PeerJ Computer Science

The multi-modal fusion in visual question answering: a review of attention mechanisms

Siyu Lu[1], Mingzhe Liu[2], Lirong Yin[3], Zhengtong Yin[4], Xuan Liu[5] and Wenfeng Zheng[1]

[1] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan Province, China
[2] School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou, China
[3] Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA, United States of America
[4] College of Resource and Environment Engineering, Guizhou University, Guiyang, China
[5] School of Public Affairs and Administration, University of Electronic Science and Technology of China, Chengdu, China

ABSTRACT

Visual Question Answering (VQA) is a significant cross-disciplinary issue in the fields of computer vision and natural language processing that requires a computer to output a natural language answer based on pictures and questions posed based on the pictures. This requires simultaneous processing of multimodal fusion of text features and visual features, and the key task that can ensure its success is the attention mechanism. Bringing in attention mechanisms makes it better to integrate text features and image features into a compact multi-modal representation. Therefore, it is necessary to clarify the development status of attention mechanism, understand the most advanced attention mechanism methods, and look forward to its future development direction. In this article, we first conduct a bibliometric analysis of the correlation through CiteSpace, then we find and reasonably speculate that the attention mechanism has great development potential in cross-modal retrieval. Secondly, we discuss the classification and application of existing attention mechanisms in VQA tasks, analysis their shortcomings, and summarize current improvement methods. Finally, through the continuous exploration of attention mechanisms, we believe that VQA will evolve in a smarter and more human direction.

Subjects Artificial Intelligence, Computer Vision
Keywords VQA, Visual question answering, Multi-modal, Fusion, Attention mechanisms, Attention

INTRODUCTION

With the rapid development of deep learning and big data techniques, quantities of remarkable achievements have been made in areas such as computer vision (*Ren et al, 2017*; *Ronneberger, Fischer & Brox, 2015*) and natural language processing (*Bahdanau, Cho & Bengio, 2015*; *Luong, Pham & Manning, 2015*), These milestones have all enabled the ability to recognize a single modality. However, numerous implementations within the field of artificial intelligence involve multiple modalities. Modality, initially, refers to a

Submitted 4 August 2022
Accepted 25 April 2023
Published 30 May 2023

Corresponding authors
Lirong Yin, yin.lyra@gmail.com
Wenfeng Zheng, winfirms@uestc.edu.cn, wenfeng.zheng.cn@gmail.com

Academic editor
Xiangjie Kong

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj-cs.1400**

biological concept. For example, human beings can receive information from the outside world by virtue of their sensory organs and experiences, including vision, hearing, touch, etc. Each of these forms of information can be called a modality. Multi-modal interaction refers to the integration of information from various senses, thus making it easy for people to communicate with the outside world.

Image caption task (*Anderson et al., 2018*; *Chen et al., 2017*; *Xu et al., 2015*) first combines computer vision with natural language processing. Similarly, VQA is a dynamic multidisciplinary field, which attracts increasing interest and has extraordinary potential (*Wu et al., 2017*; *Antol et al., 2015*). Given a picture and a natural language question including which, when, where, who, what, how and why (*Zhu et al., 2016*) that is correlative with this picture, the system need to produce a natural language answer as an output. It is clear that this is a typical difficult multi-modal task that combines computer vision (CV) and natural language processing (NLP) as two major techniques, and its main objective is for the computer to produce an answer that complies with natural language rules and has reasonable content based on the input image and question. In this task, computers need to understand and integrate information from multiple modalities, such as text features and image features, and build models that can deal with and correlate information from multiple modalities, which is obviously very challenging. It has a wide range of applications and is crucial in many real-world situations, such as medical research (*Zheng et al., 2020*; *Pan et al., 2022*; *Li et al., 2022*; *Gong et al., 2021*; *Zhan et al., 2020*; *Wang et al., 2022a*; *Wang et al., 2022b*), remote sensing (*Bazi et al., 2022*; *Zheng et al., 2022b*; *Al Rahhal et al., 2022*), and blindness help (*Gurari et al., 2018*; *Tung et al., 2021*; *Tung, Huy Tien & Minh Le, 2021*). Therefore, it is of great significance to improve existing VQA methods.

The algorithms in visual question answering can be broadly divided into three steps: extracting features from images, extracting features from questions, and finally combining image and text features to generate answers. The third of these steps, which is how to better fuse image features with text features, is the core problem of the VQA task (*Liu et al., 2019*; *Gao et al., 2019*). As a fundamental multi-modal task, VQA requires a comprehension of both the problem's constituent structure and the complex substance in the image, and should also have the ability to correctly extract and integrate information from both the image and the text, that's what makes it so challenging. Therefore, the study of cross-modal modeling methods is of broad interest, and how to better integrate text features and image feature sets into a compact multi-modal representation has become an important research topic.

In recent studies, the difference in algorithms was mainly in how to combine the features of both to do the processing. Most of the existing methods focus on learning the joint representation of image and question. Under the framework of such methods, question and image are usually encoded with RNN (LSTM most commonly used) and CNN (*Selvaraju et al., 2020*; *Gao et al., 2015*; *Malinowski et al., 2015*), respectively, and then their representation vectors are input into the multimode fusion unit. The aligned joint representation vector is obtained by training the fusion unit, and finally the representation vector is input into a neural network to get the final answer, as shown in Fig. 1A. However, this is actually a simple and straightforward combination of features that cannot correctly

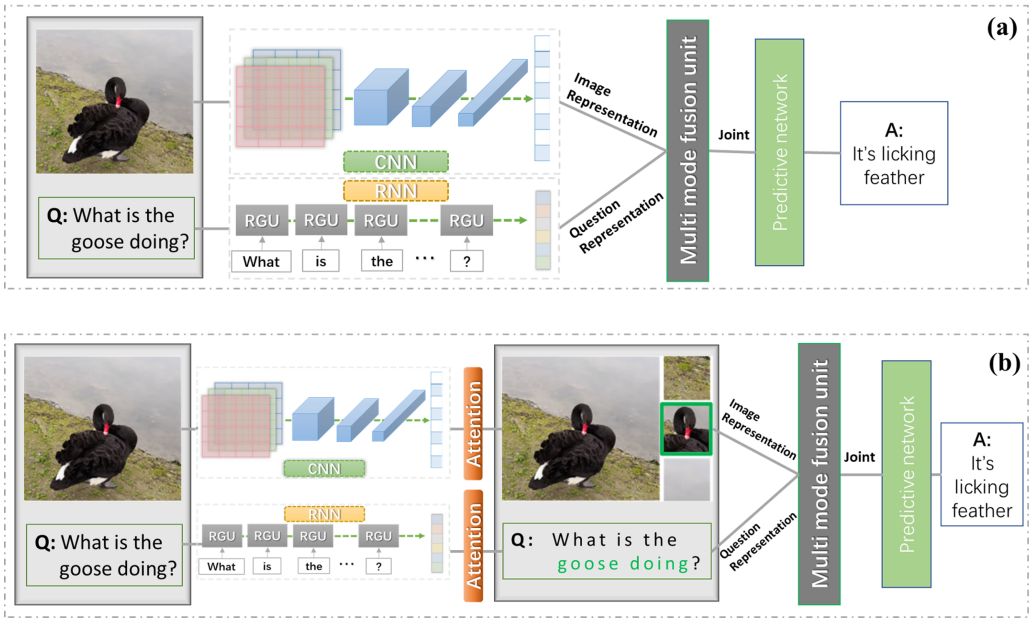

Figure 1 **VQA task with/without attention mechanism.**
Full-size DOI: 10.7717/peerjcs.1400/fig-1

discriminate what is really useful information, the concerns are all global in character, making it difficult to correctly infer the answer and greatly limiting the performance of VQA. Therefore, in order to make VQA models better grasp the problem and identify the key visual content, scholars are now more committed to studying attention mechanisms.

It can be said that attention models have become almost ubiquitous in deep neural networks. The attention mechanism is inspired by the human attention mechanism, in which people tend to selectively observe and pay attention to specific parts of information while ignoring the rest, as needed. That is, the observation can be adjusted to the more informative features according to their relative importance, focusing the algorithm on the most relevant parts of the input, moving from focusing on global features to the focused features, thus saving resources and getting the most effective information quickly. The attention mechanism has arguably become one of the most important concepts in the field of deep learning, since *Bahdanau, Cho & Bengio (2015)* used attention mechanism for the machine interpretation tasks, various variants of attention mechanism have emerged, such as Co-Attention networks (*Yang et al., 2019a*; *Han et al., 2021*; *Yu et al., 2019*; *Liu et al., 2021b*; *Lu et al., 2016*; *Sharma & Srivastava, 2022*), Recurrent Attention networks (*Osman & Samek, 2019*; *Ren & Zemel, 2017*; *Gan et al., 2019*), Self-Attention networks (*Li et al., 2019*; *Fan et al., 2019*; *Ramachandran et al., 2019*; *Xia et al., 2022*; *Xiang et al., 2022*; *Yan, Silamu & Li, 2022*), *etc.* The effectiveness of visual information processing is considerably enhanced by all of these attention mechanisms, which also optimize VQA performance.

Its advantages can be summarized in two simple points:

(1) Improvement of computational power: Because of remembering a large amount of information, the model becomes complex, and at present, the computational power remains an important factor limiting the neural network.

(2) Simplification of the model: It can make the neural network a bit simpler and effectively alleviate the contradiction between model complexity and expressive power.

While most previous work has reviewed the entire field of VQA, the focal point of this article is to give an overview of "attention mechanisms" in multi-modal fusion methods for VQA tasks. Nowadays, with the rapid development of information technology, multi-modal data has become the main form of data resource in recent years. Therefore, the study of multi-modal learning development can empower computers to understand diverse data, so this research topic has an important value: how to better utilize the information of multiple modalities is of great significance to the study of deep learning. At the same time, multi-modal data can also provide more information for model decision making, thus improving the overall accuracy of the decision. A thorough research review can be produced through the analysis and evaluation of its existing research. This article can help pertinent researchers to have a more comprehensive understanding of the attention mechanism in multimodal fusion methods.

The study is organized as follows: in Section 2, we describe how we comprehensively collect relevant literature and provide a comprehensive analysis of VQA using bibliometric methods, including analysis of keywords, countries and institutions, and timelines; in Section 3, We describe the general model of attention mechanism, and classify the attention mechanisms commonly used in VQA tasks from three dimensions; in Section 4, We describe the application of attention mechanism from two aspects; in Section 5, we analyzed and compared the attention mechanism used by some classical frontier models, and summarized the development trend of attention mechanism in VQA; in Section 6, we conclude the article and prospect the possible future directions.

