Image and Vision Computing 116 (2021) 104327

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

# A survey of methods, datasets and evaluation metrics for visual question answering

Himanshu Sharma *, Anand Singh Jalal

*Department of Computer Engineering and Applications, GLA University, Mathura, India*

ARTICLE INFO

*Article history:*
Received 8 September 2021
Received in revised form 2 October 2021
Accepted 8 October 2021
Available online 15 October 2021

*Keywords:*
Computer vision
Natural language processing
Deep neural networks
World knowledge
Attention

ABSTRACT

Visual Question Answering (VQA) is a multi-disciplinary research problem that has captured the attention of both computer vision as well as natural language processing researchers. In Visual Question Answering, a system is given an image; a question in a natural language related to that image as an input, and the VQA system is required to give an answer in natural language as an output. A VQA algorithm may require common sense reasoning over the information contained in the image and world knowledge to produce the right answer. In this paper, we have discussed some of the core concepts used in VQA systems and present a comprehensive survey of efforts in the past to address this problem. Apart from traditional VQA models, we have also discussed visual question answering models that require reading texts present in images and evaluated on recently developed datasets like TextVQA, ST-VQA, and OCR-VQA. Apart from standard datasets discussed in previous surveys, we have also discussed some new datasets developed in 2019 and 2020 such as GQA, OK-VQA, TextVQA, ST-VQA, and OCR-VQA. The new evaluation metrics such as BLEU, MPT, METEOR, Average Normalized Levenshtein Similarity (ANLS), Validity, Plausibility, Distribution, Consistency, Grounding, F1-Score are explained together with the evaluation metrics discussed by previous surveys. We conclude our survey with a discussion on open issues in each phase of the VQA task and present some promising future directions.



## 1. Introduction

Visual Question Answering (VQA) is a multi-disciplinary artificial intelligence research problem that has gained the interest of mainly two communities: computer vision and Natural Language Processing (NLP). VQA is a task of generating natural language answers when a question in natural language is asked related to an image. Also, it may require the techniques of knowledge representation and reasoning (KRR) for natural language answer generation. Fig. 1 show VQA as a multidisciplinary research problem.

On one side, the goal of computer vision is to teach a machine how to see [1]. On the other side, the goal of NLP is to enable interactions between machine and humans in natural language [1]. Applications of deep learning research in the field of computer vision has achieved outstanding performance and given promising results in various computer vision problems which include image classification [2,3], object detection [4,5], and activity recognition [6–8].

Questions in VQA may include different sub-tasks in the field of computer vision like object recognition to find the most salient object in an image irrelevant of its position. Object detection aims to find instances of a particular semantic object in an image. Counting problems aim to find the number of object instances of a particular class in an image. The sub-problems may be more complex, for example, finding the spatial relationship between two objects or problems that require commonsense reasoning [9].

Apart from visual content, images may contain a lot of other semantic information that can be further utilized for visual question answering task. Text is one of the semantic information that commonly exists in the natural scene. Examples of such natural scenes are traffic signboard, vehicle number plate, commercial shop name, and product name slip. This gives a motivation to utilize text present in an image to refine the output of a VQA system.

In Visual Question Answering (VQA) task, the questions can be a free form open-ended, in its most general form. The system has to determine the most accurate answer, either in few words or in short phrases. Other variants of questions can be binary questions, where the VQA system will predict the answer as either yes or no [10,11]. Another variant can be multiple-choice questions [10,12] in which set of possible answer is given to the system and it has to give the most accurate answers as the output.

* Corresponding author.
*E-mail addresses:* himanshu.sharma@gla.ac.in (H. Sharma), asjalal@gla.ac.in (A.S. Jalal)

https://doi.org/10.1016/j.imavis.2021.104327
0262-8856/

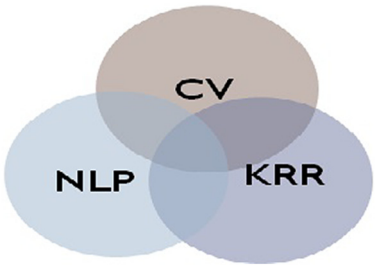



**Fig. 1.** VQA as a multidisciplinary research problem.

### 1.1. Major issues in visual question answering

***Evaluation of opened-ended and multiple-choice questions:*** Most of the state-of-the-art (SOTA) VQA systems have the capability to evaluate the multiple choice questions using conventional accuracy metric. But, these systems are not capable enough to evaluate the open-ended question. Also, there is an issue in the evaluation of multiple choice questions as the problem is reduced to just determine the correct answer instead of actually giving the answer to the question. In the case of multiple-choice questions, choices must be given to the system in such a way that questions require reasoning about the content of an image rather than just interpreting the answer from choices given.

***Need of application-oriented datasets:*** Most of the SOTA VQA models were evaluated on benchmarks datasets like DAQUAR and VQA. The models fail to handle real-world applications like helping blind people, supporting a data analyst to provide a useful content from huge amount of data, guiding children learning a concept on smart devices and communicating with a robot. So, there is a need of more publically available goal-oriented datasets like VizWiz and VQA-Med.

***Dataset Bias:*** Existing VQA datasets are strongly biased. Most of SOTA VQA models are more dependent on the question than the image content. How a question is designed strongly decides the answer. This in turn will significantly weaken the capability to assess VQA algorithms. Also, to answer questions that need the use of image content is comparatively easy as the majority of these questions are related to existence of objects or scene attributes. CNNs can handle these questions effectively. Also, they have strong language biases. The questions starting with'Why' are relatively harder to answer and rare. This will have serious implication on performance evaluation.

***Image Featurization from real VQA datasets:*** The real VQA datasets like VizWiz contains low-quality images as they are captured by blind people. To answer compositional questions based on these low-quality images, single network would not be sufficient to extract image features.

***Relatively Small size of datasets:*** The VQA datasets like VizWiz and VQA-Med that contains real-world images are relatively smaller in size as collection of data is generally more costly and need huge amount of time. On the other side, large amount of data is required to train the deep VQA models to gain reasoning capabilities and capture underlying concepts.

***Balanced Binary Questions:*** The use of binary questions to evaluate VQA systems is widely used as they are easy to answer by human annotators and lack of complex questions. But, to create balanced binary questions is a challenge as these questions are biased towards 'yes' in majority of datasets like COCO-VQA.

***Imbalance between answerable and unanswerable classes:*** Majority of publically available VQA datasets are artificially created. In such datasets, each visual question has at-least one right answer and is answerable. But, in real world application, question may not have always a right answer. Apart from VizWiz dataset, the other VQA datasets do not present unanswerable classes or samples.

***Conversational questions:*** In real world applications like helping visually-impaired users, the questions can be from verbal interactions. Such questions can have large average length and may contain many uninformative terms. To remove these terms in pre-processing step and learn techniques to better represent the question is a challenge.

***Questions requiring reading scene-text:*** Many questions in real world may require the capability to read the text present in images. The biggest challenge is to separate the text from the background is very complicated in a complex natural image. Also, diverse exotic font styles cause difficulty to achieve text localization and recognition. Due to irregular illumination, the camera sensor reaction is also irregular, which extracts distorted and deteriorated visual features and subsequently generate incorrect scene text detection and recognition.

### 1.2. Applications of visual question answering

VQA systems have many potential applications in real life. First, the most obvious is to aid blind and visually-impaired persons with identification task. The visual content is everywhere and language is how we communicate. Vision is our primary sensor. Visually impaired users don't have access to this visual content, but they have access to language interface. If we have a way to connect vision to language, then they can access this visual content. Second, VQA systems can be used to interact with, organize or go through large quantities of typically unstructured visual data. Suppose a person has given a collection of images or videos which he had captured a long time back, the aim is to help users to recover their memory about the incident captured in collection. Then if there is a bridge between vision and language, then it can be a very natural way to do this interaction. Third, it can also be used as a teaching assistant who can help the young kid to know about specific object and activities. Forth, a lot of content on the web is inherently multimodal, e.g. text, images or videos and if you a way of connecting vision to language, we can take better advantage of this information on web to learn more about the world. Fifth, it can be used to summarize a large amount of visual data for an analyst. These are some of the potential applications of the desired VQA system. In Fig. 2, we have shown th major applications of VQA in real-life.

### 1.3. Overview of related surveys

There are few survey papers on visual question answering as mentioned in Table 1.Wu et al. [1] classified the VQA methods based on their mechanism to combine both visual and textual features. They

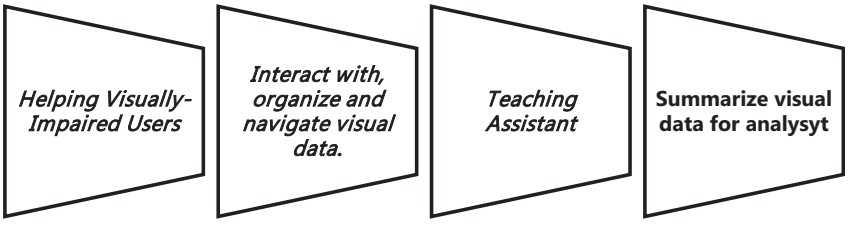



**Fig. 2.** Applications of visual question answering.

**Table 1**
Related literature survey.

| Author | Paper |
|---|---|
| Wu et al. (2017) | Visual Question Answering: A Survey of Methods and Datasets |
| Kafle and Kanan (2017) | Visual question answering: Datasets, algorithms, and future challenges |
| Zhang et al. (2019) | Information fusion in visual question answering: A Survey |
| Manmadhan and Binsu (2020) | Visual question answering: a state-of-the-art review |
| Charulata and Manasi (2020) | Visual Question Generation: The State of the Art |

also discussed memory-augmented and modular architectures that exploit structured knowledge bases. Further, they discussed datasets available for VQA task together with evaluation metrics. Kafle and Kanan [9] proposed a review which mainly covers datasets, evaluation metrics and algorithms for VQA task. They also discussed the challenges in evaluating the multi-word answers faced by VQA models. In addition, they gave an insight about how biases and other problems affect the performance of VQA systems. Zhang et al. [10] have reviewed and summarized the different fusion strategies such as simple vector operators, deep neural networks, bilinear pooling, attention mechanisms and memory networks adopted for the task of visual question answering. Manmadhan and Binsu [14] discussed the steps involved in VQA task in details such as image feature extraction, question feature extraction and fusion of both the modalities. In addition, they discussed the dataset used in VQA task with their limitations. Finally, they presented the summary of the evaluation metrics for VQA. Charulata and Manasi [15] proposed a survey which mainly focused on visual question generation task. They classified the visual question as totally grounded questions, common-sense questions and questions that need world-knowledge. In addition, they discussed methods to understand the task of VQG, available datasets and evaluation metrics for VQG task.

**Contributions of this survey**: The abovementioned surveys contribute a comprehensive analysis of the traditional study on visual question answering. Majority of these surveys mainly focused on the overview of datasets dimensions and overall models. In this paper, we have discussed all the steps involved in VQA task such as image encoding, question representation, different attention mechanisms and various fusion strategies adopted till date in detail. Fig. 3 shows all the steps involved in VQA task. The major contributions of this survey are:

- We have discussed in detail the image and question feature extraction by SOTA models including the recent (2020) SOTA models.

**Table 2**
Summary of state-of-the-art CNN models.

| Model | Year | # layers | Dimension of Input | Dimension of output (# features) | Reported error |
|---|---|---|---|---|---|
| AlexNet | 2012 | 8 | 227 × 227 | 4096 | 16.4 |
| ZFNet | 2013 | 8 | 227 × 227 | 4096 | 11.7 |
| VGGNet | 2014 | 19 | 224 × 224 | 4096 | 7.3 |
| GoogleNet | 2014 | 22 | 229 × 229 | 1024 | 6.7 |
| ResNet | 2015 | 152 | 224 × 224 | 20,148 | 3.57 |

- Apart from standard datasets discussed in previous surveys, we have also discussed some new datasets developed in 2019 and 2020 such as GQA, OK-VQA, TextVQA, ST-VQA, and OCR-VQA. Till today, these datasets are not discussed in any survey.
- The new evaluation metrics such as BLEU, MPT, METEOR, Average Normalized Levenshtein Similarity (ANLS), Validity, Plausibility, Distribution, Consistency, Grounding, F1-Score are explained together with the evaluation metrics discussed by previous surveys.
- We have also discussed the various attention mechanisms, i.e. both single-hop and multiple-hops, between the visual and question features. The different fusion strategies adopted by SOTA VQA models are discussed in depth.
- We have also explained and different baseline methods are compared based on different benchmark datasets together with these recently developed datasets. Another important contribution is, we have discussed visual question answering models that require reading texts present in images and evaluated on recently developed datasets like TextVQA, ST-VQA, and OCR-VQA. To the best of our knowledge, this is the first survey that covers the traditional visual question answering to scene-text visual question answering.
- We have presented detailed result analysis on almost all VQA datasets such as VQA 1.0, VQA 2.0, COCO-QA, DAQUAR, Vsiual7W, CLEVR, FVQA, Visual7W + KB, GQA, OK-VQA, TextVQA, ST-VQA, and OCR-VQA.
- Finally, we have analyzed a few open challenges and enlist prospective future guidelines.

## 2. Feature extraction

### 2.1. Image Featurization

Image feature extraction is one of the key tasks of VQA. To perform the different sets of mathematical operations, an image is represented as a vector using image featurization. Various convolutional neural

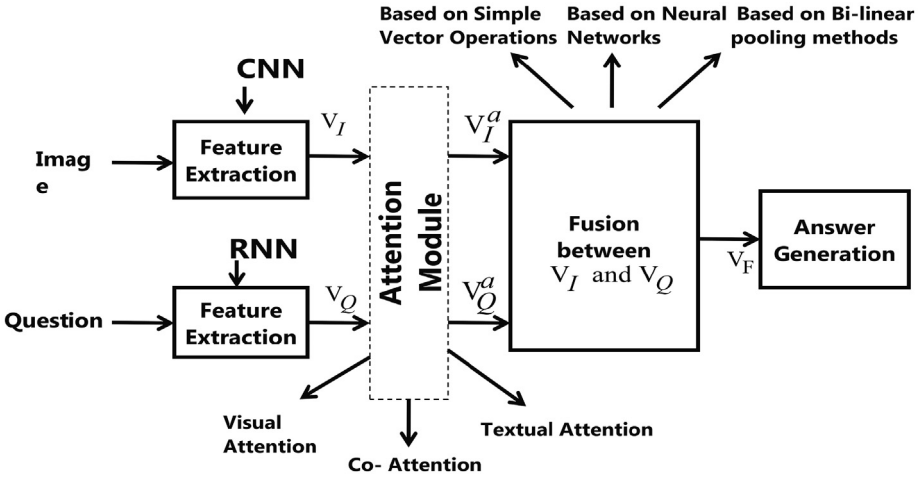



**Fig. 3.** End-end framework of VQA models.

**Table 3**
CNN Models used for extraction of image features in traditional SOTA models (2014–2017).

| VQA Model | Year | VGGNet | GoogleNet | ResNet | Faster R-CNN |
|---|---|---|---|---|---|
| VQA-AA [150] | 2014 | | | √ | |
| LSTM Q + I [10] | 2015 | √ | | | |
| ABC-CNN [30] | 2015 | √ | | | |
| iBOWING [40] | 2015 | | √ | | |
| Neural-Image QA [42] | 2015 | | √ | | |
| VIS + LSTM [50] | 2015 | √ | | | |
| VIS + BILSTM [50] | 2015 | √ | | | |
| Image_QA [51] | 2015 | √ | | | |
| mQA [55] | 2015 | | √ | | |
| SMem-VQA [61] | 2015 | | √ | | |
| Explicit-Knowledge-Base [82] | 2015 | | | √ | √ |
| NMN + LSTM [111] | 2015 | √ | | | |
| Vis_Madlibs [112] | 2015 | √ | | | |
| LSTM + Attention [12] | 2016 | √ | | | |
| Word + Region Sel [29] | 2016 | √ | | | |
| SAN [34] | 2016 | √ | √ | | |
| QRU [35] | 2016 | √ | | | |
| MCB [45] | 2016 | √ | | √ | |
| MLB [47] | 2016 | | | √ | |
| DPPnet [54] | 2016 | √ | | | |
| MRN [56] | 2016 | | | √ | |
| FDA [62] | 2016 | | | √ | |
| HieCoAtt [63] | 2016 | √ | | √ | |
| DMN [76] | 2016 | √ | | | |
| Attributes-CNN + LSTM [152] | 2016 | √ | | | |
| Re_Baseline [155] | 2016 | | | √ | |
| Answer_CNN [53] | 2016 | √ | | | |
| VQA-Caption [202] | 2016 | √ | | | |
| DAN [204] | 2016 | | √ | | |
| MFB [25] | 2017 | | | √ | |
| MLAN [27] | 2017 | | | √ | |
| High-Order [31] | 2017 | √ | | √ | |
| FVQA [83] | 2017 | √ | | | |
| MUTAN [148] | 2017 | | | √ | |
| MF + SIG + VG [151] | 2017 | | | √ | |
| SAAA [163] | 2017 | | | √ | |
| SCMC [203] | 2017 | | | √ | |

**Table 4**
CNN Models used for extraction of image features in recent SOTA models (2018–2020.

| VQA Model | Year | VGGNet | GoogleNet | ResNet | Faster R-CNN |
|---|---|---|---|---|---|
| Up-down [26] | 2018 | | | | √ |
| CVA [28] | 2018 | | | | √ |
| MAN-VGG [33] | 2018 | √ | | √ | |
| MAN-ResNet [33] | 2018 | | | √ | |
| EnsAtt [44] | 2018 | | | √ | |
| MFH [49] | 2018 | | | √ | |
| CMF [57] | 2018 | | | √ | |
| QGHC [58] | 2018 | √ | | √ | |
| DCN [65] | 2018 | | | √ | |
| Tips-Tricks [66] | 2018 | | | | √ |
| FVTA [67] | 2018 | | | √ | |
| BAN [71] | 2018 | | | | √ |
| VKMN [80] | 2018 | | | √ | |
| AVQAN [136] | 2018 | | | √ | |
| Dual-MFA [147] | 2018 | | | √ | √ |
| NMN + Caption Information [149] | 2018 | √ | | | |
| Attention-on-Attention [156] | 2018 | | | | √ |
| MLB + DA-NTN [161] | 2018 | | | √ | √ |
| CATL-QTA-M [197] | 2018 | | | √ | √ |
| MetaVQA [199] | 2018 | | | √ | |
| StrSem [200] | 2018 | | √ | | |
| Hard-Att [201] | 2018 | | | √ | |
| DRAU [32] | 2019 | | | √ | √ |
| WRAN [160] | 2019 | | | √ | |
| QAR [198] | 2019 | | | √ | |
| MAVQA [205] | 2019 | √ | | | |
| GRUC [87] | 2020 | | | | √ |
| VRR + Att [88] | 2020 | | | √ | √ |
| LRN [89] | 2020 | | | | √ |
| ODA-GCN [90] | 2020 | | | | √ |
| MOVRD [91] | 2020 | | | | √ |
| EMQA [92] | 2020 | √ | | | |
| DecomVQANet [93] | 2020 | | | | |
| MDFNet [94] | 2020 | | | √ | √ |
| IASSM[[95] | 2020 | | | | |
| ALSA [96] | 2020 | | | √ | √ |
| QLOB [97] | 2020 | | | √ | √ |
| SelRes + SelMask + bbox [98] | 2020 | | | | √ |
| QC-MLB [99] | 2020 | | | √ | |
| SANMT [100] | 2020 | | | √ | |
| RSVQA [101] | 2020 | | | √ | |

networks (CNNs) are evolved for extracting image features. Table 2 shows the various CNN models employed for the task if image feature extraction. LeNet [16] model has two convolutional layers with a simple network structure.

