual genome	COCO	108000	1445322	13.4	7	5.7	1.8	Human	Accuracy	There is no binary problem for specific image region, and the answers are diverse.
[96]	Visual Madlibs	COCO	10738	360001	33.5	12	6.9	2	Human	Accuracy	Fill in the blanks; The answer is given by 3 staff members.
[97]	VQA-balanced	Clipart	15623	33379	2.1	1	6.2	1	Human	Accuracy	Balance data of training and testing; language deviation of unused training set.
[98]	SHAPES	—	64	244	—	—	—	—	—	—	The problem involves attribute, relation and position; Binary answer, no deviation.
[99]	FM-IQA	COCO	120360	—	—	—	—	—	Human	Human	Keywords artificially generated question and answer pairs; the existing answer is sentence; difficult to evaluate.
[47]	VQAv1	Natural	204K	614K	—	—	—	—	Human	—	Multiple kinds of language bias, including some questions being heavily correlated with specific answers.
[47]	VQAv2	Natural	204K	1.1M	—	—	—	—	Human	—	Endeavors to mitigate this kind of language bias compared with vqav1.
[47]	TDIUC	Natural	167K	1. 6M	—	—	—	—	Both	—	Metrics to evaluate generalization across question types.
[47]	C-VQA	Natural	123K	369K	—	—	—	—	Human	—	Tests the ability to combine previously seen concepts in unseen ways.
[47]	VQACPv2	Natural	219K	603K	—	—	—	—	Human	—	Re-organizes vqav2 such that answers for each question type are distributed differently in the train and test sets.
[47]	CLEVR	Synthetic	100K	999K	—	—	—	—	Synthetic	—	A synthetically generated dataset, consisting of visual scenes with simple geometric shapes.

to be continued

Continued

Literature	Data set	Image source	Image	Question	Question/ Each image	Number of problem categories	Average problem length	Average answer length	Question source	Evaluation index	Characteristics
[47]	CLEVR-H	Synthetic	32K	32K	—	—	—	—	Human	—	Provide human-generated questions for CLEVR scenes to test generalization to free-form questions.
[47]	CoGenT-A	Synthetic	100K	999K	—	—	—	—	Synthetic	—	Belongs to CLEVR-cogent, tests the ability to handle unseen concept composition and remember old concept combinations.
[47]	CoGenT-B	Synthetic	30K	299K	—	—	—	—	Synthetic	—	Belongs to CLEVR-cogent, tests the ability to handle unseen concept composition and remember old concept combinations.

measurement formula is as follows:

$$\text{Accuracy}=\frac{\text{Correct answers}}{\text{Total number of questions}} \tag{1}$$

The open-ended question compares the predicted value with the real value, mainly using the following four methods:

1) Perfect match method. It is unreasonable that different levels of mistakes will be punished with the same punishment. By constraining the answer length (generally 1−3 words), the sentence matching problem can be avoided and the answer ambiguity can be reduced.

2) WUPS (Wu-Palmer similarity) [100] measures the similarity between the predicted value and the real value according to the semantic similarity, and assigns the weights from 0 to 1. The smaller the value, the lower the similarity. At the same time, set the threshold value; the value lower than the given threshold will be reduced in proportion, but there are still words with similar words and different meanings (such as white and black) with high weights, which are only applicable to strict semantic concepts, and the answer is always words.

3) Average consensus and minimum consensus. Each question is given multiple correct answers, combined with semantic similarity measurement. The average consensus is to pick a more popular answer. The minimum consensus is to agree with at least one real answer. The ratio of questions and correct answers is about 1:5 in the DAQUAR-consensus data set and 1:10 in the VQA data set. The accuracy measurement formula of VQA is as follows:

$$\text{Accuracy}_{\text{VQA}}=\min\left(\frac{n}{3},\ 1\right) \tag{2}$$

where n represents the same number of predicted values as 10 real values. This method helps to solve the fuzzy problem.

4) FM-IQA manual evaluation method. It applies to situations where there are multiple real answers. This method is extremely time-consuming and resource intensive, and the judges need to give judgment criteria. FM-IQA proposes two indicators: 1) Determine whether the correct answer is given by a person; 2) The score of 3-point system is 0 for complete error, 1 for partial correctness and 2 for complete correctness.

5) One of the main problems in VQA dataset is the skew distribution of problem types. In this case, simple precision will not work, especially for the less common types of problems. MPT (Mean-Per-Type) is proposed as a new performance index to represent the arithmetic or harmonic average accuracy of each problem type to compensate for the unbalanced distribution of problem types.

$$\text{MPT}^{\text{e,f}}=\sum_{t=1}^{T}A_t/T \tag{3}$$

where T means total number of question types; A_t means accuracy over question type t.