SURVEY METHODOLOGY

Bibliometric methods are used to comprehensively analyze the current state of attention mechanisms in VQA and conclude the article with a prospect of possible future directions. A total of 420 eligible articles in the database of Web of Science (SCIE only, since 2011) with "Visual Question Answering" and "attention mechanism" as the subject line were retrieved and screened. The retrieved articles were saved as plain text files, counting full records and cited references, and then we used the "Data/Import/Export" function of CiteSpace (*Rawat & Sood, 2021*) to convert them into an executable format and visualize the retrieved records. A visual analysis of the collected literature has been made. We selected the top 50 levels and 10.0% of most cited or occurred items from each slice, and tick at minimum spanning tree in pruning. This study aims to systematically review the development history and research status of VQA and attention mechanism with a bibliometric approach, explore the hotspots in the field, and reasonably predict the possible future directions.

Analysis of keywords

CiteSpace is used to perform keyword co-occurrence analysis on the above articles, setting the time slice as January 2010 to December 2022, one year per slice, and choosing "keyword" for the node type. The figure shows keywords with more than seven occurrences.

The size of the nodes and labels indicates the frequency of the keyword. From the figure, it can be found that "Visual Question Answering"(120), "attention mechanisms"(43), "deep learning"(35), "task analysis"(34), "video question answering"(25), "knowledge discovery" (24) "computer vision"(20) appeared most frequently.

The cluster view tells us the current research frontier topics. From the Fig. 2, we can infer that attention mechanisms, semantic analysis, visual question answering, and multi-modal learning are the current research hotspots.

Analysis of leading collaborating countries and institution

From 2010 to 2022, research institutions from 37 countries are all making irreplaceable contributions in this field. Figures 3 and 4 show the visualization of these countries and institutions.

In Fig. 3 there are a total of 37 nodes, and the larger the area of the circle of the node, the greater the number of publications from that country, with the color indicating the year of citation. As we can see from the figure, the five countries that contribute the most regarding the number of distributions are China, the United States, India, Australia, and South Korea. Undoubtedly, Chinese researchers have played an important role in the research in this field, with 242 publications by Chinese scholars accounting for 57.619% of the total number of publications (420). The United States ranks second, with 100 articles contributing to the field, accounting for 23.810%. In addition, as an example of emerging research countries, India, Australia and South Korea have also made significant contributions in the last decade. Australia ranked third with 7.143%. Other high-producing countries include Australia (5.714%), South Korea (4.762%), and others.

As we can see in Fig. 4, the five institutions that contribute the most to the number of publications are Chinese Academy of Science, Zhejiang University, University of the Chinese Academy of Sciences, Zhejiang University, University of Electronic Science and Technology of China, and Beihang University, which also shows once again that Chinese researchers are playing an irreplaceable role in research in this field.

China is the most populous country in the world, and there are many universities and research institutions. At present, AI is developing rapidly. As one of the applications of intelligent systems, VQA has attracted more Chinese researchers and scientists, focusing on this frontier field and promoting research progress in related fields (*Guo & Han, 2022*; *Guo & Han, 2023*; *Miao et al., 2022b*; *Miao et al., 2022a*; *Peng et al., 2022a*; *Shen et al., 2022*; *Liu et al., 2022a*).

Analysis of timeline

Figure 5 shows the timeline of keywords. Table 1 shows the clustering result of keywords. Each line represents a cluster, and every single line from shows the evolvement of keywords over time. Words like "visual question answering", "attention module" and "deep learning"

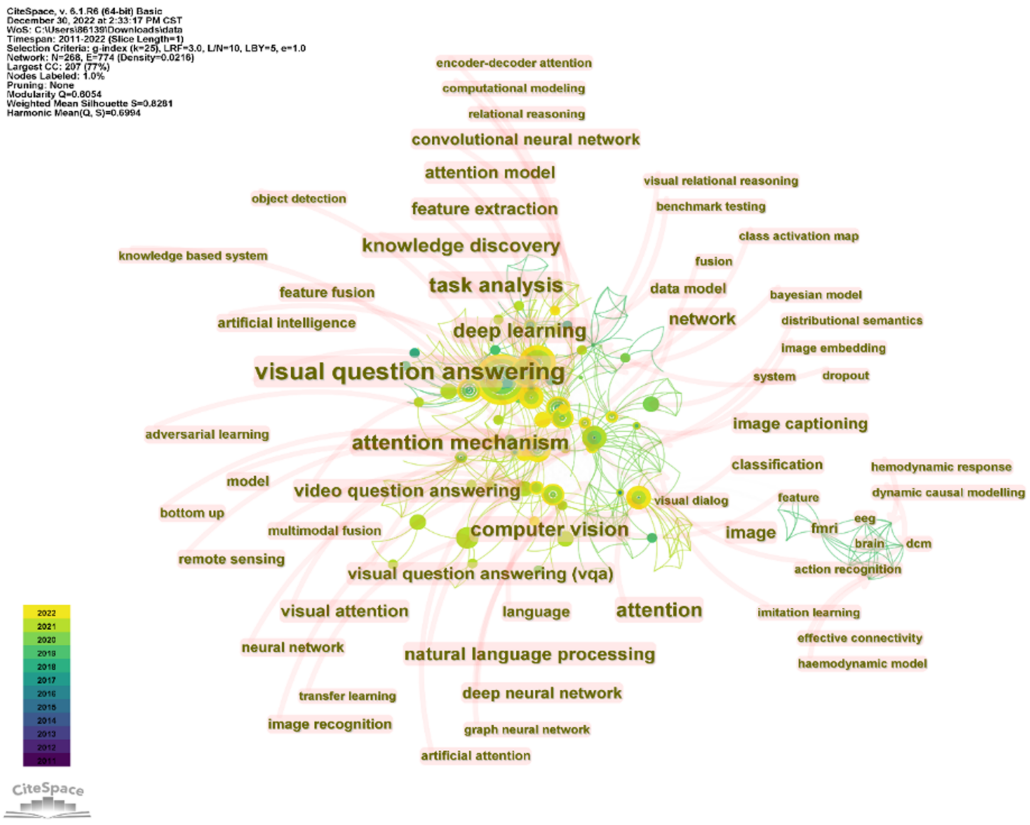

Figure 2 **Keywords co-occurrence.**

Full-size DOI: 10.7717/peerjcs.1400/fig-2

appears in 2016. This shows that researchers began to apply attention mechanism to VQA tasks gradually in 2016. So far, we have screened a total of 420 articles, indicating that the current research in this area is not saturated, and there are still many methods to be explored. In 2020, nodes have increased a lot in almost every time line, indicating that research in this field has been developing rapidly since 2020. As the largest node in 2020, knowledge discovery may become a research hotspot in the future. It can be seen that over time, people's research in this field has been gradually subdivided and deepened, and more models and methods have been applied to solve the problem of VQA. VQA research has also penetrated into different aspects, such as medical and remote sensing.

CLASSIFICATION OF ATTENTION MECHANISMS IN VQA

In this chapter, we describe the general mechanism of attention and categorize attention by scope, excitation, and dimension.

General attention mechanism

From the mathematical perspective, the attention mechanism simply assigns different weighting parameters to input items according to their importance, but this mechanism simulates the cognitive model of the human brain, that is, focusing limited attention on

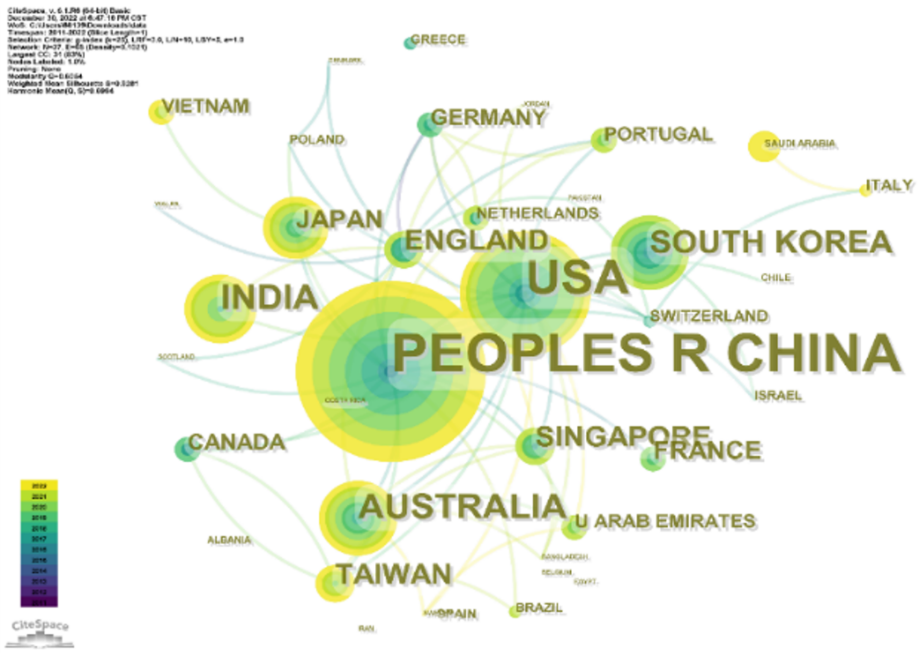

Figure 3 **Leading collaborating countries.**

Full-size DOI: 10.7717/peerjcs.1400/fig-3

the key parts of things according to actual needs, thus greatly enhancing the understanding ability of neural networks.

Suppose that the constituent elements in the source are composed of a series of key value data pairs. At this time, given an element Query in the target, calculate the similarity or correlation between Query and each key to obtain the weight coefficient of the corresponding value of each key, and then add weights to sum the values to obtain the final attention value (*Vaswani et al., 2017*). Therefore, essentially, the attention mechanism is a weighted sum of the value values of the elements in the source, while Query and Key are used to calculate the weight coefficients of the corresponding values. The keys and values are also packed into K and V. The calculation of attention is generally divided into 3 steps (*Chaudhari et al., 2021*), as shown in Fig. 6.