It is mainly used for recognizing handwritten digits. AlexNet [17] has 5 convolutional layer network structures. It was the first deep network, which was used to the increase the accuracy of classification task. The activation function used by AlexNet is ReLu (Rectified Linear Unit) which is faster than sigmoid or tanh in terms of training. It was the winner of ILSVRC2012. In ILSVRC2014, the best models develop for image classification task were GoogLeNet [18] and VGG-Net [2]. These models become the most effective models for image feature extraction. In 2016, ResNet (Residual network) [3] was proposed and became a milestone in the field of CNNs as it efficiently handles the training issue that arises in deep neural networks. It solves the problem of vanishing gradient by skipping one or more layers by identifying shortcut connections. In Table 3, the summary of various CNN models used by SOTA VQA models (2014–2017) is presented. In Table 4, the summary of recent state-of-the-art (SOTA) models (2018–20) is presented and shows the CNN models used by them.

### 2.2. Question featurization

VQA models use word embeddings to generate a feature vector for a given natural language question. By using word embeddings, we aim to map the words to numerical vectors. These vectors are easy to handle by computers. Word embeddings are mainly employed for generating language models and extracting textual features in natural language processing (NLP). The main purpose of using word embedding algorithms is to extract morphological, semantic and contextual information. In Table 5, we have presented the summary of word embedding methods used by SOTA models (2014–17) for extracting question features. Table 6 shows the mapping of recent SOTA models (2018–20) to the word embedding techniques. It can be observed that LSTM [19], GRU [20] and Bi-LSTM which belongs to the family of recurrent neural networks (RNNs), were most effective when dealing with the sequential data.

The VQA task requires word embeddings for representing questions as machine learning approaches and deep learning models are not competent enough to deal with a sequence of words or strings in their raw form. The word embedding methods can be broadly classified into three main categories: Counting based approaches, prediction based approaches and fusion bases approaches. These approaches use a large corpus of text as an input from web or research articles. Vocabulary (V) of corpus is made by the set of all distinct words. The output of word embedding is the encoded form of every word in V.

#### 2.2.1. Counting based approaches

One-hot encoding of words is the simplest word embedding approach. Each word is represented by a vector of size|V|. Fig. 4 shows an example of one-hot encoding.

Another counting based approach is co-occurrence matrix which has a size of $|V| \times |V|$. The values contained in the matrix represent the occurrence a word in the context of another word. We can define a context

**Table 5**
Mapping of traditional SOTA models (2014–2017) to word embedding methods.

| VQA Model | Year | One-hot embedding | CBOW | Skip-gram/word2vec | GloVe | CNN | LSTM | GRU | Bi-LSTM | Skip-thoughts |
|---|---|---|---|---|---|---|---|---|---|---|
| VQA-AA [150] | 2014 | | | | | | √ | | | |
| LSTM Q + I [10] | 2015 | | √ | | | | √ | | | |
| ABC-CNN [30] | 2015 | | | | | | √ | | √ | |
| iBOWING [40] | 2015 | √ | | | | | | | | |
| Neural-Image QA [42] | 2015 | | √ | | | √ | √ | √ | | |
| VIS + LSTM [50] | 2015 | | | | | | √ | | | |
| VIS + BILSTM [50] | 2015 | | | | | | | | √ | |
| Image_QA [51] | 2015 | | | | √ | | | | | |
| mQA [55] | 2015 | | | | | | √ | | | |
| SMem-VQA [61] | 2015 | | √ | | | | √ | | | |
| Explicit-Knowledge-Base [82] | 2015 | | | | | | √ | | | |
| NMN + LSTM [111] | 2015 | | | | | | √ | | | |
| Vis_Madlibs [112] | 2015 | | | √ | | | | | | |
| LSTM + Attention [12] | 2016 | √ | | | | | √ | | | |
| Word + Region Sel [29] | 2016 | | | √ | | | √ | | | |
| SAN [34] | 2016 | | | | | √ | √ | | | |
| QRU [35] | 2016 | | | | | | √ | √ | | |
| MCB [45] | 2016 | | | | | | √ | √ | | |
| MLB [47] | 2016 | | | | | | | √ | | |
| DPPnet [54] | 2016 | | | | | | √ | √ | | |
| MRN [56] | 2016 | | | | | | √ | √ | | |
| FDA [62] | 2016 | | | | | | √ | | | |
| HieCoAtt [63] | 2016 | | | | | | √ | | | |
| DMN [76] | 2016 | | | | | | | √ | | |
| Attributes-CNN + LSTM [152] | 2016 | | | | | | √ | | | |
| Re_Baseline [155] | 2016 | | | √ | | | | | | |
| Answer_CNN [53] | 2016 | | | | | √ | | | | |
| VQA-Caption [202] | 2016 | | | | | | √ | | | |
| DAN [204] | 2016 | | | | | √ | | | | |
| MFB [25] | 2017 | | | | | | √ | | | |
| MLAN [27] | 2017 | | | | | | | √ | | |
| High-Order [31] | 2017 | | | | | | √ | | | |
| FVQA [83] | 2017 | | | | | | √ | | | |
| MUTAN [148] | 2017 | | | | | | | √ | | |
| MF + SIG + VG [151] | 2017 | | | | | | | √ | | |
| SAAA [163] | 2017 | | | | | | √ | | | |
| SCMC [203] | 2017 | | | | | √ | | | | |

as k-size window around a particular word. For example, consider the same vocabulary used in Fig. 5. Then, we can represent the co-occurrence matrix as shown in Fig. 5.

*2.2.2. Prediction based approaches*

Prediction based methods learn word representations directly. These methods use neural networks module. Continuous bag-of-words (CBOW) and skip-gram models are two prediction-based word-embedding techniques proposed by Mikolov et al. [21]. Fig. 6 shows the abstract framework of CBOW (Left) and Skip-gram (Right) architectures.

In CBOW, the model uses feed forward neural network to predict a word, when (n - 1) context words are given and treat it as a multi-class classification problem. Further, this model generates an output word with respect to a bag of context words. In Skip-gram models (Fig. 6), the model predicts the context words on both sides for a given input word. In his other paper, Mikolov et al. [22] suggested different modifications to fundamental skip-gram model to handle the issue of costly operations performed at output layer. Negative sampling is popularly implemented modification, which is used in word2vec. Word2vec is an open source venture made by Google for skip-grams.

*2.2.3. Hybrid models*

Global vector (Glove) was proposed by Pennington et al. [23]. In GloVe, both count based approaches and prediction based approaches are fused to generate a word representation. Also, [23] used a weighted least squares method. Further, global information obtained from the co-occurrence matrix is used to train this model. Non-zero entries are used for the purpose of training instead of the whole sparse matrix.

*2.2.4. Recent text embedding models*

Convolutional Neural Networks (CNN) [17], long short term memory (LSTM) [19] and gated recurrent unit (GRU) [20] are also used to represent questions. In CNN based question feature extraction, concatenated encoding vectors of the entire words of a question is given as an input to the model. Then, multiple convolutional filters are applied and finally max-pooling operations are performed. Then, resultant feature maps are flattened to employ second-last layer as question vector representation.

It can be observed that LSTM which belongs to RNN family are majorly used by the researchers to represent the questions. Young et al. [24] claimed that RNN which are sequence based models perform superior to word sequence independent models such as word2vec. These methods rely on traditional word embeddings as the vectors produced by these traditional models are given as an input to LSTM or GRU. However, these models require huge quantity of labeled data for training. In Fig. 7, we have shown the an LSTM network with the flow of information in it.

## 3. Attention mechanism

In order to improve the interaction between visual and question features, attention mechanism are employed and proved to be effective. The main objective of applying attention mechanism is to focus on important words of the question and find important regions in the input image, which is crucial to generate the answer. The VQA models that do not use attention mechanism may contain noise and are not able to accurately answer fine-grained questions [25]. Depending upon the number of attention layers, attention approaches can be divided into

**Table 6**
Mapping of recent SOTA models (2018–2020) to word embedding methods.

| VQA Model | Year | One-hot embedding | CBOW | Skip-gram/word2vec | GloVe | CNN | LSTM | GRU | Bi-LSTM | Skip-thoughts |
|---|---|---|---|---|---|---|---|---|---|---|
| Up-down [26] | 2018 | | | | | | | √ | | |
| CVA [28] | 2018 | | | | | | | √ | | |
| MAN-VGG [33] | 2018 | | | | | | √ | | | |
| MAN-ResNet [33] | 2018 | | | | | | √ | | | |
| EnsAtt [44] | 2018 | | | | √ | | | | | |
| MFH [49] | 2018 | | | | | | √ | | | |
| CMF [57] | 2018 | | | | √ | | √ | | | |
| QGHC [58] | 2018 | | | | | | √ | √ | | |
| DCN [65] | 2018 | | | | | | | | √ | |
| Tips-Tricks [66] | 2018 | | | | | | | √ | | |
| FVTA [67] | 2018 | | | | | | | | √ | |
| BAN [71] | 2018 | | | | | | | √ | | |
| VKMN [80] | 2018 | | | | | | √ | | | |
| AVQAN [136] | 2018 | √ | | | | | | | | |
| Dual-MFA [147] | 2018 | | | | | | | √ | | |
| NMN + Caption Information [149] | 2018 | | | | | | √ | | | |
| Attention-on-Attention [156] | 2018 | | | | | | | √ | | |
| MLB + DA-NTN [161] | 2018 | | | | | | | √ | | |
| CATL-QTA-M [197] | 2018 | | | | | | √ | | | |
| MetaVQA [199] | 2018 | | | | √ | | | √ | | |
| StrSem [200] | 2018 | | | | | | √ | | | |
| Hard-Att [201] | 2018 | | | | | | √ | | | |
| DRAU [32] | 2019 | | | | | | | | √ | |
| WRAN [160] | 2019 | | | | | | | √ | | |
| QAR [198] | 2019 | | | | √ | | | | | |
| MAVQA [205] | 2019 | | | | | √ | | | | |
| GRUC [87] | 2020 | | | | √ | | √ | | | |
| VRR + Att [88] | 2020 | | | | √ | | | √ | | |
| LRN [89] | 2020 | | | | √ | | | √ | | |
| ODA-GCN [90] | 2020 | | | | √ | | | √ | | |
| MOVRD [91] | 2020 | | | √ | | | √ | | | |
| EMQA [92] | 2020 | | | | | | √ | | | |
| DecomVQANet [93] | 2020 | | | | | | √ | | | |
| MDFNet [94] | 2020 | | | | | | | √ | | |
| IASSM[[95] | 2020 | | | | √ | | √ | | | |
| ALSA [96] | 2020 | | | | √ | | | √ | | |
| QLOB [97] | 2020 | √ | | | | | √ | √ | | |
| SelRes + SelMask + bbox [98] | 2020 | | | | √ | | √ | | | |
| QC-MLB [99] | 2020 | | | | | | | | | √ |
| SANMT [100] | 2020 | | | | | | | | √ | |
| RSVQA [101] | 2020 | | | | | | | | | √ |

Corpus: Visual Question Answering involves both computer vision and natural language processing. $|V|=11$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **visual** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **question** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **answering** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **involves** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **both** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **computer** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **vision** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **and** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **natural** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **language** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **processing** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

**Fig. 4.** One-hot Encoding.

Corpus: Visual Question Answering involves both computer vision and natural language processing. | V | =11

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **visual** | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **question** | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **answering** | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **involves** | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **both** | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| **computer** | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| **vision** | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| **and** | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| **natural** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| **language** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| **processing** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

**Fig. 5.** Co-occurrence Matrix.

single-hop attention and multi-hop attention. In multi-hop attention, single-hop attention is applied repeatedly till specified number of times. In the subsequent subsection, these attention mechanisms are discussed in detail.

### 3.1. Single-hop attention

The main idea of attention mechanism employed in VQA models is the computation of weight vectors corresponding to one of the feature types (visual or textual) guided by the information gained from other feature types. Based upon the feature type, single-hop attention can be divided into visual attention, question attention and co-attention. In Fig. 8, each type of attention is shown in an abstract form. These attention modules are explained in detail in subsequent subsections.

VQA models that use visual attention use $V_{ques}$ as the guidance to compute the attention weight vector $W_a$.This attention weight vector $W_a$ is further used to calculate attended visual feature $V^a_{img}$ as follows $V^a_{img} = W_a V_{img}$. To obtain the attended visual features, the most common approach is to find the correlation between visual features and question features. This is further normalized by using softmax function.

$$z = W_3^T \tanh\left(W_1 V_{img} + W_2 V_{ques}\right) \tag{1}$$

$$W_a = soft\max(z) \tag{2}$$

This approach is used by [12,26–28]. Attention weight vectors can also be obtained in other ways. In reference [29], vector z is obtained as the product of $V_{img}$ & $V_{ques}$. $W_a$ is calculated by applying softmax function on z. In reference [30], attention weight vector z is computed as $z = sig\ \mathrm{mod}\ (W_{V_{ques}}) * V_{img}$, where $*$ corresponds to a convolutional operator.

In Co attention mechanism, we compute attention weight vectors for both visual and question feature types. These two attention weight vectors are derived in the same manner and performed in parallel. In [31], similarity matrix entries represent the correlation between the visual and question features. In [32], the Dual Recurrent Attention Unit (DRAU) was proposed which is also symmetric. The attention weight is computed as:

$$z = P\,\mathrm{Re}\,LU(\mathrm{W}_2 LSTM(\mathrm{P}\,\mathrm{Re}\,LU(\mathrm{W}_1 \mathrm{V}) \tag{3}$$

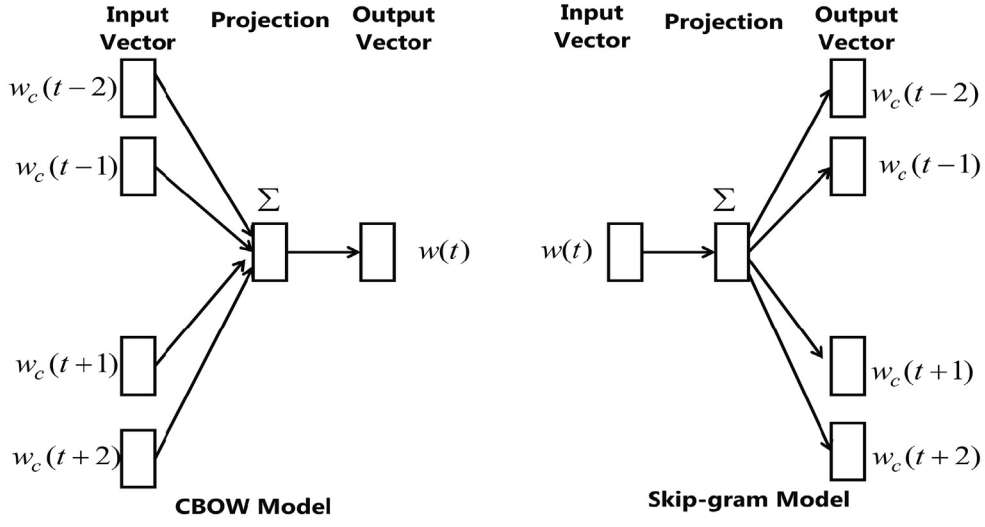



**Fig. 6.** Abstract framework of CBOW and Skip-gram Models.

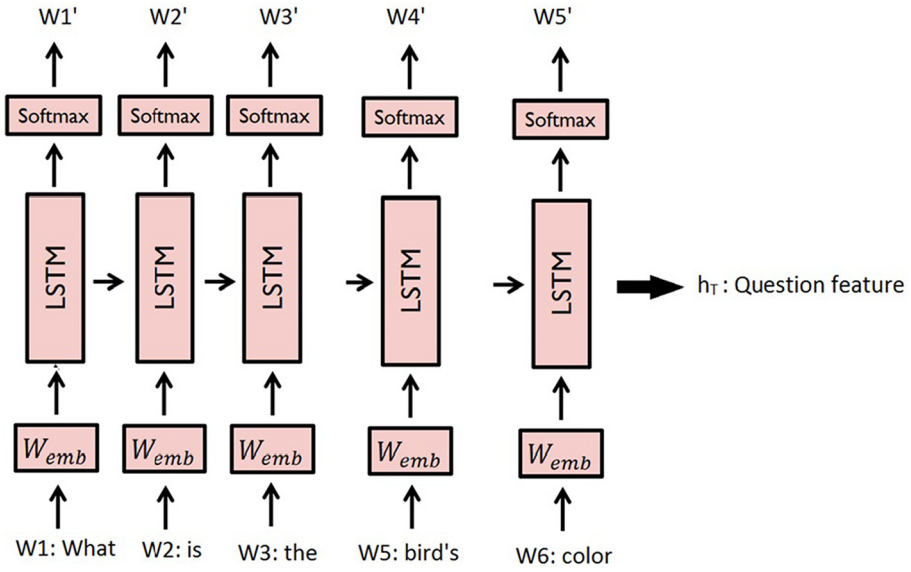



**Fig. 7.** LSTM framework for question feature extraction.

$$W_a = soft\max(z) \tag{4}$$

It uses activation function is used and is either is visual or question feature. In [33], vector is calculated as:

$$z = \tanh(W_1 V) + \tanh\left(W_2 V_{img} \odot V_{ques}\right) \tag{5}$$

To calculate the visual or question attention weight, we simply need to replace $V$ with $V_{img}$ or $V_{ques}$.