6) BLEU (bilingual evaluation understudy) and METEOR (metric for evaluation of translation with explicit ordering) were used as evaluation indexes of VQA, and were tested in VizWiz data set.

$$\text{BLEU}^{\text{g,h,i}}=\text{BP}\exp\left(\sum_{n=1}^{N}W_n\lg P_n\right) \tag{4}$$

where BP means brevity penalty; W_n means positive weights summing to one; P_n means precision score of entire corpus.

$$\mathrm{METEOR}^j=(1-\mathrm{pen})F_{\mathrm{mean}} \tag{5}$$

where j for calculation of pen and F_{mean}. The formulae (3), (4), (5) refer to Ref. [100].

5 Algorithm performance comparison

A good VQA algorithm should be capable of focusing on dataset that requires natural image understanding and synthetic datasets that test reasoning, two camps of visual question answering (VQA) research, but only a few VQA algorithms are. To solve this problem, recurrent aggregation of multimodal embeddings network (RAMEN) model for VQA is added in Table 5 to generalize across the two domains [47].

Among models in Table 5, bottom-up-attention and top-down (UpDn) are combined with bottom-up and top-down attention mechanisms to implement VQA, bottom-up mechanism generates object proposals from faster R-CNN, and top-down mechanism predicts attention distribution on proposals. Top down attention is task driven, and the problem is used to predict the attention weight of the image area. The model won the first place in the 2017 VQA workshop challenge. Question-conditioned graph (QCG) represents an image as a graph, in which the object level features proposed by bottom-up regions are regarded as graph nodes and edges to encode the interaction between regions with problem conditions. For each node, QC graph selects the neighborhood of the node with the strongest edge connection to form a problem-specific graph structure. The structure is processed by patch operator and the spatial graph is convoluted. Bilinear attention network (BAN) integrates visual and text patterns by considering the interaction between all regional proposals (visual channels) and all interrogative words (text channels). Unlike the dual attention mechanism, BAN deals with the interaction between all channels. It can be considered as an extension of the low rank bilinear pooling method which jointly represents each channel pair. BAN supports multiple glimpses of attention through the remaining connections of the connection. Relation network (RN) accepts each pair of region suggestions, embeds them, and summarizes all n2 pair embeddings to generate a vector that encodes the relationship between objects. This pairwise feature aggregation mechanism can perform component reasoning, as demonstrated by its performance on clevr datasets. However, the computational complexity of RN increases twice with the number of objects. When the number of objects is large, it will be very expensive to run it. Recently, it has been tried to reduce the number of pairwise comparisons by reducing the number of objects entered into RN.

The memory, attention and composition (MAC) networks use automatic learning to perform attention based reasoning in computational cells. Different from modular network, MAC learns inference mechanism directly from data, while modular network needs pre-defined modules to perform pre-defined reasoning function. Each MAC unit maintains a control state representing the

Table 5 Relevant algorithm comparison

Dataset or algorithm	UpDn/%	QCG/%	BAN/%	MAC/%	RN/%	RAMEN/%
VQAv1	60.62	59.90	62.98	54.08	51.84	61.98
VQAv2	64.55	57.08	67.39	54.35	60.96	65.96
TDIUC	68.82	65.57	71.10	66.43	65.06	72.52
CVQA	57.01	56.45	57.36	50.99	48.11	58.92
VQACPv2	38.01	38.32	39.31	39.36	26.70	39.21
CLEVR	80.04	46.73	90.79	98.00	95.97	96.92
CLEVR-Humans	54.51	28.12	60.23	50.20	57.65	57.87
CLEVR-CoGenT-A	82.47	59.63	92.50	98.04	96.45	96.74
CLEVR-CoGenT-B	72.22	53.45	79.48	90.41	84.68	89.07
Mean	64.18	51.69	69.00	66.05	65.26	71.02

inference operation and a memory state representing the result of the inference operation. It has a computer like structure with read, write and control units. Mac was evaluated on CLEVR dataset and reported significant improvements in challenging count and numerical comparison tasks.

Recurrent aggregation model of multimodal embedded networks (Ramen). It is designed as a simple concept architecture, which can adapt to complex natural scenes, and can also answer the questions that need complex synthetic reasoning chain. It can connect the visual features of spatial positioning and the early integration of problem features.

In order to realize the reading problem of VQA model, SINGH et al [82] took the first step and put forward a method called Look, Read, Reason & Answer (LoRRA). And in the two data sets combined with other models for comparative experiments, the results of the data are shown in Table 6.