(1) The first step is to calculate the similarity between Query and different keys, that is, to calculate the weight coefficients of different Value values; There are mainly three ways to calculate similarity.

$$\text{Similarity}_{\text{Dotproduct}}(\text{Q},\text{K}) = \text{Q}\cdot\text{K} \tag{1}$$

$$\text{Similarity}_{Cosine}(\text{Q},\text{K}) = \frac{\text{Q}\cdot\text{K}}{\|\text{Q}\|\cdot\|\text{K}\|} \tag{2}$$

$$\text{Similarity}_{MLP}(\text{Q},\text{K}) = MLP(\text{Q}\cdot\text{K}) \tag{3}$$

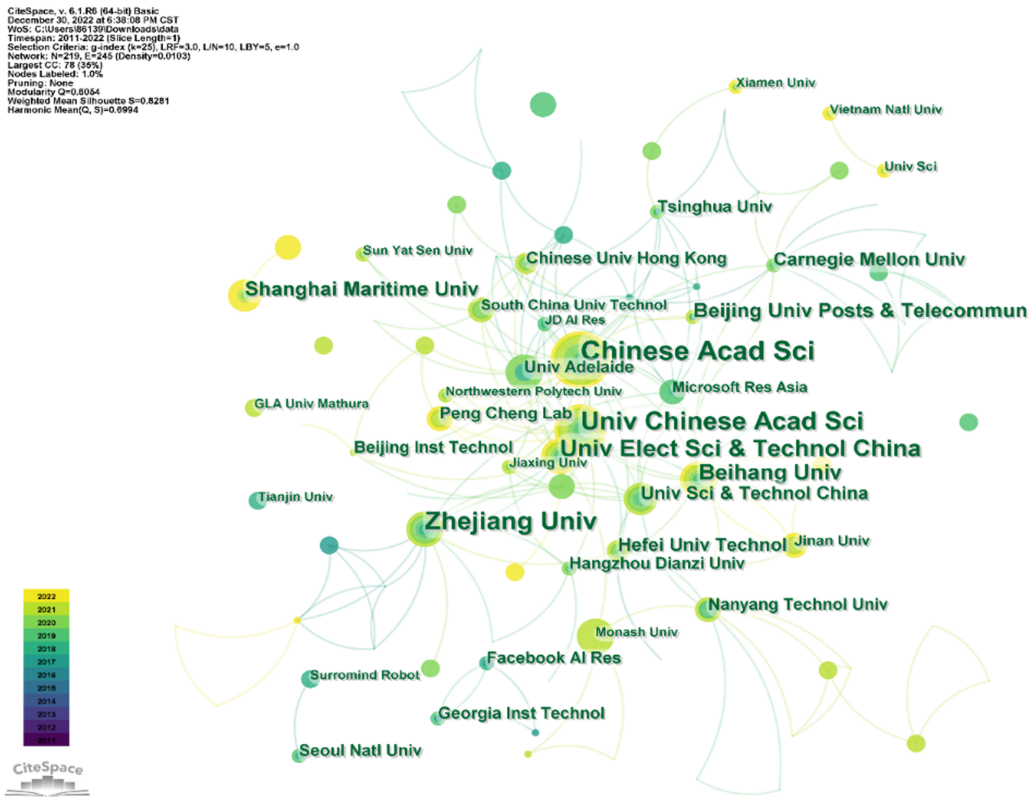

Figure 4 **Leading collaborating institutions.**

Full-size DOI: 10.7717/peerjcs.1400/fig-4

(2) In the second step, the output of the previous stage is normalized to map the range of values between 0 and 1. The most commonly used is softmax.

$$a_i = \text{softmax}\left(\text{Similarity}_i\right) = \frac{e^{\text{Sim}_i}}{\sum_{j=1}^{Lx} e^{\text{Sim}_j}} \tag{4}$$

a_i indicates the degree of attention paid to the ith information.

(3) In the third step, accumulate the result by multiplying the value and the weight corresponding to each value to obtain the attention value.

$$\text{Attention (Query, Source)} = \sum_{i=1}^{Lx} a_i \cdot \text{V}_i \tag{5}$$

Classification of attention mechanisms

We have classified the attention mechanism into three aspects, as shown in Fig. 7.

Scope of attention

According to the scope of attention, attention can be divided into soft attention and hard attention, or local attention and global attention.

Soft and hard attention were proposed by *Xu et al. (2015)* in an image caption task. The soft attention will focus on all the positions of the. Soft attention refers to when selecting

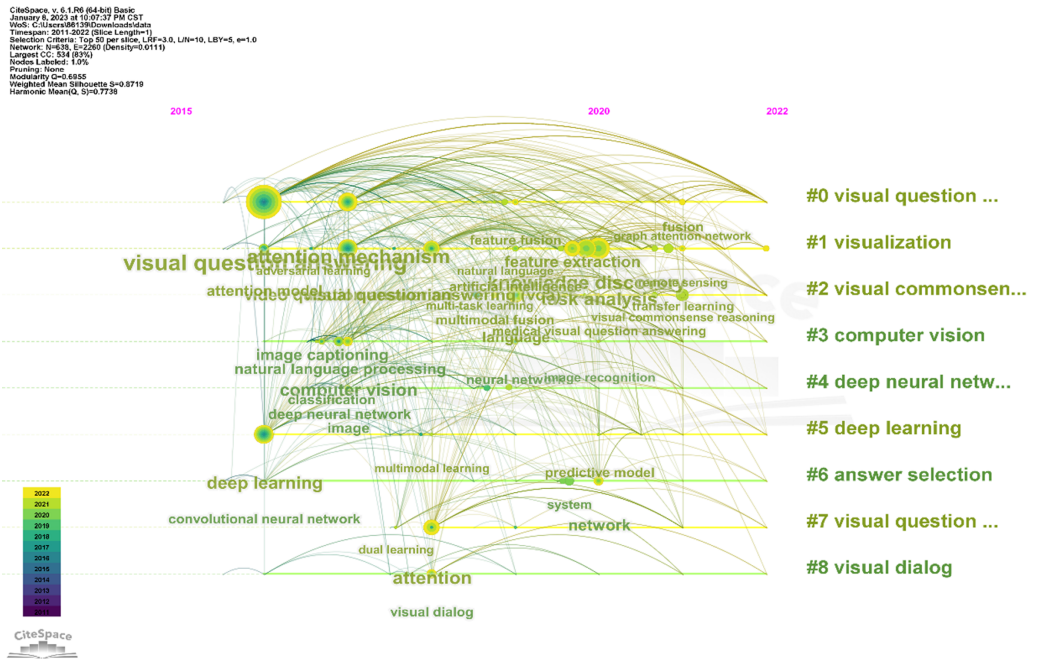

Figure 5 **Timeline of keywords clusters.**

Full-size DOI: 10.7717/peerjcs.1400/fig-5

information, instead of selecting only one of the N information, it calculates the weighted average of the N input information. For image (*Niu, Zhong & Yu, 2021*), it means giving different weights to different positions and then inputting it into the neural network for calculation. This is also similar to Bahdanau attention (*Bahdanau, Cho & Bengio, 2015*).

However, in hard attention, only one image block is considered at a time (*Malinowski et al., 2018*). Soft attention pays more attention to the overall information, which can be differentiated and expensive. Hard attention only focuses on a certain part of the information, which is not differentiable and is not expensive to calculate.

Global attention and local attention were proposed by *Luong, Pham & Manning (2015)* for a language translation task.

The initial global attention defined by *Luong, Pham & Manning (2015)* considers all the hidden states of the encoder LSTM (*Cheng, Dong & Lapata, 2016*) and decoder LSTM to calculate the "variable length context vector", while *Bahdanau, Cho & Bengio (2015)* use the previous hidden states of the unidirectional decoder LSTM and all the hidden states of the encoder LSTM to calculate the context vector. When applying the "global" attention layer, a lot of calculations will be generated. This is because all hidden states must be considered.

Local attention is a mixture of soft attention and hard attention. It does not consider all coded inputs, but only a part of the inputs. This not only avoids the expensive computation caused by soft attention, but also is easier to train than hard attention. Compared with global attention, local attention is much less computationally expensive.

Table 1 Keyword cluster analysis.

Clusters	Ranked terms
#0: visual question answering	parameter-sharing mechanism; local perception; top-down attention; cascading top-down attention | attention mechanism; relational reasoning; neural network; graph convolution; sematic relationship visual; question; answering; mechanism; bilinear attention; vision; answerability; transformer; multi-head
#1: knowledge discovery	task analysis; knowledge discovery; feature extraction; visual question; adversarial learning | visual question answering; attention model; language parsing; deep reasoning; image coding mechanism; sensing; remote; object; understanding graph; deep; bilinear; gnns; modeling
#2: video question answering	video question answering; multi-head attention; referring expression generation; knowledge discovery; multi-level feature fusion | medical visual question answering; transfer learning; data augmentation; abnormality questions; global average feature; character; multi-level; interaction; optical learning; data; modality; multi-task; planes
#3: Computer vision	Computer vision; visual question answering; sparse question self-attention; dual self-guided attention; artificial attention | natural language processing; deep learning; dense co-attention network; distributional semantics; image captioning question; attention; self-guided; sparse; self-attention semantics; knowledge; analysis; task; machine
#4: deep neural networks	deep neural networks; attention mechanisms; age recognition; gender recognition; neural network | neuromorphic computing; deep learning; artificial neural networks; spiking neural networks; simulated annealing bidirectional; literature; convolutional; transformers; self-supervision spiking; learning; neuromorphic; computing; benchmark
#5: deep learning	deep learning; visual question answering; short-term memory; mood detection; attention model | task analysis; visual question; prediction algorithms; relational reasoning; graph matching attention detection; memory; long; short-term; mood feature; description; audio; encoder–decoder; soundnet
#6: dynamic causal modelling	answer selection; attention mechanism; convolutional neural network; bidirectional lstm; siamese network | text analysis; feature construction; answer recommendation; community question answering; feature encoding answer; feature; answering; construction; biologically semantic; common; hierarchical; compositionality; knowledge
#7: dual learning	dual learning; visual question answer; policy gradient; bidirectional encoder representation; question popularity | visual question generation; attention mechanism; reasoning models; visual question answering; reinforcement learning knowledge; diffusion; network; popularity; forum bidirectional; representation; feature; multi-modal; transformers
#8: visual dialog	visual dialog; attention network; visual reference; multimodal semantic interaction; textual reference | question answering; focal attention; photo albums; vision-language understanding; heterogeneous information visual; reference; multimodal; textual; interaction spectral; reasoning; simple; convolution; dual-perspective

Generation of attention

According to the generation of attention, attention can be divided into bottom-up attention and top-down attention.

Researchers apply hard attention to image caption as bottom-up attention (*Anderson et al., 2018*; *Teney et al., 2018*). The term top-down and bottom-up attention was first proposed in neurological articles. Anderson integrated top down and bottom-up attention and presented a new model of visual question and answer (*Luong, Pham & Manning, 2015*), referring to attention mechanisms driven by non-visual or task-specific context as 'top-down', and purely visual feed-forward attention mechanisms as 'bottom-up'. Bottom up is also called focal attention (*Liang et al., 2019*; *Liang et al., 2018*).