### 3.2. Multi-hop attention

Yang et al. [34] used multi-hop attention mechanism. They proposed that VQA models may require multi-step reasoning. Suppose the input question is "what are sitting in the basket on a bicycle", the VQA model must be able to focus on the objects e.g. basket and bicycle and identify the concepts e.g. "sitting in" in the given question. Based on these concepts, the VQA model must be able to discard the irrelevant objects and focus on the image regions that are required to generate the answer. The model proposed by them was named as SAN (Stacked Attention Network) that applies the attention from one layer to second layer.

The other type of multi-layer attention network is QRU (Question Representation Update) [35]. The attention vector in QRU is computed as a function of $V_{ques}$ while in SAN, the attention vector is computed as $V_{img}$. The advantage of using multi-layer attention in textual modality is to make the question more specific. QRU model uses the similar approach as in [36,37] and employs MLP to produce the association between the question and image features.

DAN (Dual Attention Network) [38] uses co-attention mechanism into multiple hops. Two attention vectors are derived in a symmetric way, one for visual features and the other for question features. The outputs are further fused using two vector-based operators:

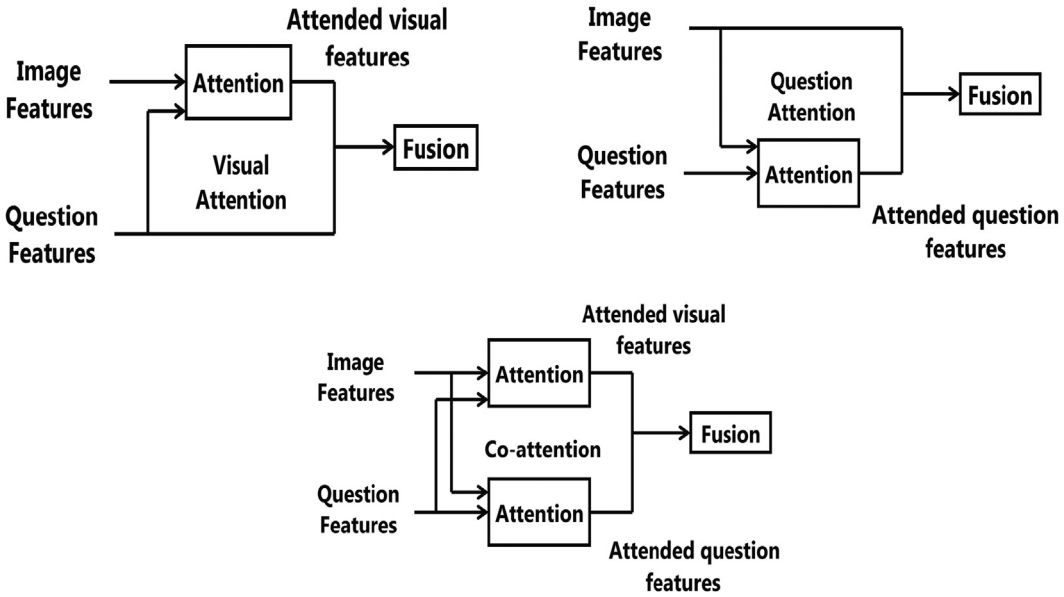



**Fig. 8.** Various forms of single-layer attention mechanisms.

**Table 7**
Single-hop and Multi-layer attention for SOTA VQA models.

| Model | Year | Single-hop attention | | Multi-layer attention | | |
|---|---|---|---|---|---|---|
| | | Attention in Visual Channel | Co-Attention in Both Channels | Attention in Visual Channel | Attention in Textual Channel | Co-Attention in Both Channels |
| VQA-AA [150] | 2014 | √ | | | | |
| ABC-CNN [30] | 2015 | √ | | | | |
| SMem-VQA [61] | 2015 | | | √ | | |
| LSTM + Attention [12] | 2016 | √ | | | | |
| Word + Region Sel [29] | 2016 | √ | | | | |
| SAN [34] | 2016 | √ | | √ | | |
| QRU [35] | 2016 | | | | √ | |
| MRN [56] | 2016 | √ | | √ | | |
| FDA [62] | 2016 | √ | | | | |
| DAN [204] | 2016 | | | | | √ |
| MFB [25] | 2017 | | √ | | | |
| MLAN [27] | 2017 | √ | | | | |
| High-Order [31] | 2017 | | √ | | | |
| MF + SIG + VG [151] | 2017 | | | √ | | |
| SAAA [163] | 2017 | √ | | | | |
| Up-down [26] | 2018 | √ | | | | |
| CVA [28] | 2018 | √ | | | | |
| MAN [33] | 2018 | | √ | | | |
| MFH [49] | 2018 | | √ | | | |
| CMF [57] | 2018 | | √ | | | |
| QGHC [58] | 2018 | √ | | | | |
| DCN [65] | 2018 | | | | | √ |
| FVTA [67] | 2018 | | √ | | | |
| VKMN [80] | 2018 | √ | | | | |
| Dual-MFA [147] | 2018 | √ | | | | |
| Attention-on-Attention [156] | 2018 | √ | | | | |
| DRAU [32] | 2019 | | √ | | | |
| MTA [206] | 2019 | √ | | | | |

$$V_F^{(k)} = V_F^{(k-1)} + V_{img}^{(k)} \odot V_{ques}^{(k)} \tag{6}$$

where, $V_{img}^{k}$ and $V_{ques}^{k}$ are the attended visual and textual features in the *k-th* step.

In Table 7, we have shown the summary presented the summary of attention mechanism used by SOTA VQA models both single-hop and multi-hop.

## 4. VQA methods based on fusion of image and question features

In the last three years, a large number of VQA algorithms are proposed. Almost all existing method consists of first extracting image features then extracting question features and combining these two modalities to generate an answer to the given question. For extracting features from an image, most methods use Convolutional Neural Networks (CNNs) that are pre-trained on a large database of images, e.g. ImageNet. The most commonly used CNN models are VGGNet [2], ResNet [3] and GOOgLeNet [18]. Extraction of question features is done through Long Short Term Memory (LSTM) [19], Bag-of-Words (BOW), Gated Recurrent Unit (GRU) [20] and skip thought vectors [39]. In order to generate the answer, VQA task is considered as classification problem. Both feature vectors are given to the classification algorithm and each predicted answer is considered as of different class. Table 8 and Table 9 show the summary of the different fusion techniques adopted by SOTA VQA models.

### 4.1. Baseline fusion methods

In this approach, the two types of features, i.e. visual feature and text feature are mapped to a common feature space. Visual features are extracted using pre-trained convolutional networks and text features are extracted using word embeddings that are trained on a large text corpus. These word embeddings are mapped to semantic space where the word those have similar meaning are grouped together. Both image and question feature vectors are then fused using simple vector based operations such as concatenation, element-wise sum and element-wise multiplication.

In vector concatenation, visual feature ($V_{img}$) and question/textual feature ($V_{ques}$) are merged together to generate an $(m + n)$-dimensional fused vector as follows:

$$V_F = [V_{img}, V_{ques}] \tag{7}$$

In element-wise addition and multiplication, $V_{img}$ and $V_{ques}$ must have the same dimensions. Otherwise, linear projection is applied to make these vectors of same dimension and then only we can apply these two operations as follows:

$$V_{img} = W_{img} V_{img} \tag{8}$$

$$V_{ques} = W_{ques} V_{ques} \tag{9}$$

Where, $W_{img}$ and $W_{ques}$ are the two embedding vectors and mapped to the same space. Thus, element-wise addition vector operations can be written as:

$$V_{img} = V_{img} + V_{ques} \tag{10}$$

and element-wise multiplication vector operations can be written as:

$$V_{img} = V_{img} \odot V_{ques} \tag{11}$$

Antol et al. [10] used LSTM with two layers to extract questions features and the VGGNet to extract features of the images. Element-wise multiplication is then performed to fuse both feature vectors. The fused vector obtained is them passed through a fully connected layer and then the softmax activation function is used to predict the answer. Zhang et al. [11] did a balancing of VQA dataset of abstract binary scenes by adding complementary scenes so that all binary questions have one answer as "yes" for one scene and have an answer as "no" for another

**Table 8**
Fusion Strategies used in SOTA VQA models (2014–2017).

| Model | Year | Simple Vector Operation | | | Neural Networks | | Bilinear Models |
|---|---|---|---|---|---|---|---|
| | | Element-wise addition | Element-wise multiplication | Vector Concatenation | LSTM-Based Fusion | CNN-Based Fusion | |
| VQA-AA [150] | 2014 | | √ | | | | |
| LSTM Q + I [10] | 2015 | | √ | | | | |
| ABC-CNN [30] | 2015 | √ | | | | | |
| iBOWING [40] | 2015 | | | √ | | | |
| Neural-Image QA [42] | 2015 | | | | √ | | |
| VIS + LSTM [50] | 2015 | | | | √ | | |
| Image_QA [51] | 2015 | | | | √ | | |
| mQA [55] | 2015 | √ | | | | | |
| SMem-VQA [61] | 2015 | √ | | | | | |
| LSTM + Attention [12] | 2016 | | √ | | | | |
| Word + Region Sel [29] | 2016 | | | | | | |
| SAN [34] | 2016 | √ | | | | | |
| MCB [45] | 2016 | | | | | | √ |
| MLB [47] | 2016 | | | | | | √ |
| DPPnet [54] | 2016 | | | | | √ | |
| FDA [62] | 2016 | | √ | | | | |
| Re_Baseline [155] | 2016 | | | √ | | | |
| Answer_CNN [53] | 2016 | | | | | √ | |
| MFB [25] | 2017 | | | | | | √ |
| High-Order [31] | 2017 | | | | | | √ |
| MUTAN [148] | 2017 | | | | | | √ |

but closely related image. They performed point-wise multiplication of extracted visual and text features to obtain a resulting language + image representation 256-dim. The obtained fused vector goes through two more FC-layers to produce a two-way softmax score for the answers 'yes' and 'no'.

Shih et al. [29] used bounding boxes to extract image features with the help of a CNN. Visual features, text features, and set of the multiple-choice answers were given as an input to the VQA system. The score is produced for every multiple-choice answer, and answer with the highest score is finally selected. Dot product is used to combine region-wise visual features and question features. Zhou et al. [40] used GoogLeNet to extract image features and represented the question in a traditional way using bag-of- words. Both the feature vectors are the concatenated and given as input to a multi-class logistic regression classifier to predict the answer.

Kafle et al. [41] proposed a model that understands a question (color, object or counting) and guesses the form of an answer. They used the concept of Bayesian framework for formulating the answer. Question is encoded using skip thought vectors. For image, they use a ResNet. During the training process, each question is assigned the type of

**Table 9**
Fusion Strategies used in recent SOTA models (2018–2020).

| Model | Year | Simple Vector Operation | | | Neural Networks | | Bilinear Models |
|---|---|---|---|---|---|---|---|
| | | Element-wise addition | Element-wise multiplication | Vector Concatenation | LSTM-Based Fusion | CNN-Based Fusion | |
| CVA [28] | 2018 | √ | | | | | |
| MAN [33] | 2018 | | | √ | √ | | |
| MFH [49] | 2018 | | | | | | √ |
| QGHC [58] | 2018 | | | | | √ | |
| DCN [65] | 2018 | | | √ | | | |
| FVTA [67] | 2018 | | | √ | | | |
| VKMN [80] | 2018 | | √ | | | | |
| NMN + Cap. Info. [149] | 2018 | √ | | | | | |
| Attention-on-Attention [156] | 2018 | | √ | | | | |
| MLB + DA-NTN [161] | 2018 | | | | | | √ |
| CATL-QTA-M [197] | 2018 | | | √ | | | |
| StrSem [200] | 2018 | | | √ | | | |
| DRAU [32] | 2019 | | | | | | √ |
| MTA [206] | 2019 | √ | | | | | |
| VRR + Att [88] | 2020 | √ | √ | | | | |
| LRN [89] | 2020 | | | √ | | | |
| ODA-GCN [90] | 2020 | | √ | | | | |
| MOVRD [91] | 2020 | | √ | | | | |
| EMQA [92] | 2020 | | | √ | | | |
| DecomVQANet [93] | 2020 | | √ | | | | |
| MDFNet [94] | 2020 | | | | | | √ |
| IASSM[95] | 2020 | | √ | | | | |
| ALSA [96] | 2020 | | √ | | | | |
| QLOB [97] | 2020 | | | √ | | | |
| SelRes + SelMask + bbox [98] | 2020 | √ | | | | | |
| QC-MLB [99] | 2020 | | | | | | √ |
| SANMT [100] | 2020 | | √ | | | | |
| RSVQA [101] | 2020 | | √ | | | | |

answer it can expect. However, questions which are of other types are difficult to solve. Malinowski et al. [42] gave an approach which they called Neural-Image-QA. They applied all the three simple vector operations to fuse both visual and question features. Element-wise multiplication operation gave the highest accuracy.

Saito et al. [43] proposed a concept of integration of two operations which are the element-wise sum and element-wise product by using a polynomial function. Image features and textual features are fused using these two operations. They integrate element-wise summation and element-wise multiplication by implementing a polynomial function. They named it 'DualNet'. They considered the VQA task as a classification task having a prebuilt set of answers. Lioutas et al. [44] proposed a multimodal fusion method by concatenating image, question and answer features where element-wise multiplication is first used to fuse the question and answer features.

### 4.2. Bilinear models

In bilinear pooling method, the outer product between two vectors is computed to provide all possible interactions between each element of both the vectors. A huge amount of memory and high computation cost is occurred in order to compute the interaction between each pair of visual feature and textual feature vectors. To handle the issue of large parameters, [45] proposed Multimodal Compact Bilinear Pooling (MCB) to compact bilinear models. They used count sketch function [46] and Fast Fourier Transformation (FFT) to avoid computing outer products explicitly. Fukui et al. [45] used a pooling approach to fuse both the modalities, i.e. image and question features. They randomly projected the visual and text (question) features to a space of higher-dimensional and then used outer product operation to convolve both vectors in the Fourier space for better efficiency. Thus they performed "Multimodal Compact Bilinear pooling" (MCB) as shown in Fig. 9 and model is depicted in Fig. 10.

Kim et al. [47] used a Hadamard product and a linear mapping for achieving low-rank bilinear pooling (MLB). They gave the reason to use the Hadamard product that MCB is computationally very expensive. Kim et al. [47] reduced the number of parameters involved in MCB by proposing multimodal low-rank bi- linear pooling method (MLB). The basic concept in MLB is modify the weight matrix by the product of two small matrices [48], that is, $W = UV$, where, $W \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{d\times n}$and $d \leq \min(m,n)$. Thus, the projected feature can be computed as:

$$f_i = V_{img}^T W_i V_{ques} = 1^T \left(U_i^T V_{img} \odot V_i^T V_{ques}\right) \quad (12)$$

To reduce the order of weight tensors by one, 1 is replaced by $P \in R^{d\times c}$:

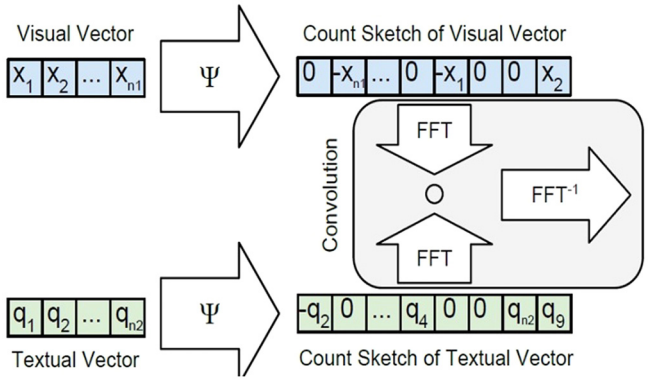



**Fig. 9.** Multimodal Compact Bilinear Pooling (MCB) for joint representation of visual and question features [45].

$$f_i = P^T \left(U_i^T V_{img} \odot V_i^T V_{ques}\right) \quad (13)$$

The results obtained by MLB are more accurate than MCB as shown by [32]. The drawback of MLB is that it has slow convergence rate and susceptible to hyper-parameters.

An improved version of MCB is MFB (Multimodal Factorized Bilinear Model) [25]. MFB maintains the dense output representation of MCB together with improved training stability. The visual and question features in MFB are first extended to high-dimensional space and further compressed by applying sum pooling into dense output feature. In comparison to MLB, MFB needs more parameters and is capable to find out more prevailing features.