Table 6 Relevant model comparison

Accuracy	Model	Test-dev/%
VQA 2.0	BUTD	65.32
	Counter	68.09
	BAN	69.08
	Pythia v0.1	68.49
	Pythia v0.3 [82]	68.71
	Pythia v0.1+LoRRA [82]	68.49
VizWiz	BAN	51.40
	Pythia v0.3 [82]	54.72

Based on the visual genome generated by scene map and VQA2.0 used for visual question answering, WORTSMAN et al [79] inferred the context and showed the experimental results. In Ref. [62], cells were integrated into a complete MuRel network, which gradually improved the interaction between vision and question, and could be used to define a more refined visualization scheme than just paying attention to maps. At the same time, the experimental results were verified. Table 6 combines the experimental results of the two literature.

6 Directions for future

VQA is now regarded as a complete task of artificial intelligence, and will be an important step towards the vision dialogue of artificial intelligence dream. The prerequisite of solving this complex task includes the mature research knowledge of basic tasks of computer vision and natural language processing.

Table 7 Relevant model comparison

Accuracy	Model	Yes/No /%	Number/ %	Other/ %	All/%	Test-std/%
VQA2.0 test-dev	Teney	81.82	44.21	56.05	65.32	—
	MUTAN	82.88	44.54	56.50	66.01	—
	MLB	83.58	44.92	56.34	66.27	—
	DA-NTN	84.29	47.14	57.92	67.56	—
	Count	83.14	51.62	58.97	68.09	—
	Chain	82.74	47.31	58.93	67.42	—
	Graph	83.53	47.09	58.6	67.56	—
	VCTREE-HL [78]	84.28	47.78	59.11	68.19	—
	Bottom-up	81.82	44.21	56.05	65.32	65.67
	Graph Att.	—	—	—	—	66.18
	MUTAN†	82.88	44.54	56.50	66.01	66.38
	MLB†	83.58	44.92	56.34	66.27	66.62
	DA-NTN	84.29	47.14	57.92	67.56	67.94
	Pythia	—	—	—	68.05	—
	Counter	83.14	51.62	58.97	68.09	68.41
	MuRel [62]	84.77	49.84	57.85	68.03	68.41

Note: † have been trained [62].

1) In the future VQA research, researchers should be committed to creating efficient, rich, unbiased, goal-oriented data sets to test the important characteristics of VQA. Object oriented VQA data sets (such as VizWiz) can appear more frequently. These data sets will be expanded and explored in the future.

2) In recent years, empirical studies on language model transfer learning show that rich and unsupervised pre training is an integral part of many language understanding systems. It can be seen that how to use transfer learning to obtain better training results is a research direction worthy of consideration.

3) VQA detects all objects in the image, and needs more natural image features for semantic segmentation. Through feature fusion, the information from multiple sources can be combined to obtain rich internal features, which opens up a broad green field for research.

4) At present, the evaluation method of open

VQA needs further study. Although Bleu is the most popular metric for machine translation (MT), the report shows that it can not use short sentences. In the current VQA system, most of the answers are short. The use of Ngram EVAluation (Neva) is worth considering.

Last but not least, a new open model which can ask any kind of problems and can carry out explicit reasoning with good performance is worth considering.

7 Conclusions

As a comprehensive research task, visual question answering involves many research problems. Although many progress has been made in all aspects, the following problems still exist:

1) Data sets and the shortcomings of the metrics themselves. Data distribution is uneven, insensitive to scarcity issues, language bias causes no images to perform well, and there are multiple correct answers. The diversity and comprehensiveness of data need to be improved. The evaluation indexes are not uniform, and the model comparison lacks comprehensiveness.

2) Most of the visual question answering algorithms focus on the improvement of each sub problem, which has no pertinence to the architecture design of the model. Therefore, the same question for different images may give the same answer.

3) We can not “find” the answer to the question like humans. Only simple questions can be answered according to the provided “image question pair”, and complex and out of range questions cannot be effectively answered.

At present, most of the researches focus on offline state, how to make the machine interact with the environment and complete the visual question answering; complete visual question answering tasks online in real time, and how to further introduce vision language into the field of action. It is a new hot issue. At present, there are also relevant researchers exploring in this field. The algorithm with both natural picture understanding ability and visual reasoning ability should be the future development of VQA trend. We hope that the machine can have the ability of asking, answering and acting. We hope that the machine can understand and process visual and language information, and complete corresponding actions, which will be more widely used in the future.

Contributors

HUANG Tong-yuan contributed to the conception of the study; YANG Yu-ling contributed significantly to analysis and manuscript preparation. YANG Xue-jiao performed the data analyses and wrote the manuscript.

Conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] ZHOU Bo-lei, TIAN Yuan-dong, SUKHBAATAR S. Simple baseline for visual question answering [EB/OL]. [2015-12-07]. https://arxiv.org/abs/1512.02167.

[2] AGRAWAL A, LU J, ANTOL S, MITCHELL M. VQA: Visual question answering [J]. International Journal of Computer Vision, 2017, 123(1): 4−31. DOI: 10.1007/s11263-018-1116-0.