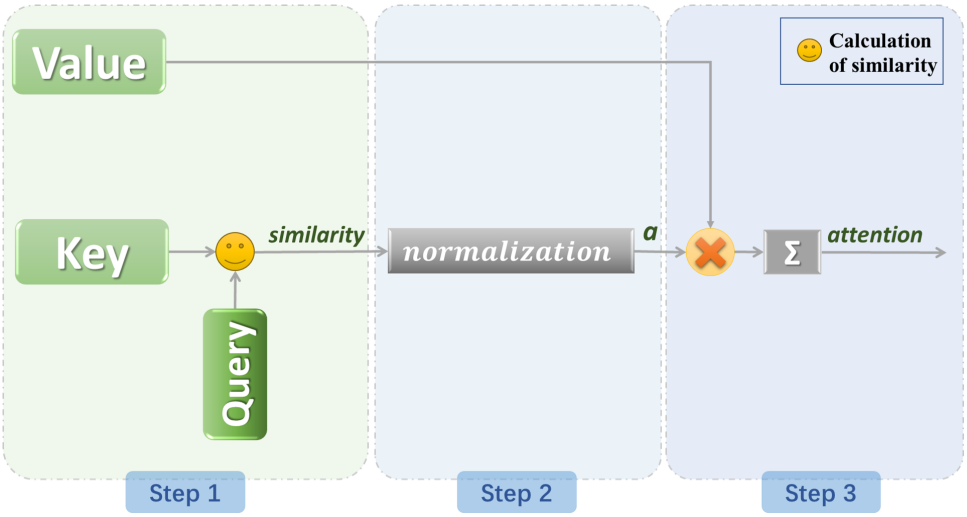

Figure 6 **General model of attention mechanism.**
Full-size DOI: 10.7717/peerjcs.1400/fig-6

The bottom up attention in *Anderson et al. (2018)* is implemented by object detection network Faster R-CNN (*Ren et al., 2017*), which divides the image into specific objects for filtering. Annotate specific elements in a given image by using the Visual Genome dataset. The resulting features can be interpreted as ResNet features centered on the first K objects in the image. That is to say, instead of using the whole picture as the visual feature, the proposal in the first K pictures is selected as the visual feature.

The top-down attention, that is, the problem feature is that after the top is concatenated with the features of each proposal, an attention is obtained through the nonlinear layer and the linear layer, and then the attention is multiplied with the visual feature to get a better feature. Top-down attention is to determine the contribution of features to the text.

Domain of attention

According to the dimension of attention, attention can be divided into spatial (*Yan et al., 2022*), channel and temporal attention (*Guo et al., 2022*) etc.

Spatial attention can be regarded as an adaptive spatial region selection mechanism and answers the question of "where". Channel attention adaptively recalibrates the weight of each channel, which can be regarded as an object selection process and answers the question of "what" or "which". As the name indicates, temporal attention can be used to answer the question of "when".

The existing VQA model usually focuses on the spatial dimension. Considering that attribute features are as important as the area of interest, *Zhang et al. (2021)* constructed an effective joint feature and spatial common attention network (UFSCAN) model for VQA.

The general Spatial visual attention only selects the most concerned visual regions and uses the same weight on the channel, which does not conform to the idea of attention (*Hu et al., 2020*). CNN is naturally characterized by space and channel. In addition, the visual attention model is usually performed at the pixel level, which may lead to the problem

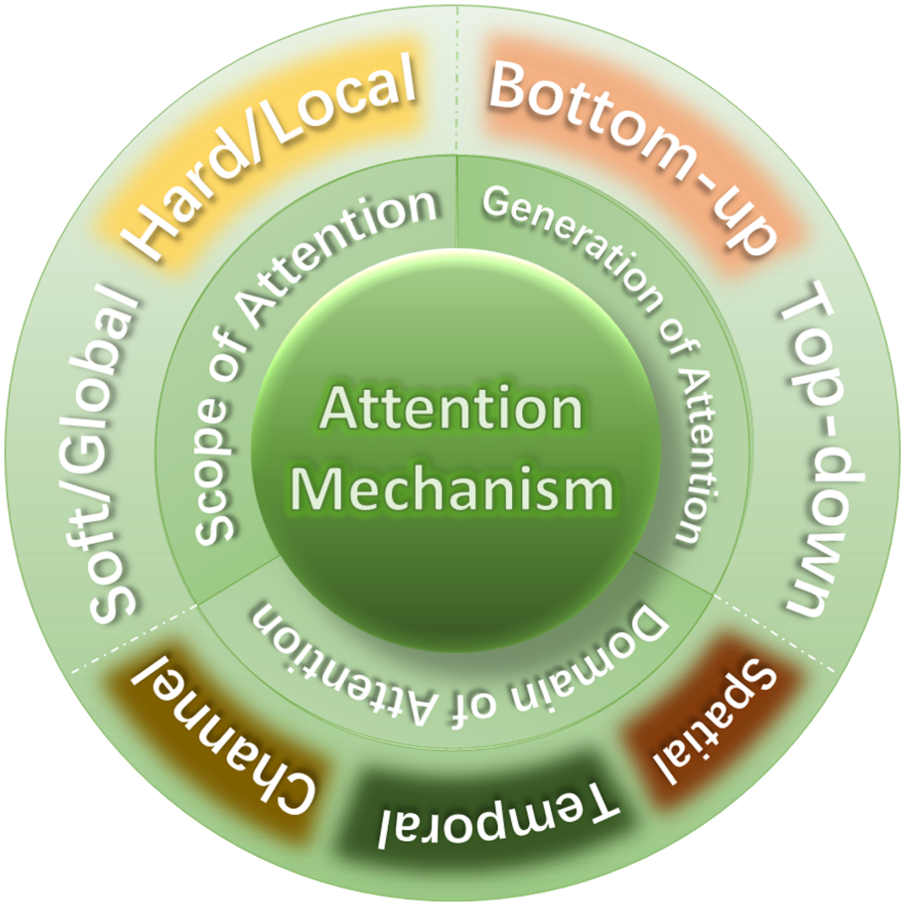

Figure 7 **Classification of attention mechanism.**
Full-size DOI: 10.7717/peerjcs.1400/fig-7

of regional discontinuity. *Song et al. (2018)* proposed a cubic visual attention to select important information and improved VQA tasks by successfully applying new channel and spatial attention to object regions.

For the Question Answering video (*Kossi et al., 2022*; *Liu et al., 2022b*; *Suresh et al., 2022*; *Varga, 2022*), which needs to pay attention to specific frames, or visual dialog (*Guo, Wang & Wang, 2019*; *Kang et al., 2019*; *Patro, Anupriy & Namboodiri, 2022*), which should be able to capture a temporary context from a dialog history, temporal attention is needed (*Yang et al., 2019b*; *Seo et al., 2017*). The existing attention mechanism of VQA mainly focuses on the attention in the spatial area of the image or in a single sequence, so it may not make full use of the properties of multiple sequences and multiple time steps (*Zhao et al., 2017*; *Jiang et al., 2020*; *Chen et al., 2019*; *Shi et al., 2018*). Combining spatial and temporal attention, an encoder decoder model is proposed by *Zhao et al. (2017)* to solve the open video question answering problem.

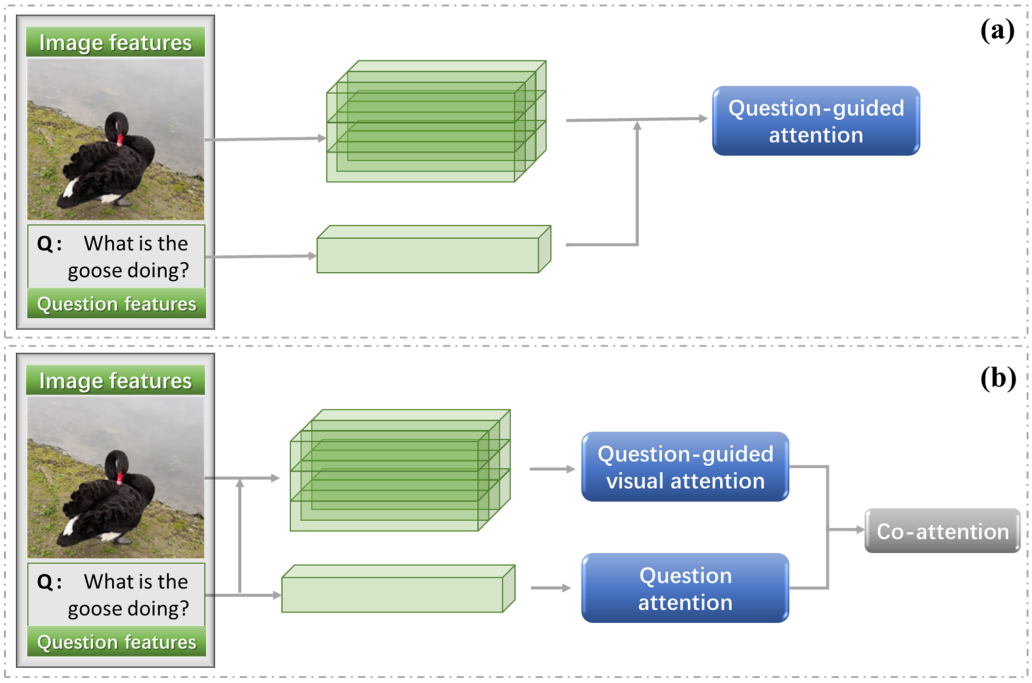

Figure 8 **Question-guided attention (A) and co-attention (B).**
Full-size DOI: 10.7717/peerjcs.1400/fig-8

APPLICATION OF ATTENTION MECHANISMS IN VQA

In this article, we divide the explanation of attention mechanism into problem-directed attention and joint attention, as shown in Fig. 8.

Question-guided visual attention

The initial attention mechanism in VQA tasks considers how to use the question to find the related area in the image. The general practice is to assign weights to image regions according to their relevance to a question (*Liu et al., 2022a*; *Chen et al., 2015*). Figure 8 shows the general model of question guided image attention. The image regions can be divided according to the location or object detection box.

Attention learned from free form regions

A stacking attention network was proposed by *Yang et al. (2016)* that searches for areas in a picture that are related to the response. It uses semantic representations of questions as queries through several iterations. This is the first time an attention mechanism has been added to a VQA task and the results are encouraging and it is a typical spatial attention method. Since then, researchers have focused more and more on employing attention models to identify significant images for efficient response inference. To a large extent, it tends to rely on strong language priority in training questions. The performance on the test image pair is significantly reduced. GVQA (*Agrawal et al., 2018*) is based on SAN and contains inductive bias and restriction in structure, which enables the model to more effectively summarize the answers of different distributions.

Li, Yuan & Lu (2021) proposed a dual-level attention network, using the problem guided attention network to select the image area related to the question, using semantics to highlight the concept related area. More relevant spatial information can be found through the combination of two kinds of attention, and thus reducing the semantic gap between vision and language.

The existing attention research answers questions by focusing on a certain image area, but it is believed that the area of attention mechanism of the existing research is not related to the image area that people will pay attention to *Das et al. (2017)*; *Qiao et al. (2018)*. Therefore, *Patro & Namboodiri (2018)* proposes to obtain a differential attention region through one or more supporting and opposing paradigms. Compared with the image-based attention method, the differential attention calculated in *Patro & Namboodiri (2018)* is closer to human attention.