Yu et al. [49] used a novel approach for the fusion of both image vector and question vector. For improving the convergence rate of MLB, Yu et al. [49] fused feature vector by matrix factorization method and named it as Multi-modal Factorized High-order pooling approach (MFH) method by reducing the parameters require. In MFH, element-wise multiplication is applied to the output of the current MFB and cascaded features by preceding blocks.

### 4.3. Fusion based on neural networks

In this fusion technique, non-linear neural networks are employed to fuse both the visual and question features.

#### 4.3.1. LSTM-based fusion models

VQA researchers used LSTM, a class of recurrent neural network to represent a question. In LSTM based fusion techniques, the idea is to consider visual feature as one of the words of a question and incorporate this visual feature into the same dimension as the question embedding.

Ren et al. [50] used LSTM based fusion strategy by projecting extracted visual feature $V_{img}$ into the same word embedding space and considering $V_{img}$ as the first term in the question. Thus, ($V_{img}$, $V_{ques_1}$, $V_{ques_2}$, ……, $V_{ques_m}$) is provided as input to the LSTM model, where $V_{ques_i}$ is i-*th* word embedding in the question.

Another fusion strategy was discussed by Malinowski et al. [42] for improving the interaction between visual features and question words. They concatenated visual features with all the question word embeddings as ([$V_{img}$,$V_{ques_1}$], [$V_{img}$,$V_{ques_2}$], ……, [$V_{img}$,$V_{ques_m}$]). Malinowski et al. [42] gave an approach which they called Neural-Image-QA. They have used LSTM to implement a recurrent neural network (RNN) which was used to handle question and multi-word answers.CNN is used to extract image features which are pre-trained for object recognition. Both the obtained feature vectors are first passed to an LSTM for encoding purpose. The output of the first encoder is a fixed size feature vector that is further given to a decoder LSTM. This decoder LSTM produces answers of variable length. This process goes on until <END> symbol is generated. In their work, the failure cases in VQA task occurs due to strong occlusion, unusual instances of objects or answer that is not in the set of ground truth answers.

Ren et al. [51] proposed the VQA task as a classification problem and modeled the visual semantic embedding to map image data to semantic space. They proposed variation by embedding visual features both at the start and end positions like ($V_{img}$, $V_{ques_1}$, $V_{ques_2}$, . . ., $V_{ques_m}$, $V_{img}$) and provided as input to the LSTM model. They also provided an option to add reverse LSTM to make RNN as bi-directional. Encoder LSTM produces a feature vector that was directly fed into a classification algorithm to generate one-word answers from a prebuilt vocabulary. They also proposed an algorithm to generate questions from description sentences. Algorithm implementation in this work is strongly dependent on the type of question. Also, it is difficult to say why the algorithm is giving a particular answer. Reference [33] uses memory-augmented neural network [52] for maintaining a relatively long memory.

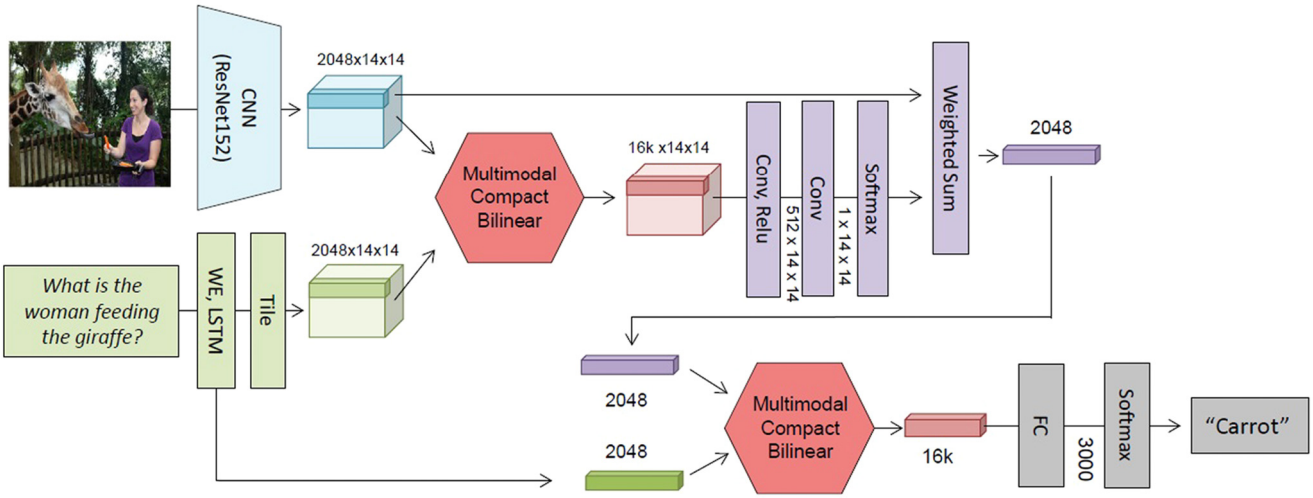



**Fig. 10.** Architecture of Multimodal Compact Bilinear (MCB) model with Attention mechanism. Conv means convolutional layers and FC means fully connected layers [45].

*4.3.2. CNN-based fusion models*

In the previous subsections, visual features are treated as an individual word. But it can be noticed that this strategy may not capture the complex relationship between the image and semantic concepts as the effect of image will disappear at each time step of LSTM. For handling this issue, an alternative fusion was used based on CNN model [53]. Multimodal convolution model is used to merge both visual and question features to effectively capture the interaction between visual and question features. Ma et al. [53] proposed the use of three CNNs for VQA task. Question features were extracted through the first CNN. The second CNN was used to extract image feature. The third CNN is used to fuse the both feature vectors thus formed an overall homogeneous convolutional architecture.

Noh et al. [54] used VGG-16 for extracting visual feature by eliminating last layer and incorporating three fully-connected (FC) layers. Dynamic parameters obtained from GRU for extracting question features were used to replace the second FC layer. The parameters of this layer are decided at run time depending upon question. They used hashing method for prediction weights in this layer. They employed GRU, which was pre-trained on a large text corpus.

Gao et al. [55] proposed that the semantic representation of a question is done by a Long Short-Term Memory (LSTM) using word embedding layer. Visual features are extracted by CNN. The linguistic context in an answer is stored using another LSTM. These three features vectors are further fused using element-wise addition to generate the answer. The answer generated by this method can be a sentence, a phrase or a single word. However, this model is not able to answer correctly when the targeting objects are too small or looks similar to other objects or model requires high level commonsense or reasoning based on the fact from daily life.

Kim et al. [56] proposed the concept of deep residual learning for VQA task and to learn the joint representation of images and question. It utilized multimodal inputs, and allowed a deeper network structure. The fusion of question vector and the visual feature vector is performed by element-wise multiplication. However, there are some limitations' of this model. It does not handle counting questions effectively. Also, it finds objects irrespective of the given question.

Lao et al. [57] proposed residual learning inspired end-to-end VQA model. They proposed multi-modal fusion architecture for achieving multiple interactions between image and question features without mounting the learning parameters. The key point of this network is that is that features are fused at every step rather than fusing them at last step.

Gao et al. [58] proposed a new fusion approach based question-guided hybrid convolution. Convolutions are performed on image feature maps and question features are used to predict the convolution kernels. The key idea of their approach is to perform the fusion of image and question features at an early phase of VQA methods to preserve additional information such as spatial-relations among objects in the image.

Andreas et al. [59] presented a method for developing and learning neural module networks (NMN), that are used further to create a set of jointly-trained neural "modules" into deep network for visual question answering (VQA). Zhu et al. [12] provided bounding boxes for the entity mentioned in the question-answer sentences with entire grounding explanations and a new QA type is introduced with image regions as visually grounded responses.

*4.4. Attention based models*

In Joint embedding based methods, global features are captured to represent image content. However, this may misguide the algorithm for a particular VQA task. Attention models handle these shortcomings. By focusing on the most relevant image regions that are related to the question asked. Attention models were earlier used in mage captioning task by Xu et al. [60]. Spatial attention is implemented by putting a rectangular grid over- all image regions and extracts the local image feature from each grid. Another way to do this is to find bounding boxes over the image and then each of these boxes is encoded using a CNN and then the bounding box is selected according its relevance with respect to the question asked.

Yang et al. [34] proposed multiple-layer stacked attention model in which they allow reasoning iteratively to find the answer. The model first focuses on all entities such as bicycle, basket and dogs in the first attention layer. In the second attention layer, it narrows down the focus to dogs as the answer The model may fail if it focuses on the wrong region. Fig. 11 shows the SAN model.

Xu et al. [61] proposed a method that attaches words with image regions in the first hop. In the second hop whole question guided attention is added to verify the visual results obtained from the first hop. They called this method as a Spatial Memory Network (SMN).

Shih et al. [29] used bounding boxes to extract image features with the help of a CNN. Visual features, text features, and set of the multiple-choice answers were given as an input to the VQA system. The score is produced for every multiple-choice answer, and answer with the highest score is finally selected. Ilievski et al. [62] proposed to identify regions depending upon the keywords in the questions. These

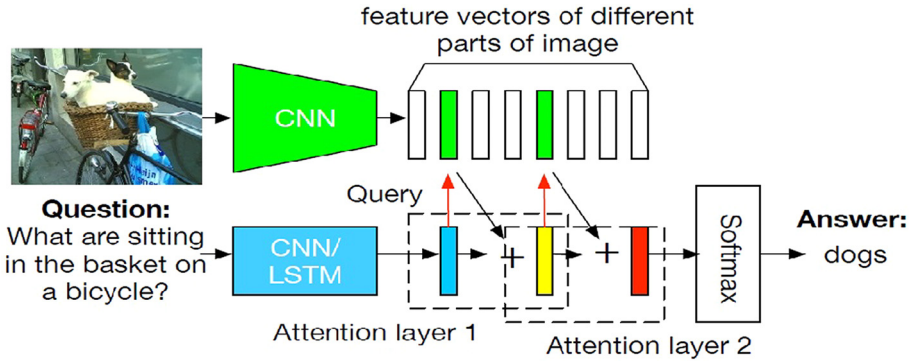



**Fig. 11.** SAN model architecture [34]. The model first focuses on all entities such as bicycle, basket and dogs in the first attention layer. In the second attention layer, it narrows down the focus to dogs as the answer.

keywords reflect the object in the image with their corresponding object labels.Word2vec was used to measure the similarity between object label and words in the question.

Lu et al. [63] proposed a "hierarchical co-attention model" in which attention on questions is also employed in addition to visual attention. In this method, a question is encoded at various level i.e. word level, multiword level and question level. Chen et al. [30] proposed a question-guided attention map (QAM) instead of word guided attention. They searched for image features that relate to the semantics of the text in the input question in the spatial image feature map. Zhu et al. [12] provided bounding boxes for the entity mentioned in the question-answer sentences with entire grounding explanations and a new QA type is introduced with image regions as visually grounded responses. Zhang et al. [11] did a balancing of VQA dataset of abstract binary scenes by adding complementary scenes so that all binary questions have one answer as "yes" for one scene and have an answer as "no" for another but closely related image.

Vinay et al. [64] proposed model having phases. The first phase is to generate a caption for an input image, divides the image into regions. This model then generate caption for each region of the image and combines these generated captions into a dense caption .In the second phase, the VQA task is just reduced to a textual question answering task. All region explanations are considered as a context for question-answer pair related to the image. Yu et al. [27] used semantic attention for finding language linked concepts from the image and context-aware image attention for finding language-related regions and visual representation of these regions is learned by the model. It fills the semantic gap between question and image, and learns detailed explanations from image regions.

Anderson et al. [26] used a combination of both bottom-up and top-down image attention approaches. The bottom-up process extracts all objects or other salient regions (attention candidates) independent of question using faster R-CNN. Top-down process weights each attention candidates given the task-specific context. Fig. 12 shows the bottom-up and top-down attention model.

Song et al. [28] used a model based on Cubic Visual Attention (CVA) by applying a spatial attention and a new channel on object regions to enhance VQA task. Osman et al. [32] proposed recursive attention units for producing soft attention for visual question answering task. Li et al. [35] suggested updating the question encoding by choosing image regions frequently related to the words in a question and thus predict the accurate answer.

Duy et al. [65] proposed a dense co-attention mechanism which calculates each relation between the words with any region of the image. In their approach, generation of the attention map on regions for each word and generating one attention map on words for every region was performed. They are stacked to form a multi-layer hierarchy for interactions between an image and question pair.

Teney et al. [66] proposed an approach that uses object detection methods on VQA algorithms. The model focused on the features and thus gives better attention to images. An R-CNN network is used to improve performance over other architectures. Junwei et al. [67] proposed the task question answering on a given set of personal photos and answer questions that are related to some past events captured in the pictures. They used a MemexQA dataset [68]. It includes of 20,860 questions related 13,591 personal photos belonging to 101 real Flickr users. These photos contain a variety of important moments of their live such as holiday trips, birthday parties, marriage functions, etc. Shah et al. [69] proposed a cycle consistency training scheme in question answering and question generation task and also proposed large scale VQA rephrasing dataset which consists of three rephrasing for approximately 40,000 questions on approximately 40,000 images from VQA 2.0 dataset. Ilievski et al. [70] suggested an adversarial self-learning based generative attention model. Kim et al. [71] utilized bilinear attention networks (BAN) which compute bilinear attention distributions for using seamless information between language and vision.

### 4.5. Compositional models

These approaches consist of linking different modules that are proposed for a particular task such as memory or a particular type of reasoning. The question in a VQA task may require many steps of reason to answer accurately. For example, a question of type" What is above

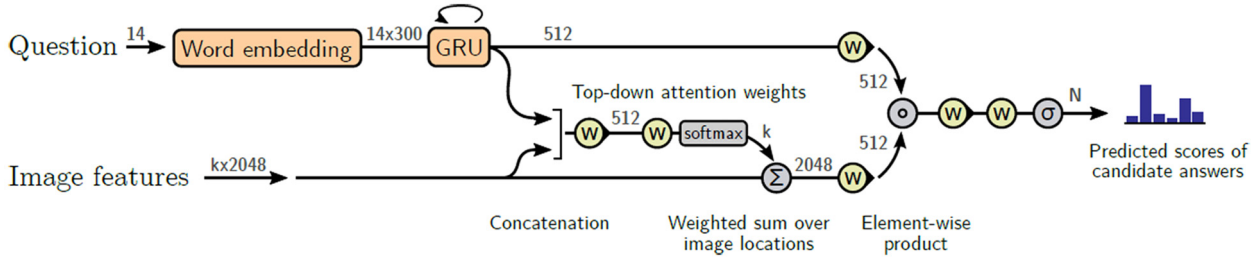



**Fig. 12.** VQA model based on combine top-down and bottom-up visual attention mechanism [26].

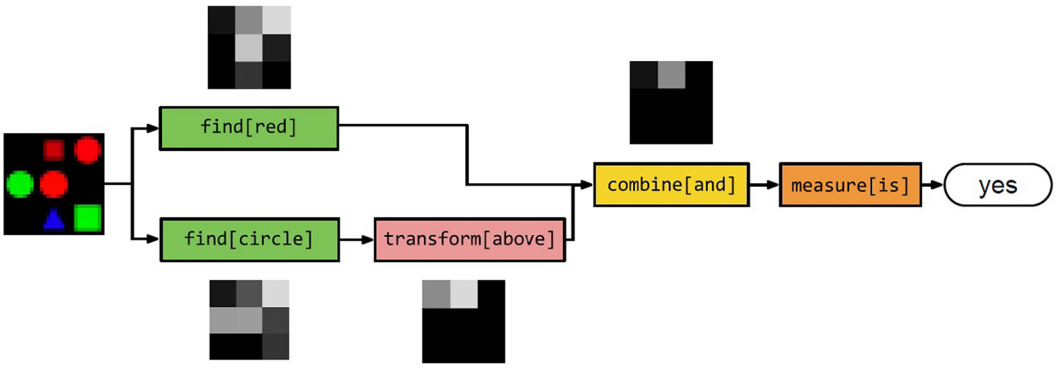



**Fig. 13.** Neural network module for answering the question Whether there is a red shape above a circle? The model uses two find modules, the transform [above] module, combine module and measure[is] module to predict the answer [59].

the table?" requires first to find the table and then naming the object above the table.

The architecture of the NMN views VQA problem as a sequence of sub-problems performed by independent neural sub-networks. Every sub-network carry out a particular and well-developed task, these modules must then be brought together into a meaningful layout. Andreas et al. [59] employed a natural language parser on the given question to both the sub-problems in question and to decide the needed design of the sub-tasks when the sub-tasks are accepted. The parsing of question is done using the Stanford dependency parser [72] that finds the relationship between different words of a given sentence. Andreas et al. [73] proposed a model that can be used for both images and ordered information bases. This model utilizes natural language sentences to automatically make neural networks from a set of defined modules. Fig. 13 explains the framework of neural network module used for answering questions.

Without relying on parser with some external language, Noh et al. [74] proposed a RAU model which can implicitly perform compositional reasoning. They used several self-contained response units in their design which can answer VQA sub-problems. They gave a novel approach for the VQA task based on a recurrent deep neural network, where each component in the network communicates to an entire answering unit with attention method by own. They used multiple passes over the internal memory-like unit and find a loss over each of these passes, instead of a single loss. At test time, the inference is carried out by only using one such pass, after the training process was completed.

Dynamic Memory Networks (DMN) [75] is a type of neural networks with modular architecture. It consists of four modules input module, question module, the episodic memory module and the answer module. Xiong et al. [76] used the DMN for VQA task. Image features are extracted using the input module with a VGGNet [2].Then the features extracted are given to a GRU. To focus on a particular region, the episodic memory module is used.

Yi et al. [77] designed a model specifically designed for CLEVR dataset. The images are transformed into organized features and their original root question approach is transferred to the question features. They used these features to filter out the required response. Remi et al. [78] presented a MuRel, a multimodal relational network that learnt reason over real images in end to end way. They represented interactions between question and image regions using a dense vector representation called it a MuRel cell and incorporate these cells into the MuRel network to define finer visualization details.