[3] MALINOWSKI M, FRITZ M. A multi-world approach to question answering about real-world scenes based on uncertain input [C]// International Conference on Neural Information Processing Systems. New York: MIT Press, 2014: 1682−1690.

[4] SHIH K J, SINGH S, HOIEM D. Where to look: Focus regions for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2016: 4613−4621. DOI: 10.1109/CVPR.2016.499.

[5] MA Chao, SHEN Chun-hua, DICK A, WU Qi, WANG Peng, HENGEL V A, REID L. Visual question answering with memory-augmented networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 6975−6984. DOI: 10.1109/cvpr.2018.00729.

[6] FUKUI A, PARK H. D, YANG D, ROHRBACH A, DARRELL T, ROHRBACH M. Multimodal compact bilinear pooling for visual question answering and visual grounding [EB/OL] [2016-07-06]. https://arxiv.org/abs/1606.01847v3. DOI: 10.18653/v1/d16-1044.

[7] GORDON D, KEMBHAVI A, RASTEGARI M, REDMON J, FOX D, FARHADI A. IQA: Visual question answering in interactive environments [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 4089−4098. DOI: 10.1109/cvpr.2018.00430.

[8] KAFLE K, KANAN C. Visual question answering: Datasets, algorithms, and future challenges [J]. Computer Vision and Image Understanding, 2017, 163(1): 3−20. DOI: 10.1016/j.cviu.2017.06.005.

[9] WU Qi, TENEY D, WANG P, SHEN Chun-hua, DICK A, HENGEL V A. Visual question answering: A survey of

methods and datasets [J]. Computer Vision and Image Understanding, 2017, 163(3): 21−40. DOI: 10.1016/j.cviu.2017.05.001.

[10] GUPTA A K. Survey of visual question answering: Datasets and techniques [EB/OL] [2017-05-10]. https://arxiv.org/abs/1705.03865.

[11] YU Jun, WANG Liang, YU Zhou. Research on visual question answering techniques [J]. Journal of Computer Research and Development, 2018, 55(9): 122−134. DOI: 10.7544/issn1000-1239.2018.20180168.

[12] WU Qi, WANG P, SHEN Chun-hua, DICK A, HENGEL V. Image captioning and visual question answering based on attributes and external knowledge [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(6): 1367−1381. DOI:10.1109/TPAMI.2017.2708709.

[13] GRAVES A. Long short-term memory [M]// Supervised Sequence Labelling with Recurrent Neural Networks. Germany: Springer Berlin Heidelberg, 2012: 1735−1780. DOI:10.1007/978-3-642-24797-2_4.

[14] JAIN U, SCHWING A G, LAZEBNIK S. Two can play this game: Visual dialog with discriminative question generation and answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 5754−5763. DOI: 10.1109/CVPR.2018.00603.

[15] ZHOU Luo-wei, KALANTIDIS Y, CHEN Xin-lei, CORSO J J, ROHRBACH M. Grounded video description [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 6578−6587. DOI: 10.1109/INFOCOMMST.2015.7357328.

[16] HEDI B Y, CADENE R, CORD M, THOME N. MUTAN: Multimodal tucker fusion for visual question answering [C]// Proceedings of the IEEE International Conference on Computer Vision. USA: IEEE Press, 2017: 2612−2620. DOI:10.1109/ICCV.2017.285.

[17] LI Yi-kang, DUAN Nan, ZHOU Bo-lei, CHU Xiao, OUYANG W, WANG Xiao-gang. Visual question generation as dual task of visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6116−6124. DOI: 10.1109/CVPR.2018.00640.

[18] LIU Feng, XIANG Tao, HOSPEDALES T M, YANG Wan-kou, SUN Chang-yin. iVQA: Inverse visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 8611−8619. DOI: 10.1109/CVPR.2018.00898.

[19] VRIES D H, STRUB F, CHANDAR S, PIETQUIN O. Guesswhat? visual object discovery through multi-modal dialogue [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2017: 5503−5512. DOI:10.1109/CVPR.2017.475.

[20] FUKUSHIMA K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position [J]. Biological Cybernetics, 1980, 36(4): 193−202. DOI: 10.1007/BF00344251.

[21] LECUN Y, BOTTOU L, BENGIO Y, HAFFNER P. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278−2324. DOI: 10.1109/5.726791.

[22] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]// International Conference on Neural Information Processing Systems. United States: Curran Associates Inc. 2012: 1097−1105. DOI: 10.1145/3065386.

[23] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL] [2014-09-04]. https://arxiv.org/abs/1409.1556.