Attention learned from bounding boxes

The research mentioned in this section use the bounding boxes to select specific objects in the picture, so as to analyze the relationship between different objects (*Ilievski, Yan & Feng, 2016*; *Zhang et al., 2020*; *Zhu et al., 2021*; *Yu et al., 2020*; *Xi et al., 2020*).

For example, although the model can detect objects, backgrounds, *etc.* in the image, it is difficult to understand the semantic information about positions and actions. In order to capture the motion and position information of objects in the image, the model needs to have a more comprehensive understanding of the visual scene in the image by analyzing the dynamic interaction between different objects in the image, not just object detection. One possible method is to align the relative geometric position of objects in the image (for example, the motorcycle is beside the car) with the spatial description information in the text. The other is to capture the dynamic interaction in the visual scene by learning the semantic dependency between objects. Based on this, *Wu et al. (2018a)* introduced a relational attention model. He compares different objects in the image through difference calculation and uses the attention mechanism to filter. *Peng et al. (2019)* devised two question-adaptive relation attention modules that can extract not only the fine-grained and precise binary relations but also the more sophisticated trinary relations.

We discussed the visual attention mechanisms based on the free form region and bounding boxes respectively above. But sometimes these two attention mechanisms can provide complementary information and should be effectively integrated to better solve VQA problems.

Lu et al. (2018a) proposed a new deep neural network for VQA, which integrates two attention mechanisms. The multi-mode multiplicative feature embedding effectively fuses the features of free form image area, detection frame and question representation, so as to answer questions more accurately.

Co-attention mechanisms

In the visual question answering task, if the model wants to correctly answer complex questions, it must have a full understanding of the visual scene in the image, especially the interaction between different objects. Although this kind of method can deal with some

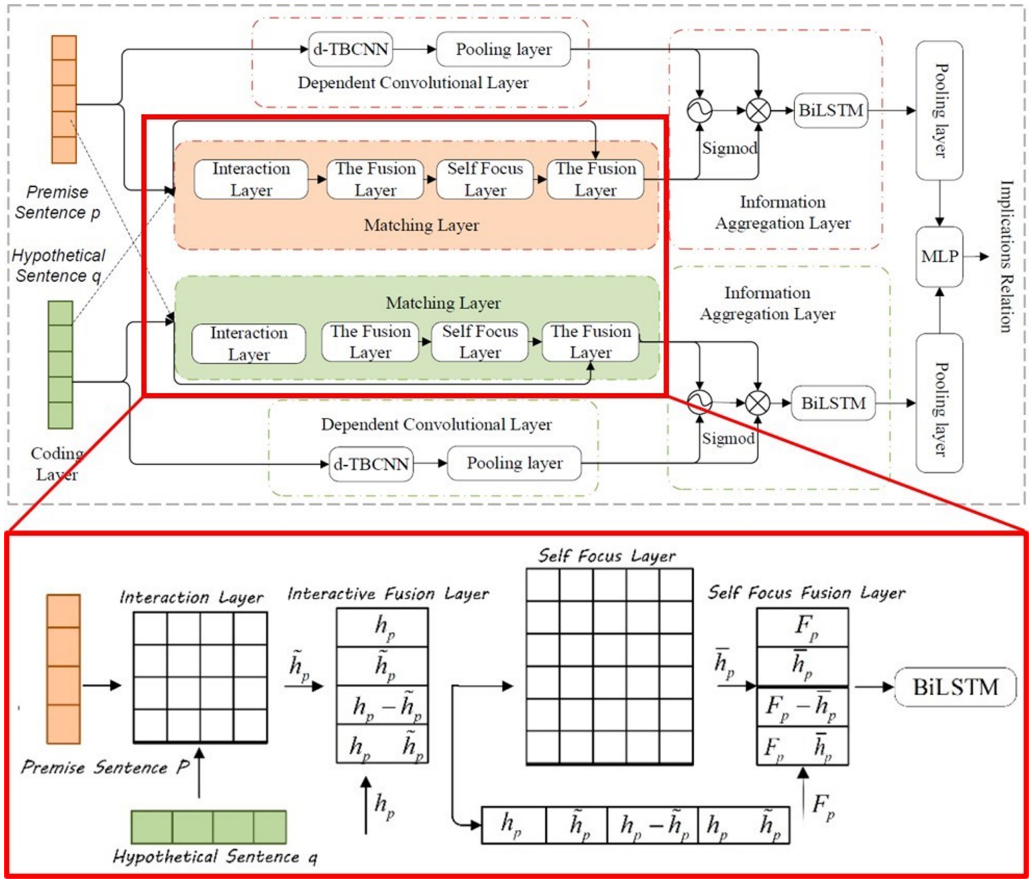

Figure 9 **The structure of SCF-DMN model and the network structure of the matching layer (*Zheng et al., 2022a*).**

Full-size DOI: 10.7717/peerjcs.1400/fig-9

VQA tasks, it still cannot solve the semantic gap between image and question, as shown in Fig. 9. The attention for question also matters (*Burns et al., 2019*; *Nam, Ha & Kim, 2017*; *Yu et al., 2018*).

Kim, Jun & Zhang (2018) proposes the BAN network to use visual and linguistic information as much as possible. BAN uses bilinear interaction in the group of two input channels, and the low rank bilinear pooling extracts the joint representation of the two channels. In addition, the author also proposes a variant of multimodal residual network to make full use of the attention map of BAN network.

Osman & Samek (2019) proposed a recurrent attention mechanism, which uses dual (textual and visual) recurrent attention Units (RAUs). With this model, the researchers demonstrated the impact of all potential combinations of recurrent and convolutional dual attention. *Lu et al. (2018b)* proposed a novel sequential attention mechanism to seamlessly combine visual and semantic clues for VQA.

Gao et al. (2019) proposed a new framework (DFAF) for VQA, which dynamically fuses attention flow within and between modes. DFAF frames alternately transfer information within or across modes according to the flow of attention between and within modes.

Lu et al. (2016) proposed a new concept of collaborative saliency of joint image and text features, which enables two features of different modes to guide each other. In addition, the author also weights the input text information from word level, phrase level and question level to build multiple image question co expression maps at different levels.

Nguyen & Okatani (2018) proposes a new cooperative attention mechanism to improve the fusion of visual and linguistic representation. Given the representation of the image and the question, first generate the attention map on the image area for each question word, and generate the attention map on the problem word for each image area.

The attention mechanism has been improved through self-attention. Typically, self-attention is used as a spatial attention mechanism to capture global information. It is simpler for the model to capture long-distance interdependent elements in a sentence after the introduction of self-attention (*Vaswani et al., 2017*). Because in the case of RNN or LSTM, it requires sequential sequence computation, and for long-distance interdependent features, linking the two requires multiple time steps of information gathering, and the greater the separation, the more difficult it is to successfully capture them. The distance between long-distance dependent features is greatly reduced, however, because of self-attention, which is an attention mechanism that connects elements at various positions of a single sequence and directly connects the connection of any two words in a sentence through one computation step directly during the computation. As a result, it lessens the reliance on external information and is better at capturing the internal relevance of data or features, which can affect the distance between long-distance dependent features (*Chen, Han & Wang, 2020*; *Khan et al., 2022*). In addition, self-attention is also helpful for increasing the parallelism of computation. It just makes up for the shortcomings of the attention mechanism.

Liu et al. (2021b) proposed dual self-attention with co-attention networks (DSACA) for the purpose of capturing the internal dependencies between images and interrogatives and the intrinsic dependencies between different interrogatives and different images. It attempts to fully use the intrinsic correlation between question words and image regions in VQA to describe the internal dependencies of spatial and sequential structures independently using the newly developed self-attentive mechanism to effectively reason with answer correlations. The model is composed of three main submodules: a visual self-attention module to capture the spatial dependencies within images for learning visual representations; a textual self-attention module to integrate the association features between sentence words to selectively emphasize the correlations between interrogative words; and finally, a visual-textual co-attention module to capture the correlations between images and question representations. This approach successfully combines local features and global representation for better representation of text and images, which enormously works on the effectiveness of the model. *Chen, Han & Wang (2020)* proposed a Multimodal Encoder-Decoder Attention Network (MEDAN). The MEDAN consists of Multimodal Encoder-Decoder Attention (MEDA) layers cascaded in depth, and can capture rich and reasonable question features and image features by associating keywords in question with important object regions in image. Each MEDA layer contains an Encoder module modeling the self-attention of questions, as well as a Decoder module modeling the

question-guided-attention and self-attention of images. *Zheng et al. (2022a)* and *Zheng & Yin (2022)* propose an SCF-DMN model that includes a self-attention mechanism, in which a model independent meta-learning algorithm was introduced and a multi-scale meta-relationship network was designed. adds the model-independent meta-learning algorithm and designs a multi-scale meta-relational network, as can be seen in Fig. 9.

Yu et al. (2019) proposed Modular Co-Attention Network (MCAN) to refine and understand both visual and textual content. Inspired by the Transformer model, the author sets two general attention units: a self-attention (SA) unit for modal internal interaction and a guided attention (GA) unit for modal interaction. Then a Modular Co-Attention layer is used to connect the two units in series. Finally, multiple module layers are connected in series. This co-attention mechanism, in the form of co-learning problems and images, can effectively reduce features that are not relevant to the target, exploit the correlation between multi-modal features, and can better capture the complex interactions between multi-modal features, which can improve the performance of VQA models to some extent.

However, behind the powerful ability of the self-attention mechanism, there is a drawback that cannot be ignored: when the model is encoding the information of the current position, it will focus excessively on its own position. To solve this problem, we can improve it by the method of multi-layer attention mechanism.

The multi-layer attention mechanism (*Zheng, Liu & Yin, 2021*), on the other hand, is an evolved version of the single-layer attention mechanism, which allows the model to distill information about the features in the problem from different dimensions by performing multiple operations. This mechanism permits the model to pay joint attention to information coming from various subspaces in various regions, an aspect that is not possible with other types of attention mechanisms.

As the difficulty of the problem increases, more and more VQA models use multiple attention levels to capture deeper visual language associations. *Yu et al. (2017)* proposed a multi-level attention mechanism, including Semantic Attention, Context-aware Visual Attention, and Joint Attention Learning. However, the negative consequences of more layers are parameter explosion and over fitting of the model. Inspired by the capsule network, *Zhou et al. (2019)* proposed a very compact alternative to achieve accurate and efficient attention modeling, called dynamic capsule attention (CapsAtt). CapsAtt regards visual features as capsules, obtains attention output through dynamic routing, and updates the attention weight by calculating the coupling coefficient between the bottom capsule and the output capsule.