Zellers et al. [79] proposed a visual commonsense reasoning task (VCR). It combines two flavors of commonsense inference: answering and explanation. Also introduced a VCR [79] dataset which consists of 290 k multiple-choice questions. They also proposed recognition to cognition networks to perform layered inference required for VCR. Ma et al. [33] utilized the concept of memory-augmented neural networks for generating answers that occur rarely in training set more accurately. Su et al. [80] presented the idea of incorporating human knowledge in structured form and visual features in memory networks for VQA task.

### 4.6. Models using external knowledge bases

Sometimes the task of answering the question requires information beyond the information given in the image such as commonsense reasoning or task-specific knowledge or encyclopedic information. For example, questions like "Which image is most related to a chef?" one must have the context knowledge about the chef. For answering such question, we linked the VQA models with knowledge bases like DBpedia [81].

Wang et al. [82] proposed a VQA model called "Ahab" which in turn uses concepts from DBpedia [81]. Ahab first identifies relevant concepts from the image and then matches it to semantic concepts available in a knowledge base. A query is generated for the natural language question given related to an image which is executed over the combined image and knowledge base concepts. Wang et al. [83] proposed a method named FVQA, which is an improvement of [82]. They used an LSTM and image-question mapping to find the significant content in an image and then generate a query based on question over the image and the knowledge base. They also used two more knowledge bases, one was ConceptNet and the other was WebChild. Wu et al. [84] used CNN for extraction of semantic attributes. DBpedia gives knowledge about these extracted attributes as it contains short descriptions. DocVec is used to embed these descriptions into fixed-size vectors. An LSTM model uses these vectors as the input that infers the question and finally generates an answer. Fig. 14 explains the VQA model using external knowledge base for generating answers.

Tatiana et al. [85] used Visual Madlibs dataset for answering fill in the-blank type multiple-choice questions. They also proposed a method for localizing phrases from possible answer to give spatial support for extracting features. Hudson et al. [86] proposed a GQA dataset for real-world visual reasoning and compositional question answering. It consists of 100 K real images and 22 M compositional questions. In this dataset, each image comes with a scene graph and each question comes with a program. Questions require diverse reasoning skills and sequence id steps for answering questions.

### 4.7. Recent state-of-the-art models

Yu et al. [87] represent multimodal knowledge by employing different knowledge graphs corresponding to semantic, visual and fact-based perspectives. Also, they proposed a memory-dependent recurrent module for multistep knowledge reasoning over graph-structured

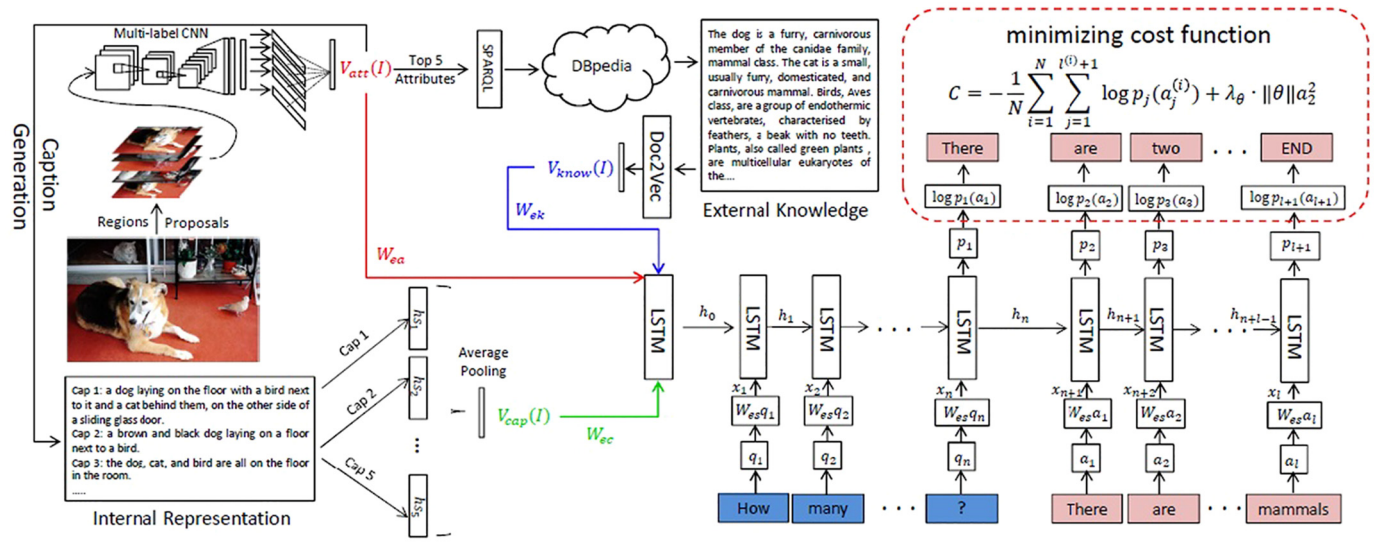



**Fig. 14.** VQA model using external knowledge from DBpedia to generate the answer [84].

multimodal information. The reasoning module (GRUC) is consists of Read, Update and Control units, which maintains more open and structured reasoning.

Zhang et al. [88] proposed a visual relation reasoning module to capture the relationship between different image regions. They employed bilinear attention together with bottom-up attention to get more significant attention maps. Finally, a multi-label classifier is used to predict the answer. The model is shown in Fig. 15 (a).

Sun et al. [89] proposed a local relation networks (LRNs) which capture deeper semantic relationship information and thus produces context-aware visual features for every image region. Further, they employ multi-level attention to combine the information obtained from LRNs to converge both original image and relational information. The model is shown in Fig. 15 (b).

Zhu et al. [90] proposed a graph based model to capture the relationship between objects which are obtained under the guidance of questions. They added soft attention layer in the graph convolution procedure to obtain question-based objects and thus handle object redundancy problem. Also, they proposed a difference-based graph learner which considers the difference between objects and semantics of questions to find the edges between graph nodes. The model is shown in Fig. 16 (a).

Xi et al. [91] proposed a VQA models that detects multiple relationships between objects. They used word vector similarity principle to

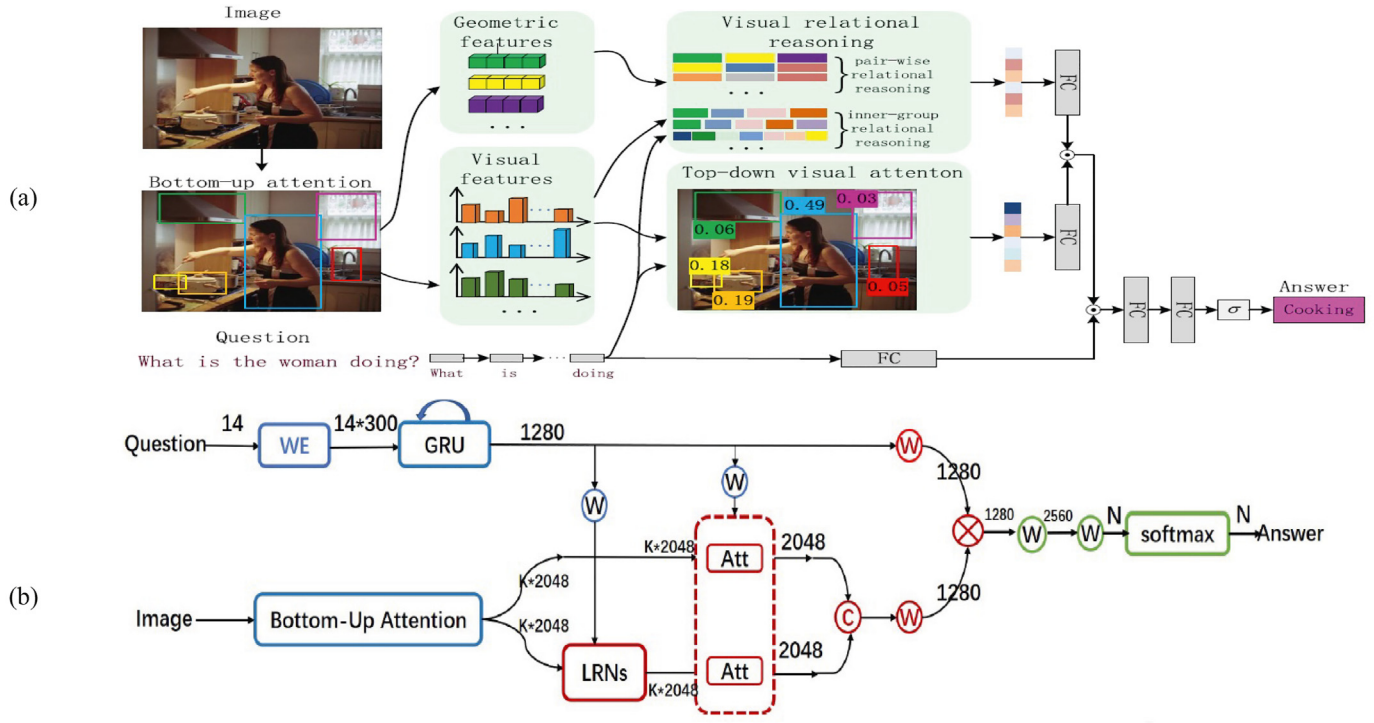



**Fig. 15.** (a) A visual relational reasoning and top-down attention based VQA model (Zhang et al. 2020). (b) The architecture of the LRN model. Symbol "⊗" denotes the Hadamard product [89].

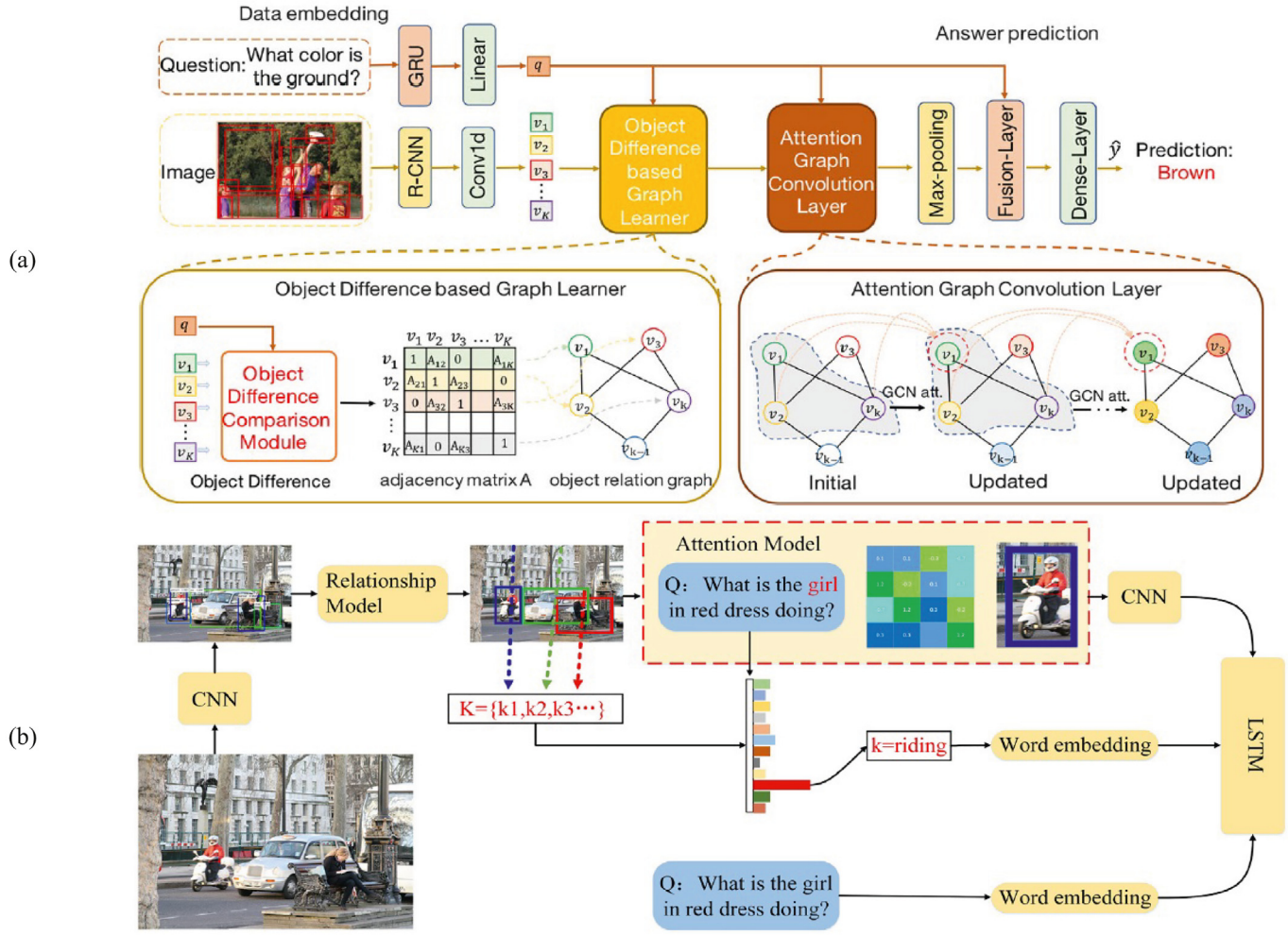



**Fig. 16.** (a) The framework of ODA-GCN model having four main modules: Data embedding, Object-Difference based Graph Learner, Attention Graph Convolution Layer, and Answer Prediction [90]. (b) The architecture of the multi-objective visual relationship detection based VQA model (MOVRD) [91].

capture the relationship between objects. Word mover's distance algorithm is used to compute the similarity between word vectors. Question-guided attention mechanism is used to focus on the regions of the image. The model is shown in Fig. 16 (b).

Hosseinabad et al. [92] propose a multiple-answer VQA model based on sliding window to find the answer to a given question corresponding to different image regions. Also, they proposed a new dataset named as "ICon Question Answering (ICQA)" for training and evaluating their VQA model. Their training approach is called glimpse training approach which can answer object based questions. Their models fail when questions are related to abstract meaning of scene.

Bai et al. [93] proposed an approach to compress and accelerate VQA systems. They applied different decomposition and regression methods such as Tucker, Canonical Polyadic and Tensor Train for decomposing FC-layers of CNN and LSTM. Tensor regression layer is used to replace FC-layer and flattening layer. Tensor Contraction layer is employed to compress the parameters.

To perform fine-grained multimodal fusion, Zhang et al. [94] proposed Multimodal Deep Fusion Network (MDFNet). They proposed Graph Reasoning and Fusion Layer (GRFL) to capture the semantic and spatial relationship between objects and combine both the relations effectively. Further, many GRFLs were used to build a Multimodal Deep Fusion Network, to improve the multimodal fusion mechanism. The model is shown in Fig. 17 (a).

Liu et al. [95] used inferential attention and semantic space mapping to build a VQA model named as IASSM. Semantic space module is used to the combination of both the labeled and unlabeled answers set. Thus, a new set of answers are generated by combining a question and visual features. On the other hand, inferential attention module is used to capture the correlation between question and image, to simulate human attention. Both these modules are combined into an end-to-end model to generate the answer. The model is shown in Fig. 17 (b).

Liu et al. [96] used two supervised attention mechanisms; one is free form-based attention and other is detection-based attention learnt from prior-knowledge of human-annotated attention maps to build a novel VQA model called adversarial learning of supervised attentions (ALSAs). Further, an adversarial network is constructed between these attention modules to answer question related to foreground objects and background forms.

Gao et al. [97] proposed a VQA model by capturing question semantics, detailed object information and the correlation between these two modalities. The model is called as Question-Led Object Attention (QLOB). Question model is used to capture the semantics of question sentence. Object detection network is used to extract local and global image features by applying top k object region proposals. Further, QLOB attention module is employed to choose object regions based on questions. Softmax classifier is used to optimize question model and QLOB attention to generate the final answer. The model is explained in Fig. 18 (a).

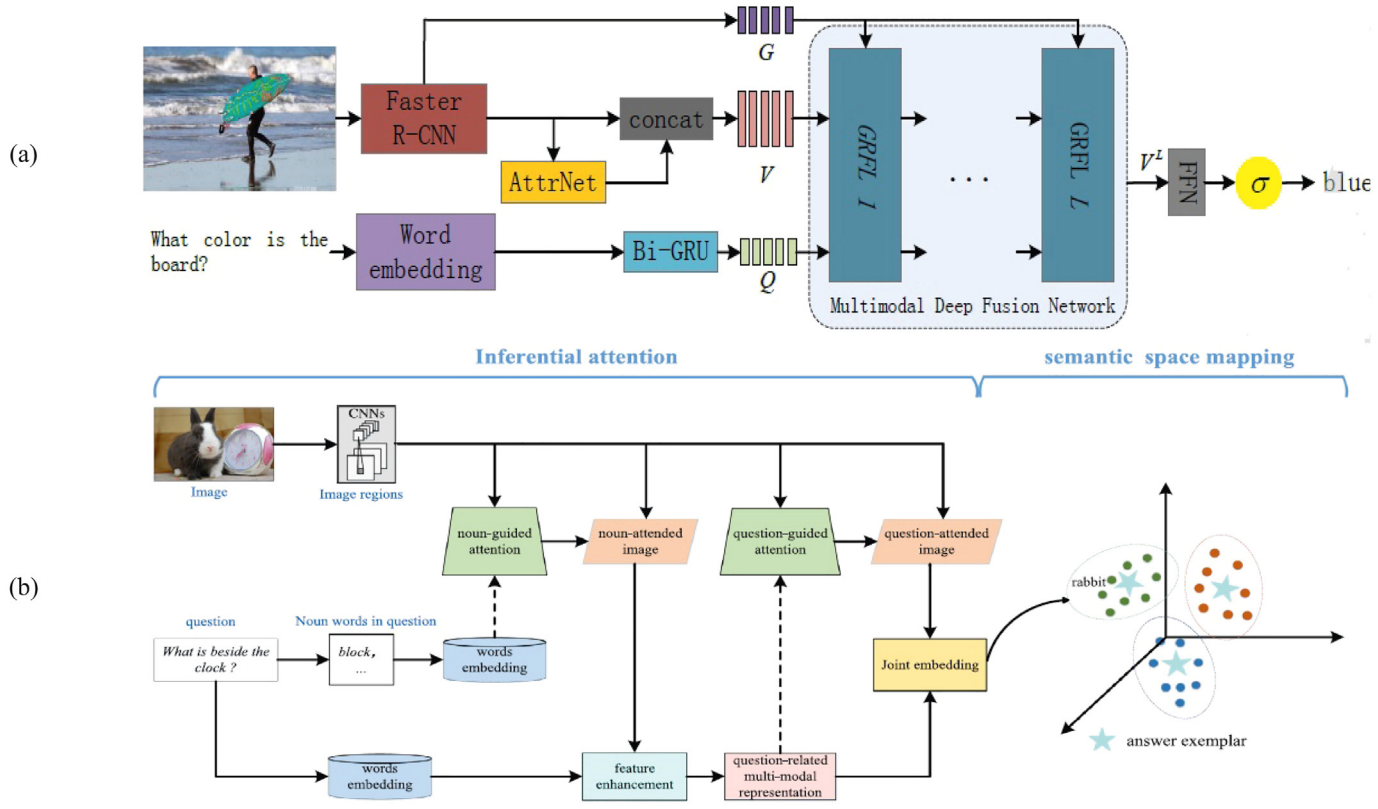



**Fig. 17.** (a) The architecture of MDFNet model using graph reasoning and fusion layers (GRFL) [88]. (b) The framework of the IASSM model, which mainly contains two modules, i.e., inferential attention and semantic space mapping [95].