[24] LIN Min, CHEN Qiang, YAN Shui-cheng. Network in network [EB/OL] [2013-12-16]. https://arxiv.org/abs/1312.4400.

[25] SZEGEDY C, LIU Wei, JIA Yang-qing, SERMANET P, REED S, ANGUELOV D, ERHAN D, VANHOUCKE V, RABINOVICH A. Going deeper with convolutions [C]// IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Computer Society, 2015: 1−9. DOI: 10.1109/CVPR.2015.7298594.

[26] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [EB/OL] [2015-02-11]. https://arxiv.org/abs/1502.03167.

[27] SZEGEDY C, VANHOUCKE V, IOFFE S, SHLENS J, WOJNA Z. Rethinking the inception architecture for computer vision [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2016: 2818−2826. DOI: 10.1109/CVPR.2016.308.

[28] HE Kai-ming, ZHANG Xiang-yu, REN Shao-qing, SUN Jian. Deep residual learning for image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2016: 770−778. DOI: 10.1109/CVPR.2016.90.

[29] HUANG Gao, LIU Zhuang, MAATEN L V D, WEINBERGER Q K. Densely connected convolutional networks [C]// IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Computer Society, 2017: 2261−2269. DOI: 10.1109/CVPR.2017.243.

[30] HU Jie, SHEN Li, SUN G. Squeeze-and-excitation networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Computer Society, 2018: 7132−7141. DOI: 10.1109/TPAMI.2019.2913372. DOI:10.1109/CVPR.2018.00745.

[31] MASSICETI D, SIDDHARTH N, DOKANIA P K, TORR H P. FLIPDIAL: A generative model for two-way visual dialogue [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 6097−6105. DOI: 10.1109/CVPR.2018.00638.

[32] CAO Qing-xing, LIANG Xiao-dan, LI Bai-ling, LI Guan-bin, LIN Liang. Visual question reasoning on general dependency tree [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 7249−7257. DOI: 10.1109/CVPR.2018.00757.

[33] HU He-xiang, CHAO Wei-lun, SHA Fei. Learning answer embeddings for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 5428−5436. DOI: 10.1109/CVPR.2018.00569.

[34] ANTOL S, AGRAWAL A, LU J, MITCHELL M. VQA:

Visual question answering [J]. International Journal of Computer Vision, 2017, 123(1): 4−31. DOI: 10.1007/s11263-018-1116-0.

[35] SHIN A, USHIKU Y, HARADA T. Customized image narrative generation via interactive visual question generation and answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 8925−8933. DOI: 10.1109/CVPR.2018.00930.

[36] LI Rui-yu, JIA Jia-ya. Visual question answering with question representation update [C]// Advances in Neural Information Processing Systems. Spain: Curran Associates Inc, 2016: 4655−4663.

[37] TENEY D, ANDERSON P, HE Xiao-dong, HENGEL V A. Tips and tricks for visual question answering: Learnings from the 2017 challenge [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 4223−4232. DOI: 10.1109/CVPR.2018.00444.

[38] ANDERSON P, HE Xiao-dong, BUEHLER C. Bottom-up and top-down attention for image captioning and visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 6077−6086. DOI: 10.1109/CVPR.2018.00636.

[39] ELMAN J L. Finding structure in time [J]. Cognitive Science, 1990, 14(2):179−211.

[40] KIROS R, ZHU Yu-kun, SALAKHUTDINOV R, et al Skip-thought vectors [J]. Computer Science, 2015, 27(28): 23−36. DOI: 10.1207/s15516709cog1402_1.

[41] MALINOWSKI M, ROHRBACH M, FRITZ M. Ask your neurons: A neural-based approach to answering questions about images [C]// IEEE International Conference on Computer Vision. USA: IEEE Press, 2015: 1−9. DOI: 10.1109/ICCV.2015.9.

[42] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNNS for fine-grained visual recognition [EB/OL] [2015-04-29]. 2015: 1449−1457. DOI: 10.1109/ICCV.2015.170.

[43] GAO Yang, BEIJBOM O, ZHANG Ning, DARRELL T Compact bilinear pooling [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2016: 317−326. DOI: 10.1109/CVPR.2016.41.

[44] KIM J H, ON K W, LIM W, KIM J, HA J W, ZHANG B T. Hadamard product for low-rank bilinear pooling [EB/OL] [2016-11-14]. https://arxiv.org/abs/1610.04325.

[45] YU Zhou, YU Jun, FAN Jian-ping, TAO Da-cheng. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering [C]// IEEE International Conference on Computer Vision. USA: IEEE Computer Society, 2017: 1839−1848. DOI: 10.1109/ICCV.2017.202.