Meanwhile, we found that some previous models overly rely on attention mechanisms to associate query words with image content to answer relevant questions. However, they are usually oversimplified to linear transformations, leading to erroneous correspondence between questions and visuals, poor generalization capacity, and possible limitations, and it may not be enough to fully capture the complexity of multi-modal data. In this context, reseachers introduced the method of adversarial learning (*Liu et al., 2018*; *Liu et al., 2021a*; *Sarafianos et al., 2019*; *Wu et al., 2018b*; *Xu et al., 2018*; *Ilievski & Feng, 2017*). *Ilievski & Feng (2017)* suggest an attention mechanism based on adversarial learning in this situation

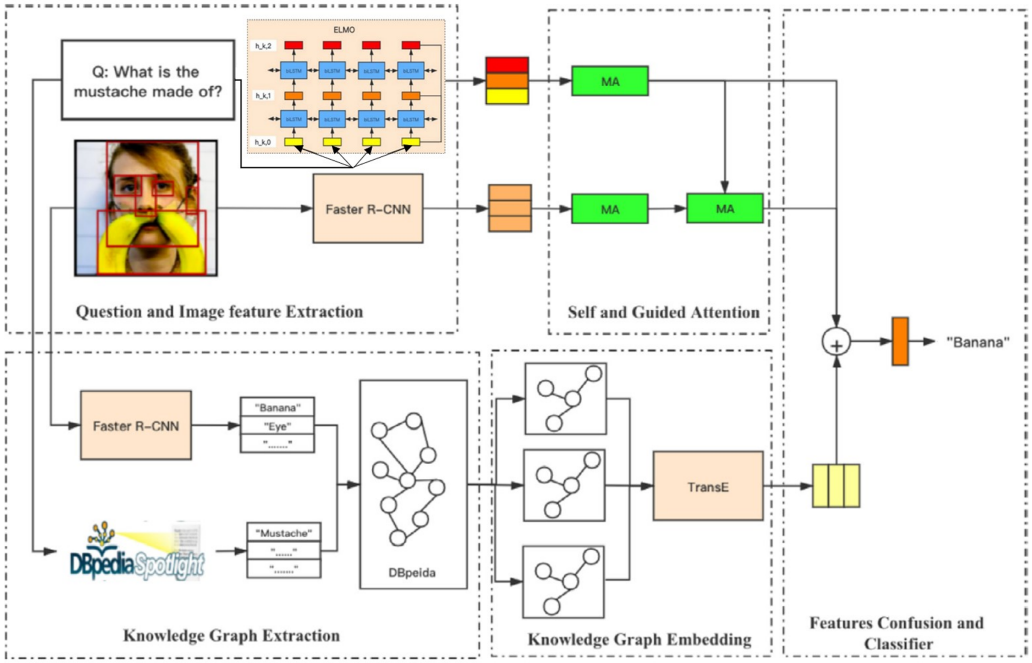

Figure 10 **The infrastructure of N-KBSN model (*Ma et al., 2021*; *Zheng et al., 2021*).**
Full-size DOI: 10.7717/peerjcs.1400/fig-10

that generates more varied visual attention maps and improves the generalization of attention to new issues, leading to improved learning to capture complicated multi-modal data linkages. *Liu et al. (2021a)* proposed a framework based on adversarial learning to learn joint representation to effectively reflect information related to answers. *Ma et al. (2021)* proposed an N-KBSN model that introduced dynamic word vectors based on self-attention and guided attention-based multi-head attention mechanisms, and found that the accuracy of the results exceeded that of the winners (gloves) models of 2017 and 2019, *Zheng et al. (2021)* then introduced a higher-level representation named semantic representation and obtained a better result. As shown in Fig. 10.

DISCUSSION

VQA is the basic field for realizing artificial intelligence, and improving the efficiency and accuracy of models is a hot spot in current research, The attention mechanism is very important for achieving effect improvement in VQA (*Zheng et al., 2022b*; *Patro, Anupriy & Namboodiri, 2022*; *Peng et al., 2022b*). Therefore, we review the development status of V QA and find that the attention mechanism has received more and more attention from scientists, and it is widely used in VQA models and has multiple forms of existence.

We selected the models with the highest citation rates in each year from 2016 to 2022 for collation and comparison, and our findings are shown in Table 2.

In Table 2, we have sorted out the attention-based model that were the mostly cited every year since 2016. The relational attention representation model pays attention to the relationship between different areas of the picture or different objects, and (graph)

Table 2 Comparison of the highest cited models in different years.

Model	Year	Free region	Object-Detection	Co-attention	Relation-attention	Self-attention
Ren et al. (2017)	2016	✓				
Ronneberger, Fischer & Brox (2015)	2016	✓		✓		
Bahdanau, Cho & Bengio (2015)	2016	✓				
Luong, Pham & Manning (2015)	2016		✓			
Anderson et al. (2018)	2017	✓		✓		
Chen et al. (2017)	2017	✓		✓		
Xu et al. (2015)	2017	✓				
Wu et al. (2017)	2017	✓		✓		
Antol et al. (2015)	2017	✓			✓	
Zhu et al. (2016)	2017	✓		✓	✓	
Zheng et al. (2020)	2018		✓			
Pan et al. (2022)	2018		✓	✓		
Li et al. (2022)	2018	✓		✓		
Gong et al. (2021)	2018	✓		✓	✓	✓
Zhan et al. (2020)	2018	✓				
Wang et al. (2022a)	2019		✓	✓	✓	✓
Wang et al. (2022b)	2019		✓	✓	✓(graph)	✓
Bazi et al. (2022)	2019		✓	✓	✓	✓
Zheng et al. (2022b)	2019	✓		✓	✓(graph)	
Al Rahhal et al. (2022)	2020	✓		✓	✓(graph)	
Gurari et al. (2018)	2020		✓	✓	✓	
Tung et al. (2021)	2020		✓		✓	
Tung, Huy Tien & Minh Le (2021)	2021		✓	✓		✓
Liu et al. (2019)	2021		✓	✓		
Gao et al. (2019)	2021	✓		✓	✓	
Selvaraju et al. (2020)	2021	✓		✓	✓	✓
Gao et al. (2015)	2022	✓	✓	✓		
Malinowski et al. (2015)	2022		✓	✓	✓(graph)	
Yang et al. (2019a)	2022	✓	✓	✓	✓	✓
Han et al. (2021)	2022		✓	✓	✓(graph)	
Yu et al. (2019)	2022		✓	✓	✓	✓

represents the model uses knowledge-based or factor-based graph attention. It can be seen that over time, researchers have proposed a variety of attention modules, and the complexity of the model has also increased.

For image attention, researchers initially focused on finding the image area related to the problem. Later, with the development of object detection technology, researchers also introduced object detection into the VQA task, so as to find the object concerned by the problem and the semantic relationship between different objects. For text attention, researchers put forward the view that problem attention and image attention are equally important. They used attention maps to rank each word in the question and select the most important word. The introduction of relational attention and self-attention also

enables VQA tasks to more accurately understand the semantic relationship inter and intra modes. However, this is still far from enough for such a complex cross-modal task. In order to further understand the high-level semantic information in the image, such as attributes and visual relationship facts, researchers introduce a factor based priori condition or knowledge graph, encode each image into a graph, and learn the attention relationship between different nodes and edges. This also greatly improves the accuracy and interpretability of VQA tasks, which may become the direction of further development of future work.

CONCLUSION AND EXPECTATION

This review provides a brief overview of the attention mechanism approach to the VQA tasks, describing its classification, shortcomings, and existing methods of improvement, and we also use bibliometric methods to comprehensively analyze the current state of the VQA field and reasonably predict the attention mechanisms' possible future directions.

Through a comprehensive study of the literature, we can reasonably infer that with the growing demand for real-time applications in AI, real-time attention may become a hot issue for future research. At the same time, we find that emotional attention and cross-modal retrieval will also become a hot issue for research in the future. Finally, the attention mechanism can be stand-alone (*Ramachandran et al., 2019*) and take the lead, and in the future, it may be possible to perform attention mechanism in a larger area or globally, and it can also open up the relationship between various modalities, enhance the adaptability to various dynamic scenarios, meet the requirements of different scenarios, process data from multiple modalities, reduce the training cost of machine learning, and form a general "intelligent machine" model with multidimensional interaction. The method based on graph attention may become a research hotspot in the future.

We found that although the attention mechanism is improving, there are still some shortcomings, and it still needs continuous exploration and research on how to break through these problems. By exploring hard in these aspects, we can make the VQA task develop toward a more intelligent, more accurate, and more humane direction.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was jointly supported by the Sichuan Science and Technology Program (2021YFQ0003). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The Sichuan Science and Technology Program: 2021YFQ0003.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Siyu Lu performed the experiments, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Mingzhe Liu analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Lirong Yin performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Zhengtong Yin performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Xuan Liu performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Wenfeng Zheng conceived and designed the experiments, performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This is a literature review.

REFERENCES

Agrawal A, Batra D, Parikh D, Kembhavi A. 2018. Don't just assume; look and answer: overcoming priors for visual question answering. In: *31st IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE, 4971–4980.

Al Rahhal MM, Bazi Y, Alsaleh SO, Al-Razgan M, Mekhalfi ML, Al Zuair M, Alajlan N. 2022. Open-ended remote sensing visual question answering with transformers. *International Journal of Remote Sensing* **43(18)**:6809–6823 DOI 10.1080/01431161.2022.2145583.

Anderson P, He XD, 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: *31st IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D. 2015. VQA: visual question answering. In: *IEEE international conference on computer vision*. Piscataway: IEEE.

Bahdanau D, Cho K, Bengio Y. 2015. Neural machine translation by jointly learning to align and translate. In: *3rd international conference on learning representations, ICLR 2015*.

Bazi Y, Al Rahhal MM, Mekhalfi ML, Al Zuair MA, Melgani F. 2022. Bi-modal transformer-based approach for visual question answering in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **60** DOI 10.1109/tgrs.2022.3192460.

Burns A, Tan R, Saenko K, Sclaroff S, Plummer BA. 2019. Language features matter: effective language representations for vision-language tasks. In: *IEEE/CVF international conference on computer vision (ICCV)*. Piscataway: IEEE.

Chaudhari S, Mithal V, Polatkan G, Ramanath R. 2021. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology* **12(5)**:32 DOI 10.1145/3465055.

Chen BY, Li PX, Sun C, Wang D, Yang G, Lu HC. 2019. Multi attention module for visual tracking. *Pattern Recognition* **87**:80–93 DOI 10.1016/j.patcog.2018.10.005.

Chen CQ, Han DZ, Wang J. 2020. Multimodal encoder-decoder attention networks for visual question answering. *IEEE Access* **8**:35662–35671 DOI 10.1109/access.2020.2975093.

Chen K, Wang J, Chen L-C, Gao H, Xu W, Nevatia R. 2015. ABC-CNN: an attention based convolutional neural network for visual question answering. ArXiv preprint. arXiv:1511.05960.