Hong et al. [98] proposed a VQA model named as Selective Residual learning (SelRes) which uses self-attention mechanism. The model applies the residual learning to capture most important relationships. They also proposed Selective masking for masking attention maps corresponding to the significance of previous stack's vector. The model is explained in Fig. 18 (b).

Vu et al. [99] proposes a VQA model namely Question-Centric Multimodal Low-rank Bilinear (QC-MLB), applied in the field of medical imaging. The model combines question and visual features guided by question sentence. They have also shown that the proposed VQA model can be combined with CAM-like techniques to emphasize which regions of the image are exploited by the model to predict the answer. The model is explained in Fig. 19 (a).

Zhong et al. [100] presented a novel feed-forward encoder-decoder pipeline using Self-Adaptive Neural Module Transformer (SANMT) by substituting feed-forward encoder-decoder pipeline. To encode question features, they employed transformer module to create dynamic question feature embedding which grows over reasoning steps. Also, they used the intermediate results to choose the question features and for dynamically adjusting the layout selection. The model is explained in Fig. 19 (b).

Lobry et al. [101] proposed a VQA system names as RSVQA to capture the high level information from remote sensing data. They have developed two datasets image/question/answer triplets using low resolution and high resolution data. OpenStreetMap (OSM) is used to construct the question and answers.

Sharma et al. [195] proposed a VQA model that uses graph neural network for improve image representation by capturing the relationship between the objects followed by context-aware attention model to predict an answer. Sharma et al. [212] used image captioning module to improve the accuracy of the visual question answering task. The VQA model uses the knowledge learnt from the image captioning module and uses this knowledge to improve the accuracy of answer prediction.

Chen et al. [213] proposed a VQA model which is able to predict the scan-paths during the answer generation process. They have integrated a task regulation map to generate a series of the model learns to predict a sequence of task-oriented scan-paths which direct to right or wrong answers. Whitehead et al. [214] proposed a skill-concept composition, which uses both supervised VQA idea with self supervised learning, for more accurate and effective evaluation for VQA models on real image datasets. Urooj et al. [215] came with an idea of visual capsule component which uses a query-dependent selection method having capsule features, for focusing on significant regions depending on textual information contained in a given question. Zhang et al. [216] proposed a way of improving the representations of visual information for tasks such as VQA and image captioning. They proposed a method to represent an image in based on objects present an image in an improved way.

## 5. Datasets

For research on VQA, various datasets have been proposed purposely. Since 2014, many significant datasets for VQA have been freely made available. These datasets empower VQA frameworks to be trained and assessed. These datasets contain minimum three elements, i.e. an image, a question, and a correct answer to that question. Questions within the datasets may have different complexity levels. Many questions require common sense and reasoning over the visual information given in an image to conclude the correct answer. The major datasets for VQA are DAQUAR [109], COCO-QA [50], the VQA Dataset [11], FM-IQA [55], Visual Genome [110], Visual7W [12], Shapes [111], KB-VQA [82], FVQA [83], Visual Madlibs [112], CLEVR [177], FigureQA [208], DVQA [209], Diagram [210], TDIUC [211], VizWiz [122], VQA-Med [128] and ICQA [92].

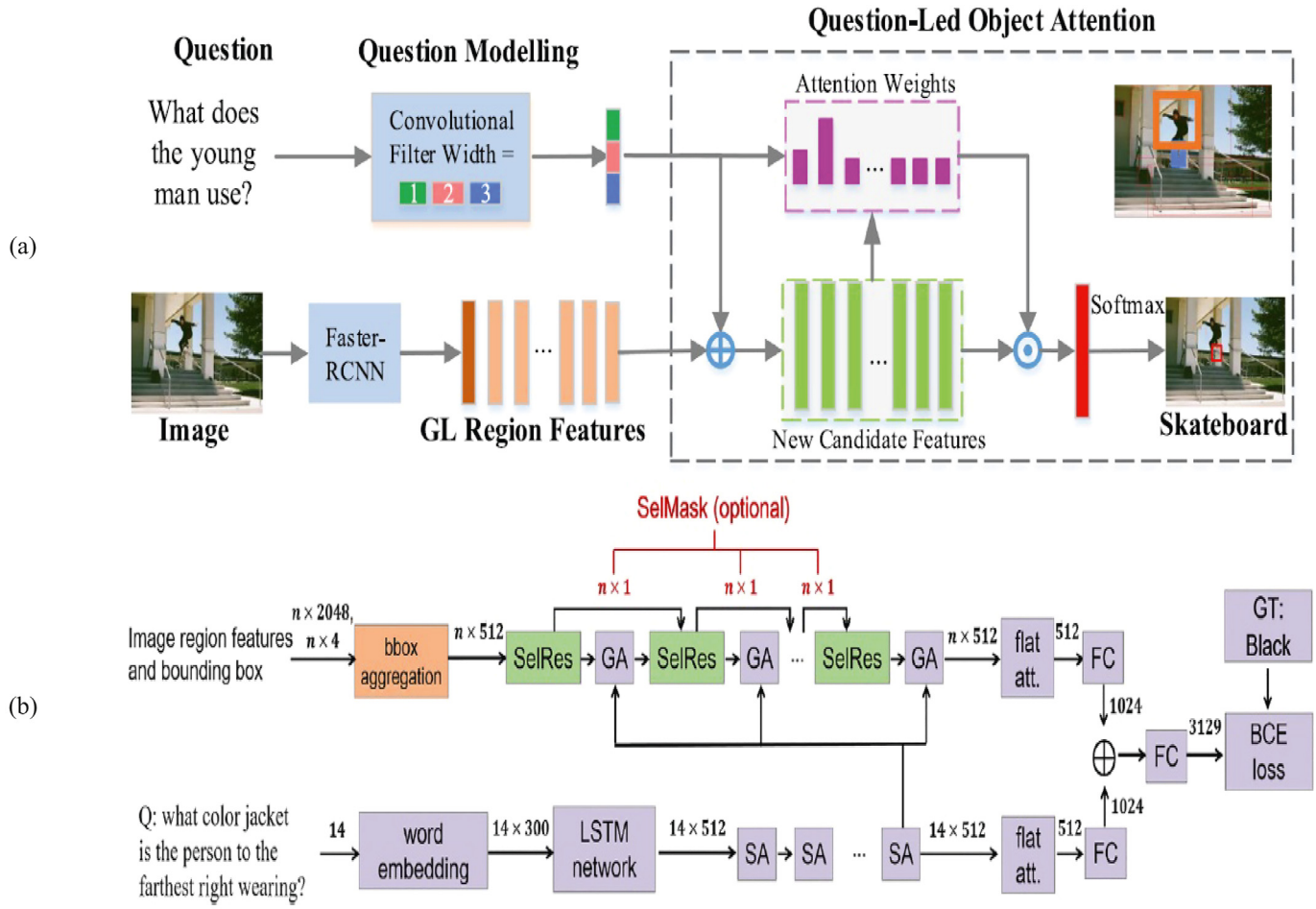



**Fig. 18.** (a) The framework of QLOB model. There are mainly four parts: (1) image feature extraction; (2) question encoding; (3) attention layer; and (4) answer generation [97]. (b) The framework of SelRes + SelMask VQA model, which has three main modules: (1) SelRes is the selective residual learning module. (2) SelMask is the selective masking module. (3) bbox aggregation is the bounding box aggregation module [98].

Most of the datasets include images from the Microsoft Common Objects in Context (COCO) dataset [113], which consists of 328,000 images, 91 common object classes with over 2 million labeled instances, and in general five captions per image on an average, except the DAQUAR dataset. Apart from COCO images, Visual Genome and Visual7W also make use of images from Flickr100M.In the subsequent subsections; we critically assess the available datasets and also present their limitations. Table 10 shows the statistics of VQA datasets. Figs. 20–33 shows the images from the different datasets and corresponding QA pairs.

### 5.1. DAQUAR

The Dataset for Question Answering on Real-world images (DAQUAR) [109] was designed as the first major VQA dataset used as the benchmark. It is considered as one of the smallest datasets for VQA tasks. It is based on NYU-DepthV2 Dataset [114] images; consists of overall 12,468 questions-answer pairs out of which 6795 QA pairs are used for training and 5673 QA pairs for testing. The images in this dataset are divided into 795 for training purpose and 654 for testing purpose, thus contains a total of 1449 images of indoor scenes. There are a total of 894 object classes which are assigned to every pixel of the image.

DAQUAR-37 dataset is a smaller version of DAQUAR that consists of only 37 object classes. It consists of overall 4122 question/answer pairs out of which 3825 QA pairs are used for training and 297 QA pairs for testing. There are certain limitations of this dataset. First, though the DAQUAR was first dataset for VQA task, it does not effectively train and assess VQA models with higher complexity due to its small dataset size. Second, the variety of questions available is limited as this dataset contains exclusively indoor scenes and third, many questions are challenging to answer due to extreme lighting conditions in some cases and cluttered images.

### 5.2. COCO-QA

In COCO-QA [50], Natural Language Processing (NLP) algorithms are used for generating QA pairs, based on the MS-COCO image captions. The MS-COCO dataset contains five descriptions for an image in a single sentence. Suppose we have a caption for an image as a girl is riding a bicycle, we can generate a question as “What is the girl riding?” with a natural language answer as a bicycle. It consists of overall 1, 17,684 question-answer pairs out of which 78,736 QA pairs are used for training and 38,948 QA pairs for testing. The questions in this dataset broadly belongs to four categories: object (69.84%), color (16.59%), number (7.47%) and location (6.10%) [9]. The total number of images in the dataset is 1, 23,287. The main limitation of COCO-QA is to generate QA pairs using NLP algorithm, which have flaws. Many questions in COCO-QA have grammatical errors and are absurd. Another limitation is the high repetition of questions due to automatic conversion from captions. The other major limitation is that the dataset contains only four types of questions, and these questions are bounded to the class of things depicted in COCO's captions.

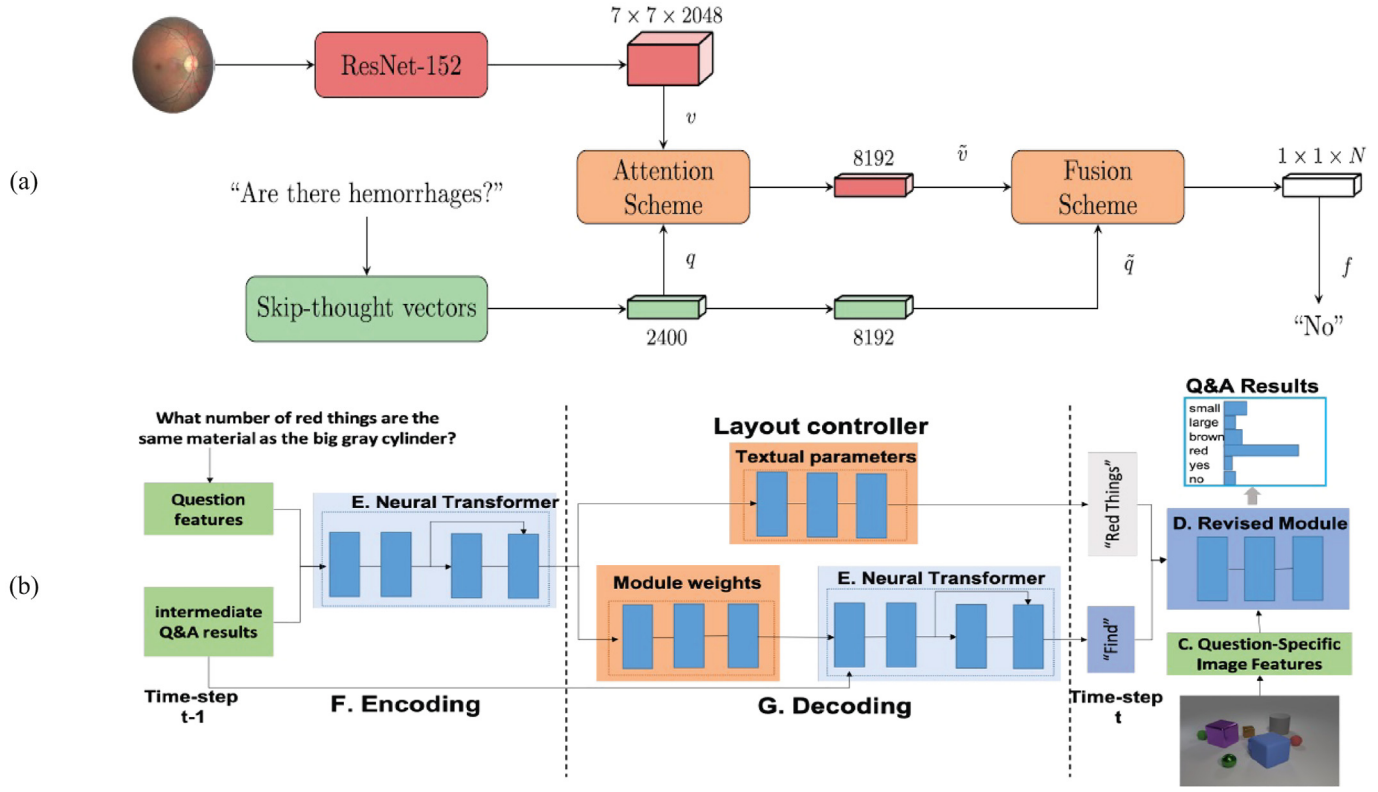



**Fig. 19.** (a) The architecture QC-MLB model together with the attention mechanism [99]. (b) The architecture of self-adaptive encoder-decoder mechanism based VQA model (SANMT) [100].

### 5.3. The VQA dataset

The VQA Dataset [11,115] includes both real images from MS-COCO and abstract clipart images. It is most widely used dataset for VQA task and released publically as a part of VQA challenge. VQA dataset for real images consists of overall 6, 14,163 questions out of which 2, 483, 49 questions are used for training 1, 21, 512 for validation and 2, 44, 302 for testing. Questions for each image were generated by Amazon Mechanical Turk (AMT), and answers to these questions were given by a different group of workers. This dataset includes three questions corresponding to an image and ten answers corresponding to a given question. Ten independent annotators answered to each question given to them. The total number of images is 2, 04,721. The longest question consists of 32 words and the longest answer is of 20 words. VQA dataset for clipart images consists of 50,000 abstract images with 1, 50,000 questions. These abstract images are made from over 20

**Table 10**
VQA dataset and their statistics.

| Dataset Name | Image Source | # Images | # Questions | Type of Questions | Question Length | Answer Length | OE/MC | Question Collection |
|---|---|---|---|---|---|---|---|---|
| DAQUAR [109] | NYU-Depth V2 | 1449 | 12,468 | 4 | 11.5 | 1.2 | OE | Both |
| COCO-QA [50] | MS-COCO | 117,684 | 117,684 | 4 | 8.6 | 1 | OE | Auto |
| VQA [10] | MS-COCO | 204,721 | 614,163 | 20+ | 6.2 | 1.1 | Both | Manual |
| FM-IQA [55] | MS-COCO | 158,392 | 316,193 | – | 7.38 | 3.82 | OE | Manual |
| Visual Genome [110] | MSCOCO, YFCC | 108,000 | 145,322 | 7 | 5.7 | 1.8 | OE | Manual |
| Visual7W [12] | MS-COCO | 47,300 | 327,939 | 7 | 6.9 | 1.1 | MC | Manual |
| Shapes [111] | Synthetic Shapes | 15,616 | 244 | – | – | – | Binary | Auto |
| Visual Madlibs [112] | MS-COCO | 10,738 | 360,001 | 12 | 6.9 | 2 | Fill-in-the blanks | Manual |
| KB-VQA [82] | MS-COCO | 700 | 2402 | 23 | 6.8 | 2 | OE | Manual |
| FVQA [83] | MS-COCO, ImageNet | 2190 | 5826 | 12 | 9.5 | 1.2 | OE | Manual |
| CLEVR [177] | Synthetic Shapes | 100,000 | 999,968 | 90 | – | – | OE | – |
| VizWiz [122] | – | – | 31,173 | – | 6.68 | 1.66 | OE | Spoken |
| VQA-Med [128] | – | 4200 | 15,292 | 4 | – | – | OE | – |
| Diagrams [210] | – | 5000 | 15,000 | – | – | – | MC | – |
| TDIUC [211] | MSCOCO, YFCC100M | 167,437 | 1,654,167 | 12 | – | – | OE | Both |
| FigureQA [208] | – | 100,000 | 1.3 M | – | – | – | MC | – |
| DVQA [209] | – | 3,000,000 | 3,487,194 | 3 | – | – | OE | – |
| OK-VQA [187] | MSCOCO | 14,031 | 14,055 | – | 8.1 | 1.3 | OE | Manual |
| GQA [86] | Visual Genome | 113,018 | 22,669,678 | – | – | – | OE | Manual |
| ICQA [92] | Synthetic Shapes | 42,021 | 260,840 | – | – | – | OE | Auto |

Question: How many objects are of white color?
Answer: 9

Question: What is on the right side of cabinet?
Answer: Bed

**Fig. 20.** Images from DAQUAR [13] dataset with corresponding question answer pairs.

Question: What is the color of horses?
Answer: Brown


Question: How many beer bottles are there?
Answer: Three

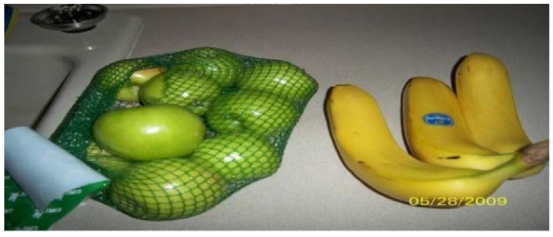
**Fig. 21.** Images from COCO-QA [14] dataset with corresponding question answer pairs.