[46] WU Qi, WANG P, SHEN Chun-hua, DICK A, HENGEL V. Ask me anything: Free-form visual question answering based on knowledge from external sources [C]// Computer Vision and Pattern Recognition. USA: IEEE Press, 2016: 4622−4630. DOI: 10.1109/CVPR.2016.500.

[47] KIM J H, LEE S W, KWAK D, HEO M O, KIM J, HA J W, ZHANG B T. Multimodal residual learning for VQA [C]// Advances in Neural Information Processing Systems. Spain: Curran Associates Inc, 2016: 361−369.

[48] SHRESTHA R, KAFLE K, KANAN C. Answer them all! Toward universal visual question answering models [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 10472−10481. DOI: 10.1109/CVPR.2019.01072.

[49] PENG Gao, JIANG Zheng-kai, YOU Hao-yuan, LU Pan, HOI S, WANG Xiao-gang, LI Hong-sheng. Dynamic fusion with intra- and inter- modality attention flow for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 6639−6648. DOI: 10.1109/CVPR.2019.00680.

[50] KAZEMI V, ELQURSH A. Show, ask, attend, and answer: A strong baseline for visual question answering [EB/OL] [2017-04-11]. https://arxiv.org/abs/1704.03162.

[51] YANG Zi-chao, HE Xiao-dong, GAO Jian-feng, DENG Li, SMOLA A. Stacked attention networks for image question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2016: 21−29. DOI: 10.1109/CVPR.2016.10.

[52] ZHU Chen, ZHAO Yan-peng, HUANG Shuai-yi, TU Ke-wei, MA Yi. Structured attentions for visual question answering [C]// Proceedings of the IEEE International Conference on Computer Vision. USA: IEEE Press, 2017: 1291−1300. DOI: 10.1109/ICCV.2017.145.

[53] YIN Wen-peng, SCHÜTZE H, XIANG Bing, XIANG Bing. ABCNN: Attention-based convolutional neural network for modeling sentence pairs [J]. Transactions of the Association for Computational Linguistics, 2016, 4(4): 259−272. DOI: 10.1162/tacl_a_00097.

[54] ILIEVSKI I, YAN Shui-cheng, FENG Jia-shi. A focused dynamic attention model for visual question answering [EB/OL] [2016-04-06]. https://arxiv.org/abs/1604.01485.

[55] XU Hui-juan, SAENKO K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering [C]// European Conference on Computer Vision. Netherlands: Springer International Publishing, 2016: 451−466. DOI: 10.1007/978-3-319-46478-7_28.

[56] YU Dong-fei, FU Jian-long, MEI Tao. Multi-level attention networks for visual question answering [C]// Computer Vision and Pattern Recognition. USA: IEEE Press, 2017: 4187−4195. DOI: 10.1109/CVPR.2017.446.

[57] JANG Y, SONG Y, YU Y, KIM Y, KIM G. TGIF-QA: Toward spatio-temporal reasoning in visual question answering [C]// IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Computer Society, 2017: 1359−1367. DOI: 10.1109/CVPR.2017.149.

[58] LU J, YANG Jian-wei, BATRA D, PARIKH D. Hierarchical question-image co-attention for visual question answering [C]// Advances in Neural Information Processing Systems. Spain: Curran Associates Inc, 2016: 289−297.

[59] LIANG Jun-wei, JIANG Lu, CAO Liang-liang, LI Li-Jia, HAUPTMANN A. Focal visual-text attention for visual

question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 6135−6143. DOI: 10.1109/TPAMI. 2018.2890628. DOI: 10.1109/TPAMI. 2018.2890628.

[60] NGUYEN D K, OKATANI T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 6087−6096. DOI: 10.1109/ CVPR.2018.00637.

[61] PATRO B, NAMBOODIRI V P. Differential attention for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 7680−7688. DOI: 10.1109/CVPR. 2018.00801.

[62] CADENE R, BEN Y H, CORD M, THOME N. MUREL: Multimodal relational reasoning for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 1989−1998. DOI: 10.1109/CVPR.2019.00209.

[63] XU K, BA J, KIROS R, CHOK, COURVILLEA, SALAKHUTDINOV R, ZEMEL R，BENGIOY. Show, Attend and tell: Neural image caption generation with visual attention [EB/OL] [2015-02-10]. https://arxiv.org/abs/1502. 03044v3.

[64] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation [J]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015. DOI:10.18653/v1/d15-1166.

[65] AUER S, BIZER C, KOBILAROV G, LEHMANN J, CYGANIAK R, IVES Z. DBpedia: A nucleus for a web of open data [J]. Semantic Web, 2007, 4825(2): 11−15. DOI: 10.1007/978-3-540-76298-0_52.

[66] BOLLACKER K, EVANS C, PARITOSH P, STURGE T, TAYLOR J. Freebase: A collaboratively created graph database for structuring human knowledge [C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Stanford University: ACM, 2008: 1247−1250. DOI: 10.1145/1376616.1376746.