Chen L, Zhang HW, Xiao J, Nie LQ, Shao J, Liu W, Chua TS. 2017. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: *30th IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Cheng J, Dong L, Lapata M. 2016. Long short-term memory-networks for machine reading. In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. Austin: Association for Computational Linguistics.

Das A, Agrawal H, Zitnick L, Parikh D, Batra D. 2017. Human attention in visual question answering: do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* **163**:90–100 DOI 10.1016/j.cviu.2017.10.001.

Fan CY, Zhang XF, Zhang S, Wang WS, Zhang C, Huang H. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In: *32nd IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE Computer Soc.

Gan Z, Cheng Y, El Kholy A, Li LJ, Liu JJ, Gao JF. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In: *57th annual meeting of the association-for-computational-linguistics (ACL)*. Florence: Assoc Computational Linguistics-ACL.

Gao HY, Mao JH, Zhou J, Huang ZH, Wang L, Xu W. 2015. Are you talking to a machine? Dataset and methods for multilingual image question answering. In: *29th annual conference on neural information processing systems (NIPS)*. Montreal: Neural Information Processing Systems (Nips).

Gao P, Jiang ZK, You HX, Lu P, Hoi S, Wang XG, Li HS. 2019. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In: *32nd IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Gong H, Chen G, Liu S, Yu Y, Li G. 2021. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In: *11th international conference on multimedia retrieval (ICMR)*. Electrical Network.

Guo D, Wang H, Wang M. 2019. Dual visual attention network for visual dialog. In: *28th international joint conference on artificial intelligence*. Macao: International Joint Conference on Artificial Intelligence.

Guo M-H, Xu T-X, Liu J-J, Liu Z-N, Jiang P-T, Mu T-J, Zhang S-H, Martin RR, Cheng M-M, Hu S-M. 2022. Attention mechanisms in computer vision: a survey. *Computational Visual Media* **8(3)**:331–368 DOI 10.1007/s41095-022-0271-y.

Guo Z, Han D. 2022. Multi-modal co-attention relation networks for visual question answering. *Visual Computer* DOI 10.1007/s00371-022-02695-9.

Guo Z, Han D. 2023. Sparse co-attention visual question answering networks based on thresholds. *Applied Intelligence* **53**:586–600 DOI 10.1007/s10489-022-03559-4.

Gurari D, Li Q, Stangl AJ, Guo AH, Lin C, Grauman K, Luo JB, Bigham JP. 2018. VizWiz grand challenge: answering visual questions from blind people. In: *31st IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Han DZ, Zhou SL, Li KC, De Mello RF. 2021. Cross-modality co-attention networks for visual question answering. *Soft Computing* **25(7)**:5411–5421 DOI 10.1007/s00500-020-05539-7.

Hu J, Shen L, Albanie S, Sun G, Wu E. 2020. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42(8)**:2011–2023 DOI 10.1109/TPAMI.2019.2913372.

Ilievski I, Feng J. 2017. Generative attention model with adversarial self-learning for visual question answering. In: *1st International ACM multimedia thematic workshops*. Mountain View, CA.

Ilievski I, Yan S, Feng JJAE-P. 2016. A focused dynamic attention model for visual question answering. ArXiv preprint. arXiv:1604.01485.

Jiang JW, Chen ZQ, Lin HJ, Zhao XB, Gao Y. 2020. Divide and conquer: question-guided spatio-temporal contextual attention for video question answering. In: *34th AAAI conference on artificial intelligence/32nd innovative applications of artificial intelligence conference/10th AAAI symposium on educational advances in artificial intelligence*. New York: Association of the Advancement of Artificial Intelligence.

Kang GC, Lim J, Zhang BT. 2019. Dual attention networks for visual reference resolution in visual dialog. In: *Conference on empirical methods in natural language processing/9th international joint conference on natural language processing (EMNLP-IJCNLP)*. Hong Kong: Association of Computational Linguistics-ACL.

Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. 2022. Transformers in vision: a survey. *ACM Computing Surveys* **54(10s)**:1–41 DOI 10.1145/3505244.

Kim JH, Jun J, Zhang BT. 2018. Bilinear attention networks. In: *32nd conference on neural information processing systems (NIPS)*. Montreal: Neural Information Processing Systems (Nips).

Kossi K, Coulombe S, Desrosiers C, Gagnon G. 2022. No-reference video quality assessment using distortion learning and temporal attention. *IEEE Access* **10**:41010–41022 DOI 10.1109/access.2022.3167446.

Li XP, Song JK, Gao LL, Liu XL, Huang WB, He XN, Gan C. 2019. Beyond RNNs: positional self-attention with co-attention for video question answering. In: *33rd*

AAAI conference on artificial intelligence/31st innovative applications of artificial intelligence conference/9th AAAI symposium on educational advances in artificial intelligence. Honolulu: Assoc Advancement Artificial Intelligence.

Li XL, Yuan AH, Lu XQ. 2021. Vision-to-language tasks based on attributes and attention mechanism. *IEEE Transactions on Cybernetics* **51(2)**:913–926 DOI 10.1109/tcyb.2019.2914351.

Li Y, Long S, Yang Z, Weng H, Zeng K, Huang Z, Wang FL, Hao T. 2022. A Bi-level representation learning model for medical visual question answering. *Journal of Biomedical Informatics* **134**:104183 DOI 10.1016/j.jbi.2022.104183.

Liang JW, Jiang L, Cao LL, Li LJ, Hauptmann A. 2018. Focal visual-text attention for visual question answering. In: *31st IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Liang JW, Jiang L, Cao LL, Kalantidis Y, Li LJ, Hauptmann AG. 2019. Focal visual-text attention for memex question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**:1893–1908 DOI 10.1109/tpami.2018.2890628.

Liu F, Liu J, Fang Z, Hong R, Lu H. 2019. Densely connected attention flow for visual question answering. In: *28th international joint conference on artificial intelligence*. Macao.

Liu Y, Zhang XM, Huang FR, Li ZJ. 2018. Adversarial learning of answer-related representation for visual question answering. In: *27th ACM international conference on information and knowledge management (CIKM)*. Torino: Assoc Computing Machinery.

Liu Y, Zhang XM, Huang FR, Cheng L, Li ZJ. 2021a. Adversarial learning with multi-modal attention for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems* **32(9)**:3894–3908 DOI 10.1109/tnnls.2020.3016083.

Liu Y, Zhang XM, Zhang QY, Li CZ, Huang FR, Tang XH, Li ZJ. 2021b. Dual self-attention with co-attention networks for visual question answering. *Pattern Recognition* **117**:107956 DOI 10.1016/j.patcog.2021.107956.

Liu Y, Guo Y, Yin J, Song X, Liu W, Nie L, Zhang M. 2022a. Answer questions with right image regions: a visual attention regularization approach. *ACM Transactions on Multimedia Computing Communications and Applications* **18(4)**:93 DOI 10.1145/3498340.

Liu Y, Wu J, Li A, Li L, Dong W, Shi G, Lin W. 2022b. Video quality assessment with serial dependence modeling. *IEEE Transactions on Multimedia* **24**:3754–3768 DOI 10.1109/tmm.2021.3107148.

Lu JS, Yang JW, Batra D, Parikh D. 2016. Hierarchical question-image co-attention for visual question answering. In: *30th conference on neural information processing systems (NIPS)*. Barcelona: Neural Information Processing Systems (Nips), ArXiv preprint. arXiv:1606.00061.

Lu P, Li HS, Zhang W, Wang JY, Wang XG. 2018a. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In: *32nd AAAI conference on artificial intelligence/30th innovative applications of artificial intelligence conference/8th AAAI symposium on educational*

advances in artificial intelligence. New Orleans: Association of the Advancement of Artificial Intelligence.

Lu P, Ji L, Zhang W, Duan N, Zhou M, Wang JY. 2018b. R-VQA: learning visual relation facts with semantic attention for visual question answering. In: *24th ACM SIGKDD conference on knowledge discovery and data mining (KDD)*. London: Association of Computing Machinery.

Luong T, Pham H, Manning CD. 2015. Effective approaches to attention-based neural machine translation. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. Lisbon, Portugal.

Ma Z, Zheng W, Chen X, Yin L. 2021. Joint embedding VQA model based on dynamic word vector. *PeerJ Computer Science* **7**:e353 DOI 10.7717/peerj-cs.353.

Malinowski M, Rohrbach M, Fritz M. 2015. Ask your neurons: a neural-based approach to answering questions about images. In: *IEEE international conference on computer vision*. Piscataway: IEEE.

Malinowski M, Doersch C, Santoro A, Battaglia P. 2018. Learning visual question answering by bootstrapping hard attention. In: *15th European conference on computer vision (ECCV)*. Munich: Springer International Publishing Ag.

Miao Y, Cheng W, He S, Jiang H. 2022a. Research on visual question answering based on dynamic memory network model of multiple attention mechanisms. *Scientific Reports* **12(1)**:16758 DOI 10.1038/s41598-022-21149-9.

Miao Y, He S, Cheng W, Li G, Tong M. 2022b. Research on visual question answering based on GAT relational reasoning. *Neural Processing Letters* **54(2)**:1435–1448 DOI 10.1007/s11063-021-10689-2.

Nam H, Ha JW, Kim J. 2017. Dual attention networks for multimodal reasoning and matching. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Nguyen DK, Okatani T. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: *31st IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Niu ZY, Zhong GQ, Yu H. 2021. A review on the attention mechanism of deep learning. *Neurocomputing* **452**:48–62 DOI 10.1016/j.neucom.2021.03.091.

Osman A, Samek W. 2019. DRAU: dual recurrent attention units for visual question answering. *Computer Vision and Image Understanding* **185**:24–30 DOI 10.1016/j.cviu.2019.05.001.

Pan H, He S, Zhang K, Qu B, Chen C, Shi K. 2022. AMAM: an attention-based multi-modal alignment model for medical visual question answering. *Knowledge-Based Systems* **255**:109763 DOI 10.1016/j.knosys.2022.109763.

Patro BN, Anupriy , Namboodiri VP. 2022. Explanation vs. attention: a two-player game to obtain attention for VQA and visual dialog. *Pattern Recognition* **132**:108898 DOI 10.1016/j.patcog.2022.108898.

Patro B, Namboodiri VP. 2018. Differential attention for visual question answering. In: *31st IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Peng L, Yang Y, Wang Z, Wu X, Huang Z. 2019. CRA-Net: composed relation attention network for visual question answering. In: *27th ACM international conference on multimedia (MM)*. Nice: Association of Computing Machinery.

Peng L, Yang Y, Zhang X, Ji Y, Lu H, Shen HT. 2022a. Answer again: improving VQA with cascaded-answering model. *IEEE Transactions on Knowledge and Data Engineering* **34(4)**:1644–1655 DOI 10.1109/tkde.2020.2998805.