Question: What is the shape of bench?
Answer: oval, semi circle, curved, curved, double curve, banana, curved, wavy, twisting, curved

Question: What color is the stripe on the train ?
Answer: white, white, white, white, white, white, white, white, white, white

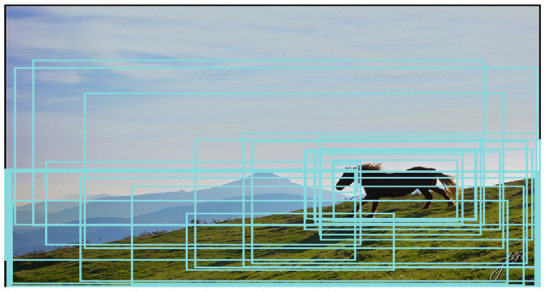
**Fig. 22.** Images from VQA [11] dataset with corresponding question answer pairs. Every question has 10 answers, each answer is provided by different annotators.


Question: What is the color of bus?
公共汽车是什么颜色的？
Answer: Red 公共汽车是红色的。




Question: Which fruit is of color yellow?
黄色的是什么？
Answer: Banana 香蕉

**Fig. 23.** Images from FM-IQA [15] dataset with Chinese and English question answer pairs.



Question: What is the color of sky?
Answer: Blue

Question What is the color of the clock ?
Answer: Green

**Fig. 24.** Images from Visual Genome [16] dataset along with annotated image regions with bounding boxes as the choices.

Question: When is the function in image?
Answer: Wedding

Question How many women under umbrella?
Answer: Two

**Fig. 25.** Images from Visual7W [16] dataset.

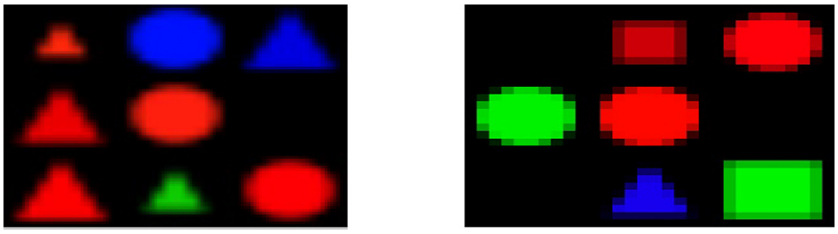


**Fig. 26.** Image from Shapes [18] dataset. Questions can be type counting (How many circles are there?), spatial reasoning (Is there a green shape below a circle?), and inference (Is there a blue shape circle?).

Question: What is the common characteristic of the animal in this image and elephant?
Answer: mammal, African animals

Question: Is it a tourist place ?
Answer: Yes

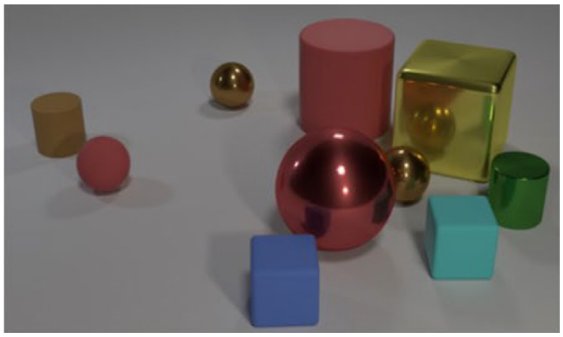
**Fig. 27.** Images from datasets using knowledge bases. KB-VQA [20] dataset.

Question: What are eatables in the image?
Answer: apples


Question: Which thing is used for romantic party?
Answer: Wine

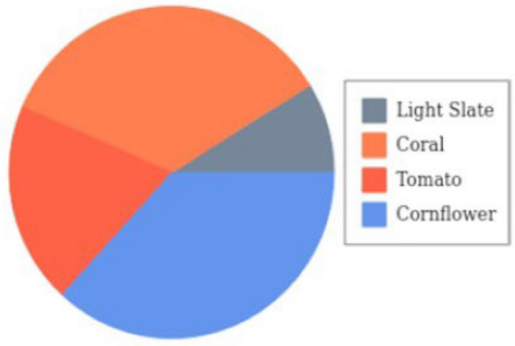
**Fig. 28.** Images from datasets using knowledge bases. FVQA [20] dataset.


The place is a ground.
Person B has frisbee.

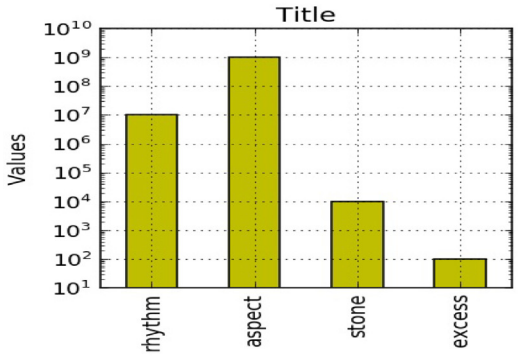


Question: Are there equal number of cylinder and circle?
Answer: No

**Fig. 29.** Visual Madlibs (Left) and CLEVR (Right) datasets.




Question: Is cornflower maximum?
Answer: Yes




Question: Which bar has the smallest value?
Answer: excess

**Fig. 30.** FigureQA (Left) and DVQA (Right) datasets.

human cartoon models.100 different objects and 30 different animal models.

This dataset has both open-ended and multiple-choice questions for both types of images, i.e. real and clipart images. Eighteen different choices are also provided for the multiple-choice questions. These questions consist of the same QA pairs. The choices provided are consisted of:

**Correct Answer**: It is the most common answer given by the ten independent annotators.

**Plausible Answers**: It consists of 3 answers gathered from independent annotators without seeing the image.

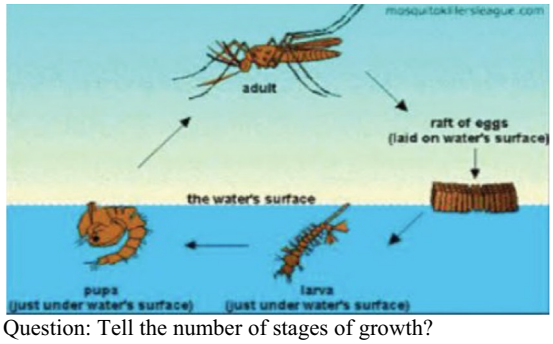
**Popular Answers**: These are the top 10 most popular answers provided in the dataset.

**Random Answers**: These are the randomly preferred correct answers for other type of questions.

However, there are many limitations of this dataset. First, due to language biases, many questions can be answered correctly without considering the images. For example, binary questions cover 38% of all questions, out of which 59% of them have 'Yes' as the answer. It is very hard to say that an algorithm is really explaining a VQA task or just guessing the answer.

### *5.4. FM-IQA*

The Freestyle Multilingual Image Question Answering (FM-IQA) [55] dataset is based on MS-COCO. In this dataset, answers and questions are




Question: Tell the number of stages of growth?
Answer: 5

Question: Is the dog jumping?
Answer: Yes

**Fig. 31.** Diagram (Left) and TDIUC (Right) datasets.

Question: Is it a sunscreen?
Answer: Yes

Question: What is shown in CT-scan?
Answer: bilateral multiple pulmonary nodules.

**Fig. 32.** VizWiz (Left) and VQA-Med (Right) datasets.

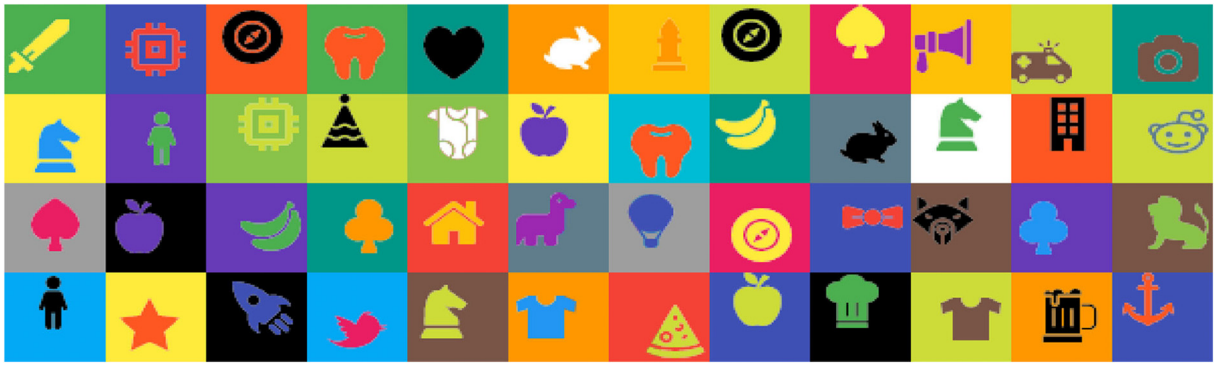


**Fig. 33.** Few iconic shape utilized in the ICQA dataset.

generated by humans. The QA pairs in the dataset are available in both Chinese and English language. It uses the Baidu crowd sourcing server to construct questions and answers. Answers in this dataset are full sentences. This dataset includes a broad range of AI-related questions that requires commonsense reasoning over visual content (e.g. "Why does the bus park here?).This dataset consists of 1, 58,392 images and 3, 16,193 QA pairs which were originally available in Chinese and later on translated into English. Due to this, automatic evaluation using general metrics is difficult. Consequently, the authors proposed utilizing human judges for assessment, where the judges are entrusted with choosing whether or not the appropriate response is given by a human or not and also evaluating the nature of an answer on a scale of 0–2. This methodology is illogical for most research gatherings and makes algorithms difficult.

### 5.5. Visual genome

Visual Genome dataset [110] consists of 108,249 images and 1.7 million QA pairs are available for images. For an image, 17 QA pairs for an image are available on an average. Visual Genome is one of the largest datasets for VQA task. In Visual Genome dataset, questions can start with six Ws': What, Where, How, When, Who, and Why. The dataset was made by collecting data by two different modes. For open ended free form questions, annotators can ask any question related to an image. In this dataset, two types of questions are available: Region-based and free form open-ended. In free form questions, human annotator is shown an image and asked to generate 8 QA pairs. In region based questions, the human annotator must provide QA pair for a specific region in the image.

Visual Genome dataset has a much bigger range of answers in comparison to other datasets. The top 1000 answers that have a high probability of occurrence in Visual Genome only cover 65% of all answers that is present in the dataset. The diversity in answers that exist in the dataset involves challenges in the evaluation of open-ended questions. Furthermore, since the question categories themselves are expected to belong exclusively to one of the six 'W 'forms, the response heterogeneity may sometimes artificially result merely from differences in phrasing that could be avoided by encouraging annotators to select more descriptive responses. This dataset has no binary (yes/no) questions.

### 5.6. Visual7W

The Visual7W [12] dataset is a subset of Visual Genome dataset. This dataset has 47,300 images from Visual Genome which are also available in MS-COCO. Visual7W contains seven categories of questions: What, Where, How, When, Who, Why, and which. The dataset includes two different types of questions. Questions about 'telling 'are the same as questions in Visual Genome dataset, and their response is text-based. The questions that start with 'Which' are considered as the 'point' questions and the system will pick the accurate bounding box between choices available for these questions.

In this dataset, the questions are assessed in a multiple-choice format, with four candidate answers to each question, only one of which is correct. However, all the objects listed in the questions are visually grounded, i.e. aligned in the images with bounding boxes of their depictions. Again the dataset does not contain binary question like Visual Genome dataset.

### 5.7. SHAPES

The SHAPES dataset [111] consists of objects of various arrangements, forms, and color. Questions are about the characteristics, relationships, and locations of the shapes. It emphasizes on learning of spatial and logical relations among different objects. This method makes it possible to build a vast amount of data, free of many of the limitations that affect other datasets to varying degrees.

Shapes consist of 244 unique questions and 15,616 images in total. All questions are binary questions with yes or no answer. This dataset is fully balanced and does not have language biases.

### 5.8. KB-VQA

The KB-VQA dataset [82] has been developed for the purpose of evaluating the performance of Ahab VQA algorithm [20]. It contains questions that require knowledge about a particular which is present in DBpedia. Because of the rich contextual information and different object classes in it, 700 images were selected for the validation from COCO image dataset [113] and 3 to 5 QA pairs were collected for each image and thus contain 2042 questions in total. The images were chosen to cover about 150 classes of objects and 100 classes of scenes, usually showing 6 to 7 objects each.

### 5.9. FVQA

The commonsense knowledge is used to answer questions in FVQA dataset [83]. A lot of supporting-facts (commonsense knowledge) are provided to the concepts in the given images. These facts are formulated as a triplet (val, rel, val2).Object, scene and action are the different visual concept found in this datasets. The understanding of each visual concept is taken out from a range of existing structured knowledge bases, such as DBpedia [81], ConceptNet [116] and WebChild [117,118].

Human annotators chose an image and a visual content of the image and then selected one of the pre-extracted supporting facts relevant to the visual definition. Finally, they had to present a QA pair that explicitly contains the supporting facts selected. FVQA dataset includes 193,005 candidate supporting facts related to 580 visual concepts (234 objects, 205 scenes and 141 attributes) having a total of 4608 questions.

### 5.10. Visual Madlibs

To evaluate systems on a "fill in the blank" task, the Visual Madlibs dataset [112] is designed. The goal is to identify words to complete a statement describing a given image. Multiple choices are also given as an extra evaluation benchmark. The dataset contains 10,738 images from COCO and 360,001 directed natural language explanations. Incomplete sentences are automatically generated from these explanations. Both open-ended and multiple-choice assessments are possible.

### 5.11. CLEVR

CLEVR (Compositional Language and Elementary Visual Question Reasoning) is a dataset similar to SHAPES that is it is a collection of 100, 000synthetic images of 3D shapes such as spheres and cylinders. The questions included in this dataset are used to test the visual reasoning capabilities of a VQA model. There are different categories of questions associated with each image. In training set, there are 70,000 images with 699,989 question-answer pairs. The validation set and test contains 15,000 images with 149,991 and 14,988 question answer pairs respectively.

### 5.12. FigureQA

FigureQA dataset contains graphical plots and figures with 5 classes. These classes are line plots, dot-line plots, bar graphs both horizontal and vertical and pie charts. There are 15 categories of questions which are used to find different relationships between objects in graph. These questions can be used to inspect properties such as the maximum, the minimum, smoothness, area-under-the-region, and intersection.

### 5.13. DVQA

DVQA (Data Visualization Question Answering) is a synthetic dataset which is used to evaluate the different aspects of bar charts only. There are three categories of questions in this dataset: structure understanding, data retrieval and reasoning. Example of structure understanding question is "Are the bars vertical?" Example of data retrieval question is "What is label of second horizontal bar from right?" Example of data retrieval question is "Which algorithm has highest accuracy for VQA dataset?"

### 5.14. Diagram

Diagram (AI2D) dataset focuses on evaluation of diagram interpretation capabilities of VQA systems. It contains more than 5000 diagrams which represents grade school science, each annotated with component segmentation, their associations to each other and connection to the diagram canvas. There are annotations for more than 118 K components and 53 K associations in AI2D dataset. There are more than 15,000 multiple choice questions related to the diagrams. There are 4000 images in training set and 1000 images in blind test set.

### 5.15. TDIUC

TDIUC (Task Directed Image Understanding Challenge) dataset has 12 question-types that represent traditional computer vision tasks and a set of new high level tasks which requires reasoning capabilities. Yes/no object presence detection related questions are balanced. The dataset includes absurd questions to verify whether the question is valid fo the given image. The questions in TDIUC dataset are taken from COCO-VQA, Visual Genome and human annotators.

### 5.16. VizWiz

VizWiz is the first goal-oriented VQA dataset to handle the issue of blind users. It is originated from visually impaired users. The images captured by blind users are generally of poor quality. The questions in the dataset are collected in spoken forms and may suffer from auditory imperfections. Many questions in the dataset are unanswerable as the blind users cannot validate the captured images and their visual content.

### 5.17. VQA-Med

VQA-Med dataset is first step towards medical domain VQA. VQA-Med dataset contains medical images with medically relevant question-answer pairs. The interpretation of medical images through patient engagement is improved by the success in this task. Also, the doctors can take second opinion in case of complex images. Semi-automatic methods were used to generate question-answer pairs. The questions were first generated using rule-based methods followed by manual authentication by human experts.