[67] LIU H, SINGH P. Concept Net—A Practical Commonsense Reasoning Tool-Kit [J]. BT Technology Journal, 2004, 22(4): 21−33. DOI: 10.1023/B: BTTJ.0000047600.45421.6d.

[68] ZHU Yu-ke, ZHANG Ce, RÉ C, LI Fei-fei. Building a large-scale multimodal knowledge base system for answering visual queries [EB/OL] [2015-05-20]. https://arxiv. org/abs/1507.05670v1.

[69] TANDON N, MELO G D, SUCHANEK F. WebChild: Harvesting and organizing commonsense knowledge from the web [C]// ACM International Conference on Web Search and Data Mining. New York: ACM, 2014: 523−532. DOI: 10.1145/2556195.2556245.

[70] DONG X, GABRILOVICH E, HEITZ G, HORN W. Knowledge vault: A web-scale approach to probabilistic knowledge fusion [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 601−610. DOI: 10.1145/2623330.2623623.

[71] CHEN Xin-lei, SHRIVASTAVA A, GUPTA A. NEIL: Extracting visual knowledge from web data [C]// IEEE International Conference on Computer Vision. USA: IEEE Press, 2014: 1409−1416. DOI:10.1109/ICCV.2013.178.

[72] KUMAR A, IRSOY O, ONDRUSKA P, IYYER M, BRADBURY J, GULRAJANI I, ZHONG V, PAULUS R, SOCHER R. Ask me anything: Dynamic memory networks for natural language processing [C]// International Conference on Machine Learning. New York: ICML, 2016: 1378−1387.

[73] SHEN Chun-hua, DICK A, WU Qi, WANG Peng, HENGEL V A. Explicit knowledge-based reasoning for visual question answering [EB/OL] [2015-11-29]. https://arxiv.org/abs/1511. 02570.

[74] WANG Peng, WU Qi, SHEN Chun-hua, DICK A, HENGEL V A. FVQA: Fact-based visual question answering [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(10): 2413−2427.

[75] SU Zhou, ZHU Chen, DONG Yin-peng, CAI Dong-qi, CHEN Yu-rong, LI Jiang-guo. Learning visual knowledge memory networks for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2018: 7736−7745. DOI: 10.1109/CVPR.2018.00807.

[76] TENEY D, LIU Ling-qiao, VAN D H A. Graph-structured representations for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2017: 1−9. DOI:10.1109/CVPR.2017.344.

[77] SHIN A, USHIKU Y, HARADA T. The color of the cat is gray: 1 million full-sentences visual question answering [EB/OL] [2016-09-07]. https://arxiv.org/abs/1609.06657.

[78] TANG Kai-hua, ZHANG Han-wang, WU Bao-yuan, LUO Wen-han, LIU Wei. Learning to compose dynamic tree structures for visual contexts [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 6619−6628. DOI: 10.1109/CVPR. 2019.00678. DOI: 10.1109/CVPR. 2019.00678.

[79] WORTSMAN M, EHSANI K, RASTEGARI M, FARHADI A, MOTTAGHI R. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 6750−6759. DOI: 10.1109/CVPR.2019.00691.

[80] NOH H, KIM T, MUN J, HAN B. Transfer learning via unsupervised task discovery for visual question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 8385−8394. DOI: 10.1109/CVPR.2019.00858.

[81] WANG Zhong-dao, ZHENG Liang, LI Ya-li, WANG Sheng-jin. Linkage based face clustering via graph convolution network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 1117−1125. DOI: 10.1109/ CVPR.2019.00121.

[82] SINGH A, NATARAJAN V, SHAH M, JIANG Yu, CHEN Xin-lei, BATRA D, PARIKH D, ROHRBACH M. Towards VQA models that can read [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 8317−8326. DOI: 10.1109/CVPR.2019.00851.

[83] ZHOU Bo-yuan, CUI Quan, WEI Xiu-shen, CHEN Zhao-min. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2020: 9716−9725. DOI: 10.1109/CVPR42600.2020.00974.

[84] SHEPARD R N. Toward a universal law of generalization for psychological science [J]. Science, 1988, 242(4880): 944−944. DOI: 10.1126/science.242.4880.944b.

[85] SUKHBAATAR S, WESTON J, FERGUS R. End-to-end memory networks [C]// Advances in Neural Information Processing Systems. Canada: Curran Associates Inc, 2015: 2440−2448. DOI: 10.7551/mitpress/1120.003.0035.

[86] KUMAR A, IRSOY O, ONDRUSKA P, IYYER M, BRADBURY J, GULRAJANI I, ZHONG V, PAULUS R, SOCHER R. Ask me anything: Dynamic memory networks for natural language processing [C]// International Conference on Machine Learning. New York: ICML, 2016: 1378−1387.