Peng L, Yang Y, Wang Z, Huang Z, Shen HT. 2022b. MRA-Net: improving VQA via multi-modal relation attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44(1)**:318–329 DOI 10.1109/tpami.2020.3004830.

Qiao TT, Dong JF, Xu DQ. 2018. Exploring human-like attention supervision in visual question answering. In: *32nd AAAI conference on artificial intelligence/30th innovative applications of artificial intelligence conference/8th AAAI symposium on educational advances in artificial intelligence*. New Orleans: Association of Advancement Artificial Intelligence.

Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J. 2019. Stand-alone self-attention in vision models. In: *33rd conference on neural information processing systems (NeurIPS)*. Vancouver: Neural Information Processing Systems (Nips).

Rawat KS, Sood S. 2021. Knowledge mapping of computer applications in education using CiteSpace *Computer Applications in Engineering Education* **29(5)**:1324–1339 DOI 10.1002/cae.22388.

Ren MY, Zemel RS. 2017. End-to-End instance segmentation with recurrent attention. In: *30th IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Ren S, He K, Girshick R, Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39(6)**:1137–1149 DOI 10.1109/TPAMI.2016.2577031.

Ronneberger O, Fischer P, Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation. In: *18th international conference on medical image computing and computer-assisted intervention (MICCAI)*. Munich, Germany.

Sarafianos N, Xu X, Kakadiaris IA. 2019. Adversarial representation learning for text-to-image matching. In: *IEEE/CVF international conference on computer vision (ICCV)*. Piscataway: IEEE.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* **128(2)**:336–359 DOI 10.1007/s11263-019-01228-7.

Seo PH, Lehrmann A, Han B, Sigal L. 2017. Visual reference resolution using attention memory for visual dialog. In: *31st annual conference on neural information processing systems (NIPS)*. Long Beach: Neural Information Processing Systems (Nips).

Sharma H, Srivastava S. 2022. Visual question answering model based on the fusion of multimodal features by a two-wav co-attention mechanism. *Imaging Science Journal* **69**:177–189 DOI 10.1080/13682199.2022.2153489.

Shen X, Han D, Chang C-C, Zong L. 2022. Dual self-guided attention with sparse question networks for visual question answering. *Ieice Transactions on Information and Systems* **E105D(4)**:785–796 DOI 10.1587/transinf.2021EDP7189.

Shi Y, Furlanello T, Zha S, Anandkumar A. 2018. Question type guided attention in visual question answering. In: *15th European conference on computer vision (ECCV)*. Munich: Springer International Publishing Ag.

Song JK, Zeng PP, Gao LL, Shen HT. 2018. From pixels to objects: cubic visual attention for visual question answering. In: *27th international joint conference on artificial intelligence (IJCAI)*. Stockholm: International Joint Conference on Artificial Intelligence.

Suresh N, Mylavarapu PM, Mahankali NS, Channappayya SS. 2022. A fast and efficient no-reference video quality assessment algorithm using video action recognition features. In: *28th National conference on communications (NCC)*. Electric Network.

Teney D, Anderson P, He X, Hengel Avd. 2018. Tips and tricks for visual question answering: learnings from the 2017 challenge. In: *2018 IEEE/CVF conference on computer vision and pattern recognition*. Piscataway: IEEE.

Tung L, Thong B, Huy Tien N, Minh Le N. 2021. Bi-direction co-attention network on visual question answering for blind people. In: *14th international conference on machine vision (ICMV)*. Rome, Italy.

Tung L, Huy Tien N, Minh Le N. 2021. Multi visual and textual embedding on visual question answering for blind people. *Neurocomputing* **465**:451–464 DOI 10.1016/j.neucom.2021.08.117.

Varga D. 2022. No-reference video quality assessment using multi-pooled, saliency weighted deep features and decision fusion. *Sensors* **22(6)**:2209 DOI 10.3390/s22062209.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin IJA. 2017. Attention is all you need. In: *31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA*.

Wang H, Pan H, Zhang K, He S, Chen C. 2022a. M2FNet: multi-granularity feature fusion network for medical visual question answering. In: *19th Pacific Rim international conference on artificial intelligence (PRICAI)*. Shanghai.

Wang M, He X, Liu L, Qing L, Chen H, Liu Y, Ren C. 2022b. Medical visual question answering based on question-type reasoning and semantic space constraint. *Artificial Intelligence in Medicine* **131**:102346 DOI 10.1016/j.artmed.2022.102346.

Wu CF, Liu JL, Wang XJ, Dong X. 2018a. Object-difference attention: a simple relational attention for visual question answering. In: *26th ACM multimedia conference (MM)*. Seoul: Association of Computing Machinery.

Wu Q, Teney D, Wang P, Shen CH, Dick A, Van den Hengel A. 2017. Visual question answering: a survey of methods and datasets. *Computer Vision and Image Understanding* **163**:21–40 DOI 10.1016/j.cviu.2017.05.001.

Wu Q, Wang P, Shen C, Reid I, Van den Hengel A. 2018b. Are you talking to me? Reasoned visual dialog generation through adversarial learning. In: *31st IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Xi YL, Zhang YN, Ding ST, Wan SH. 2020. Visual question answering model based on visual relationship detection. *Signal Processing-Image Communication* **80**:115648 DOI 10.1016/j.image.2019.115648.

Xia Q, Yu C, Hou Y, Peng P, Zheng Z, Chen W. 2022. Multi-modal alignment of visual question answering based on multi-hop attention mechanism. *Electronics* **11(11)**:1778 DOI 10.3390/electronics11111778.

Xiang Y, Zhang C, Han Z, Yu H, Li J, Zhu L. 2022. Path-wise attention memory network for visual question answering. *Mathematics* **10(18)**:3244 DOI 10.3390/math10183244.

Xu K, Ba J, Kiros R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y. 2015. Show, attend and tell: neural image caption generation with visual attention. In: *Proceedings of the 32nd international conference on international conference on machine learning*. Lille, France.

Xu XJ, Chen XY, Liu C, Rohrbach A, Darrell T, Song D. 2018. Fooling vision and language models despite localization and attention mechanism. In: *31st IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Yan F, Silamu W, Li Y, Chai Y. 2022. SPCA-Net: a based on spatial position relationship co-attention network for visual question answering. *Visual Computer* **38(9–10)**:3097–3108 DOI 10.1007/s00371-022-02524-z.

Yan F, Silamu W, Li Y. 2022. Deep modular bilinear attention network for visual question answering. *Sensors* **22(3)**:1045 DOI 10.3390/s22031045.

Yang C, Jiang MQ, Jiang B, Zhou WX, Li KQ. 2019a. Co-attention network with question type for visual question answering. *IEEE Access* **7**:40771–40781 DOI 10.1109/access.2019.2908035.

Yang TH, Zha ZJ, Zhang HW. 2019b. Making history matter: history-advantage sequence training for visual dialog. In: *IEEE/CVF international conference on computer vision (ICCV)*. Piscataway: IEEE Computer Society.

Yang ZC, He XD, Gao JF, Deng L, Smola A. 2016. Stacked attention networks for image question answering. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Yu DF, Fu JL, Mei T, Rui Y. 2017. Multi-level attention networks for visual question answering. In: *30th IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.

Yu J, Zhang WF, Lu YH, Qin ZC, Hu Y, Tan JL, Wu Q. 2020. Reasoning on the relation: enhancing visual representation for visual question answering and cross-modal retrieval. *IEEE Transactions on Multimedia* **22(12)**:3196–3209 DOI 10.1109/tmm.2020.2972830.

Yu Z, Yu J, Xiang CH, Fan JP, Tao DC. 2018. Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems* **29(12)**:5947–5959 DOI 10.1109/tnnls.2018.2817340.

Yu Z, Yu J, Cui YH, Tao DC, Tian Q. 2019. Deep modular co-attention networks for visual question answering. In: *32nd IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE Computer Society.

Zhan L-M, Liu B, Fan L, Chen J, Wu X-M. 2020. Medical visual question answering via conditional reasoning. In: *28th ACM international conference on multimedia (MM)*. Electrical Network.

Zhang S, Chen M, Chen JC, Zou FH, Li YF, Lu P. 2021. Multimodal feature-wise co-attention method for visual question answering. *Information Fusion* **73**:1–10 DOI 10.1016/j.inffus.2021.02.022.

Zhang WF, Yu J, Hu H, Hu HY, Qin ZC. 2020. Multimodal feature fusion by relational reasoning and attention for visual question answering. *Information Fusion* **55**:116–126 DOI 10.1016/j.inffus.2019.08.009.

Zhao Z, Yang QF, Cai D, He XF, Zhuang YT. 2017. Video question answering via hierarchical spatio-temporal attention networks. In: *26th international joint conference on artificial intelligence (IJCAI)*. Melbourne: International Joint Conference on Artificial Intelligence.

Zheng SJ, Li YJ, Chen S, Xu J, Yang YD. 2020. Predicting drug-protein interaction using quasi-visual question answering system. *Nature Machine Intelligence* **2(2)**:134–140 DOI 10.1038/s42256-020-0152-y.

Zheng W, Yin L, Chen X, Ma Z, Liu S, Yang B. 2021. Knowledge base graph embedding module design for Visual question answering model. *Pattern Recognition* **120**:108153 DOI 10.1016/j.patcog.2021.108153.

Zheng W, Zhou Y, Liu S, Tian J, Yang B, Yin L. 2022a. A deep fusion matching network semantic reasoning model. *Applied Sciences* **12(7)**:3416 DOI 10.3390/app12073416.

Zheng W, Liu X, Yin L. 2021. Sentence representation method based on multi-layer semantic network. *Applied Sciences* **11(3)**:1316 DOI 10.3390/app11031316.

Zheng W, Yin L. 2022. Characterization inference based on joint-optimization of multi-layer semantics and deep fusion matching network. *PeerJ Computer Science* **8**:e908 DOI 10.7717/peerj-cs.908.

Zheng XT, Wang BQ, Du XQ, Lu XQ. 2022b. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing* **60**:1–14 DOI 10.1109/tgrs.2021.3079918.

Zhou YY, Ji RR, Su JS, Sun XS, Chen WQ. 2019. Dynamic capsule attention for visual question answering. In: *33rd AAAI conference on artificial intelligence/31st innovative applications of artificial intelligence conference/9th AAAI symposium on educational advances in artificial intelligence*. Honolulu: Association of the Advancement of Artificial Intelligence.

Zhu X, Mao ZD, Chen ZN, Li YY, Wang ZH, Wang B. 2021. Object-difference drived graph convolutional networks for visual question answering. *Multimedia Tools and Applications* **80(11)**:16247–16265 DOI 10.1007/s11042-020-08790-0.

Zhu YK, Groth O, Bernstein M, Li FF. 2016. Visual7W: grounded question answering in images. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. Piscataway: IEEE.