[87] XIONG Cai-ming, MERITY S, SOCHER R. Dynamic memory networks for visual and textual question answering [C]// International Conference on Machine Learning. New York: ICML, 2016: 2397−2406.

[88] SANTORO A, BARTUNOV S, BOTVINICK M, WIERSTRA D, LILLICRAP T. Meta-learning with memory-augmented neural networks [C]// International Conference on Machine Learning. New York: ICML, 2016: 1842−1850.

[89] GUO Da-lu, XU Chang, TAO Da-cheng. Image-question-answer synergistic network for visual dialog [J]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 10434−10443. DOI: 10.1109/CVPR.2019.01068.

[90] WANG Xin, HUANG Qiu-yuan, CELIKYILMAZ A, GAO jian-feng, SHEN Ding-han, WANG Yuan-Fang, WANG W Y, ZHANG Lei. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 6630−6638. DOI: 10.1109/CVPR.2019.00679.

[91] KE Li-yi-ming, LI Xiu-jun, BISK Yonatan, HOLTZMAN A, GAN Z, LIU Jing-jing, GAO Jian-feng, CHOI Ye-jin, SRINIVASA S. Tactical rewind: Self-correction via backtracking in vision-and-language navigation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2019: 6741−6749. DOI: 10.1109/CVPR.2019.00690.

[92] REN Meng-ye, KIROS R, ZEMEL R. Image question answering: A visual semantic embedding model and a new dataset [J]. Litoral Revista De La Poesía Y El Pensamiento, 2015(6): 8−31.

[93] AGRAWAL A, LU Jia-sen, ANTOL S, MITCHELL M. VQA: Visual question answering: www. visualqa. org [J]. International Journal of Computer Vision, 2016, 123(1): 12−24. DOI: 10.1007/s11263-016-0966-6.

[94] ZHU Yu-ke, GROTH O, BERNSTEIN M, Li Fei-fei. Visual7w: Grounded question answering in images [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Press, 2016: 4995−5004. DOI: 10.1109/CVPR.2016.540.

[95] KRISHNA R, ZHU Yu-ke, GROTH O, JOHNSON J, HATA K, KRAVITZ J, CHEN S, KALANTIDIS Y, LI LI-JIA, SHAMMA A D, BERNSTEIN S M, LI Fei-fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations [J]. International Journal of Computer Vision, 2017, 123(1): 32−73. DOI: 10.1007/s11263-016-0981-7.

[96] YU Li-cheng, PARK E, BERG A C, BERG T L. Visual Madlibs: Fill in the blank description generation and question answering [C]// IEEE International Conference on Computer Vision. USA: IEEE Computer Society, 2016: 2461−2469. DOI: 10.1109/ICCV.2015.283.

[97] ZHANG Peng, GOYAL Y, SUMMERSSTAY D, PARIKH D. Yin and Yang: Balancing and answering binary visual questions [C]// Computer Vision and Pattern Recognition. USA: IEEE Computer Society, 2016: 5014−5022.

[98] ANDREAS J, ROHRBACH M, DARRELL T, KLEIN D. Learning to compose neural networks for question answering [EB/OL] [2016-01-07]. https://arxiv.org/abs/1601.01705.

[99] GAO Hao-yuan, MAO Jun-hua, ZHOU Jie, HUANG Zhi-heng, WANG Lei, XU Wei. Are you talking to a machine? Dataset and methods for multilingual image question answering [J]. Computer Science, 2015, 27(28): 2296−2304.

[100] MANMADHAN S, KOVOOR B C. Visual question answering: A state-of-the-art review [J]. Artificial Intelligence Review, 2020, 53(8): 5705−5745. DOI: 10.1007/s10462-020-09832-7.

(Edited by HE Yun-bin)

中文导读

基于深度学习的视觉问答研究综述

摘要：随着机器学习特别是深度学习的兴起和不断发展，对视觉问答领域的研究取得了重大进展，具有重要的理论研究意义和实际应用价值。因此，有必要对目前的研究进行总结，为该领域的研究者提供一些参考。本文对视觉答疑领域的相关研究和典型方法进行了详细而深入的分析和总结。首先，介绍了视觉问答的相关背景知识。其次，讨论了视觉答疑存在的问题和挑战，并对具体的方法进行探讨。第三，对影响视觉答疑的关键子问题进行了总结和分析。然后，对目前常用的数据集和评价指标进行了总结。针对 VQA 研究中流行的算法和模型，对各种算法和模型进行了总结和比较。最后，对视觉答疑的发展趋势和结论进行了展望。

关键词：计算机视觉；自然语言处理；视觉问答；深度学习；注意力机制