### 5.18. ICQA

In Icon question answering (ICQA) dataset, there are about 100 different $30 \times 30$ resolution iconic shapes from the Internet. It also defines 21 different colors for these shapes and their backgrounds. The authors of ICQA have created multiple sets of data. Set A includes 260,840 questions corresponding to 42,021 images. Set B includes 226,406 questions corresponding to 42,300 large images, while set C includes 5408 questions corresponding to 1000 images (Fig. 34).

## 6. Extension of VQA models for scene-text reading

Apart from visual content, images may contain a lot of other semantic information that can be further utilized for visual question answering task. Text is one of the semantic information which commonly exists in the natural scene. Closely related to VQA task, Gupta et al. [102] integrates textual cues with the visual content for image captioning using deep CNN and LSTM. Bai et al. [103] exploits scene text for visual understanding for fine grained classification of images. They used attention mechanism to establishing the relationship between textual and visual content. Dey et al. [104] proposed a framework for understanding Ad Images by extracting both visual and textual information present in the images.

**Table 11**
Statistics of TextVQA dataset.

| Number of Images | 28,408 |
|---|---|
| Number of Questions | 45,336 |
| Number of unique answers | 26,263 |
| Average length of question | 7.18 |
| Minimum length of question | 3 |
| Answer for each QI pair | 10 |
| Source of images | Open Images v3 dataset |

Singh et al. [105] proposes a new dataset called as TextVQA containing 45,336 questions on 28,408 images that require scene text detection and reasoning over textual and visual content to answer the questions. TextVQA has collected all the images from Open images dataset. It has 37,912 unique questions from a total of 45, 336 questions. The average length of question in this dataset is 7.18. The minimum question length is 3. The unique answers are 26,263. It has 21,953 training, 3166 validation and 3289 test set images. It also proposes a novel model of combining a usual VQA model with an OCR module which is independently trained. This module has a "copy" which is based on pointer networks that permit to utilize words recognized by OCR as the predicted answers if required. Table 11 shows the statistics of TextVQA dataset.

Biten et al. [106] proposed a method of answering questions depending upon the text present in an image. They also proposed a new dataset for visual question answering called Scene Text VQA (ST-VQA) [23]. The ST-VQA dataset include images from different publically available datasets such as ICDAR 2013 [119] and ICDAR2015 [120], ImageNet [121], VizWiz [122], IIIT Scene Text Retrieval [123], Visual Genome [110] and COCO-Text [124]. It includes 23,038 images from these six datasets which are both related to general computer vision datasets and scene text understanding datasets. It contains total 31,791 questions/answer pairs from these datasets. Out of these 19,027 images and 26,308 questions are used for training and 2993 images and 4163 questions for testing. Table 12 shows the statistics of ST-VQA dataset. Table 13 depicts the image sources used for creating ST-VQA dataset.

The images in OCR-VQA-200 K dataset are collected from the dataset constructed by Iwana et al. [171]. This dataset contains cover images for books, book author names, book titles and the categories of books. The categories of a book can be art, religion, science, comics etc. The statistics of the OCR-VQA dataset is shown in Table 14. The questions are prepared by asking questions related to author names, title of book, edition of book etc. To make the variations in questions, paraphrasing of question is performed. For example a question like 'who is the author of the book?' can be paraphrases as 'who wrote the book?'

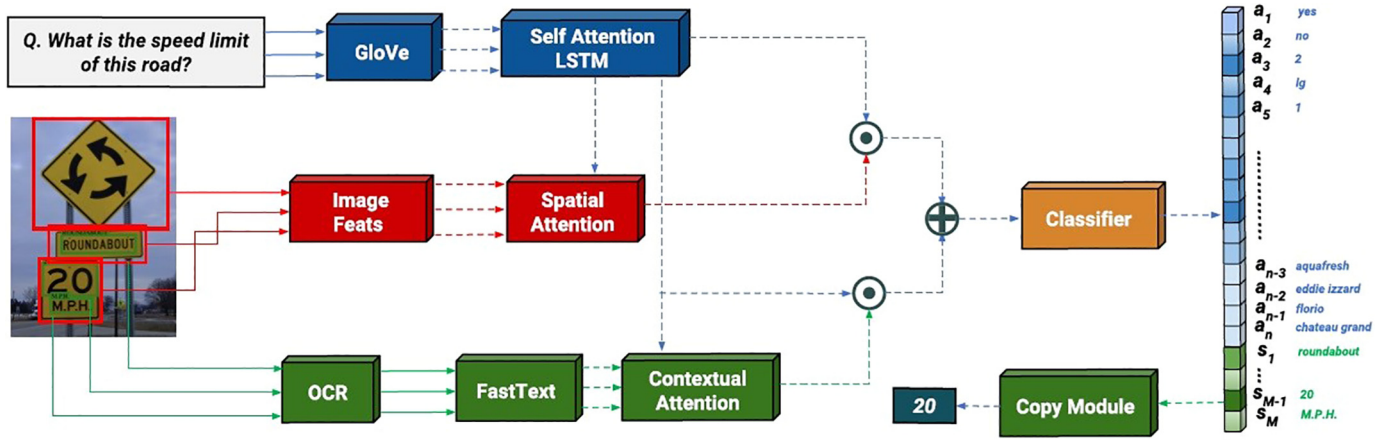



**Fig. 34.** Framework of Look, Read, Reason & Answer (LoRRA) [105]. The model is able to read the text in images and predict answer either from the fixed answer vocabulary or by selecting one of the OCR token.

**Table 12**
Statistics of ST-VQA dataset.

| | |
|---|---|
| Number of Images | 23,038 |
| Number of Questions | 31,791 |
| Number of QA pairs | 31,791 |

**Table 13**
Image Sources in ST-VQA dataset.

| Source Dataset | No. of Images | No. of questions |
|---|---|---|
| COCO-Text | 7520 | 10,854 |
| Visual Genome | 8490 | 11,195 |
| VizWiz | 835 | 1303 |
| ICDAR | 1088 | 1423 |
| ImageNet | 3680 | 5165 |
| IIIT-STR | 1425 | 1890 |

## 7. Evalaution metrics

It is really a complex task to evaluate natural language sentences generated automatically by a VQA system. It is necessary to take into account both the syntactic (grammatical) and semantic correctness. A question in a VQA task can be open-ended in which a system has to generate a string to answer the question or a multiple-choice where the system selects a choice out of given choices.

Simple accuracy can be used to evaluate the multiple-choices VQA task when an algorithm gets the right answer if it makes a correct choice. Simple accuracy can also be used to evaluate an open-ended VQA task when the predicted answer is given by an algorithm exactly matches with ground truth answer. There is a limitation of this simple accuracy metric as it requires an exact match. Consider the question asked about an image as, 'What fruits are present in the image?' and the algorithm outputs 'apple' but the correct label is 'apples,' it is considered wrong and it will be considered equally wrong when the system outputs 'mango'.

$$Accuracy = \frac{\#questions\ answered\ correctly}{\#total\ questions} \tag{14}$$

The second evaluation metric is Wu-Palmer Similarity (WUPS) [125] that was given as a substitute to simple accuracy. This metric aims to evaluate the difference between an answer predicted by the algorithm and the available ground truth answer in the dataset depending upon the difference in their semantic connotation. Based on their similarity to each other, WUPS will allocate value between 0 and 1 depending upon ground truth answer in the dataset and the answer predicted by the algorithm to a question. For example, apple and apples have a similarity score of 0.98, whereas apple and fruit have a similarity score of 0.86.

**Table 14**
Statistics of OCR-VQA-200 k dataset.

| | |
|---|---|
| Number of Images | 207,572 |
| Number of QA pairs | 1,002,146 |
| Number of unique answers | 320,794 |
| Average length of question | 6.46 |
| Average length of answer | 3.31 |
| Average number of questions per image | 4.83 |
| Number of unique authors | 117,378 |
| Number of unique titles | 203,471 |
| Number of unique genres | 32 |

$$WUPS(a,b) = \frac{1}{N_Q}\sum_{i=1}^{N_Q} \min\left\{ \prod_{a\in P_A} \max_{t\in G_A} WUP(a,t), \prod_{t\in G_T} \max_{a\in P_A} WUP(a,t), \right\} \cdot 100 \tag{15}$$

Here, $N_Q$: total number of questions, $P_A$: set predicted answers, $G_A$: set of ground truth answers and $WUP(a,b)$: It will return the location of words' 'a and 'b' based on taxonomy tree in relation to the location of the Least Common Subsumer (a, b).

There are certain limitations of WUPS metric that makes it difficult to use in VQA task. First, a certain set of words are lexically very alike, but still, they may have very different meaning. This problem may arise in color question. For example, if an answer to a certain question is white and the system predicts black as an answer, this answer will still get a WUPS score of 0.92, which seems high. Another limitation is WUPS cannot be used for phrases or sentence answers as it always works with rigid semantic concepts which are most likely to be single words.

Another way to evaluate the VQA system is by collecting multiple independent ground- truth answers for every question. This is called consensus metric. It was followed for VQA dataset [11]. In VQA dataset, ten ground-truth answers were collected for each question by ten different subjects. This Evaluation is done on VQA dataset by comparing a generated answer with these ten ground truth answer given by ten different subjects as given by:

$$Accuracy_{VQA} = \min\left(\frac{A\ specific\ answer\ is\ given\ by\ \#\ subjects}{3}, 1\right) \tag{16}$$

The above equation implies that an answer is considered 100% accurate if at least three subjects provide that answer. This metric again has certain limitations. First, it can allow two correct answers for some questions. Second, it is very expensive to collect ground truth answer for each question. Third, for 'Why' type of question, inter-human consensus is poor as it is really very difficult that three subjects will give exactly same answer.

An alternative to evaluate a VQA system is to use human judges to assess multi -words answers as suggested by developers of FM-IQA dataset. But it requires a lot of time, resources and is really expensive. It can include the subjective opinion of each human involved in the process. Multiple choice paradigms can be an alternative way to evaluate multi-word answers as used in VQA dataset, Visual7W and Visual Genome. In this, a system has to just select which of the given choices is correct instead of generating an answer.

One of the key limitations of VQA datasets is unbalanced question types distribution. For rarer question types, simple accuracy is not effective evaluation metric. So Kafle and Kanan (2017a) proposed a mean-per-type (MPT) evaluation metric for handling unbalanced question-type distribution. MPT indicates the evaluated arithmetic or harmonic mean accuracy for each question type. They also suggested to use normalized metrics, for example arithmetic normalized MPT and harmonic normalized MPT, to tackle biasness in distribution of answers per question type.

The metric proposed by [106] is Average Normalized Levenshtein Similarity (ANLS) defined as:

$$ANLS = 1 - dis_L(ans_{pred}, ans_{GT} / \max([ans_{pred}], [ans_{GT}]) \tag{17}$$

Where, ($ans_{pred}$ and $ans_{GT}$ are predicted answers and ground-truth answers respectively and $dis_L$ is the edit distance) which are averaged over all questions. The scores which are less than the threshold value 0.5 are truncated to value 0 before calculating the average.

BLEU (BiLingual Evaluation Understudy) proposed by Papineni et al. (2002) and METEOR (Metric for Evaluation of Translation with Explicit ORdering) proposed by Denkowski and Lavie (2014) are uses as

evaluation metrics for automatic assessment of machine translation. Gurari et al. (2018) discussed that both metrics can be employed for VQA task and tested with the VizWiz dataset. BLEU examined the co-occurrences of n-grams between ground truth label and the predicted answer. Usually, it is not suitable for short sentences. On the other side, METEOR can be used by finding the alignment between the GT answer words and the predicted answer words. Sometimes, this one-to-one correspondence is difficult to capture.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{18}$$

$$METEOR = (1 - Pen) * F_{mean} \tag{19}$$

Consistency metric evaluates responses consistency across diverse questions. When a new question is given, a VQA system does not contradict with its previous answers. The validity metric verifies whether a given answer is in the scope of question, e.g. replying as a fruit to a fruit related question. The plausibility score verifies whether a generated answer is reasonable or justifies, given the question (e.g. cat usually do not drink, say, wine).The distribution metric computes the alignment between ground truth answer distribution and the system generated distribution by applying Chi-Square statistic [207]. This metric is used to analyze whether the model is predicting the most frequent answer together with less frequent ones.

Accuracy is not found to be efficient measure for biased data. It means that if we have any class either positive or negative is more in the given input data, then accuracy is not accepted as a significant measure. Therefore, F-measure is used to evaluate the weighted mean involving precision and recall. Assuming $t_p$, $t_n$, $f_p$, and $f_n$ are consecutively true positive, true negative, false positive and false negative to compare a single question answers with the ground truth, the F1 measure can be calculated as follows:

$$F_1 = 2 \bullet \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{20}$$

Here, precision can be given as:

$$\text{Precision} = \frac{t_p}{t_p + f_p} \tag{21}$$

and recall can be estimated as:

$$\text{Recall} = \frac{t_p}{t_p + t_n} \tag{22}$$

F1 Score needs to be maximized.

Table 15, shows the evaluation metrics for VQA task and the supporting datasets.

## 8. Results

### 8.1. Comparison of the SOTA methods on VQA 2.0 dataset

Table 16 shows the comparison of the SOTA VQA models on VQA 1.0 dataset. In the VQA v1.0 dataset, the models are trained on the train + val sets and tested on the test-dev and test-std sets. For VQA v1.0 dataset, the evaluation method given by equation 16 is used as the metric. On the test-dev set, it can be observed that ALSA model (Liu et al. 2020) attained the highest accuracy of 69.32% in Open-Ended task and 73.67% in Multiple-Choice task. The second best performance is attained by the IASSM model (Liu et al. 2020) for both Open-Ended and Multiple-Choice task. As for the test-std set, again the ALSA model (Liu et al. 2020) attained the highest accuracy of 69.52% in Open-Ended task and 73.61% in Multiple-Choice task.

**Table 15**
CNN Models used for extraction of image features in VQA models.

| Evaluation Metric | Supporting Dataset |
|---|---|
| Accuracy | DAQUAR, COCO-QA, VQA-abstract, Visual Genome and Madlibs, FVQA, Visual7W, Shapes, CLEVR, Diagrams, DVQA, VizWiz, TextVQA, ST-VQA, OCR-VQA, GQA |
| WUPS | COCO-QA, DAQUAR |
| Consensus | VQA, DAQUAR, VizWiz |
| Human Judgment | FM-IQA, KB-VQA, FigureQA |
| MPT | TDIUC |
| BLEU | VizWiz |
| METEOR | VizWiz |
| Average Normalized Levenshtein Similarity (ANLS) | ST-VQA |
| Validity | GQA |
| Plausibility | GQA |
| Distribution | GQA |
| Consistency | GQA |
| Grounding | GQA |
| F1-Score | Icon question answering (ICQA) |

**Table 16**
Comparison of different SOTA methods on VQA 1.0 dataset.

| Model | Year | Open-ended test-std | Open-ended test-dev | Multiple test-std | Multiple test-dev |
|---|---|---|---|---|---|
| Goodfellow et al. [150] | 2014 | 65.90 | – | – | 69.80 |
| Antol et al. [10] | 2015 | 58.20 | 57.75 | 63.10 | 62.70 |
| Chen et al. [30] | 2015 | 48.38 | – | – | – |
| Zhou et al. [40] | 2015 | 55.90 | 55.70 | 62.00 | – |
| Xu et al. [61] | 2015 | 28.24 | 57.99 | – | – |
| Andreas et al. [111] | 2015 | 55.10 | 54.80 | – | – |
| Shih et al. [29] | 2016 | 62.43 | 62.44 | – | – |
| Yang et al. [34] | 2016 | 58.90 | 58.70 | – | – |
| Li et al. [35] | 2016 | 60.76 | 60.72 | 65.43 | 65.43 |
| Fukui et al. [45] | 2016 | 66.50 | 66.70 | 70.10 | 70.20 |
| Kim et al. [47] | 2016 | 66.89 | – | – | – |
| Noh et al. [54] | 2016 | 57.36 | 57.22 | 62.69 | 62.48 |
| Kim et al. [56] | 2016 | 61.84 | 61.68 | 66.33 | – |
| Ilievski et al. [62] | 2016 | 59.54 | 59.24 | 64.18 | 64.01 |
| Lu et al. [63] | 2016 | 62.10 | 61.80 | 66.10 | 65.80 |
| Andreas et al. [73] | 2016 | 59.40 | 59.40 | – | – |
| Noh et al. [74] | 2016 | 63.20 | 63.30 | 67.30 | 67.70 |
| Xiong et al. [76] | 2016 | 60.40 | 60.30 | – | – |
| Wu et al. [152] | 2016 | 59.50 | 59.20 | – | – |
| Jiang et al. [153] | 2016 | – | 52.60 | – | – |
| Wu et al. [154] | 2016 | 55.80 | 55.60 | – | – |
| Jabri et al. [155] | 2016 | – | – | 65.20 | |
| Yu et al. [25] | 2017 | 66.60 | 66.90 | 71.40 | 71.30 |
| Yu et al. [27] | 2017 | 65.30 | 65.20 | 70.00 | 70.00 |
| Schwartz et al. [31] | 2017 | – | – | 69.30 | 69.40 |
| Nam et al. [38] | 2017 | 64.20 | 64.30 | – | – |
| Saito et al. [43] | 2017 | 61.72 | 61.47 | 66.72 | 66.66 |
| Ben-Younes et al. [148] | 2017 | 67.36 | 67.42 | – | – |
| Zhu et al. [151] | 2017 | 68.14 | 67.19 | 72.08 | – |
| Kazemi et al. [163] | 2017 | 64.60 | 64.50 | – | – |
| Song et al. [28] | 2018 | 66.20 | 65.92 | 70.41 | 70.30 |
| Ma et al. [33] | 2018 | 64.10 | 63.80 | 69.40 | 69.50 |
| Yu et al. [49] | 2018 | 67.50 | 67.70 | 72.10 | 72.30 |
| Lao et al. [57] | 2018 | – | 66.40 | – | – |
| Gao et al. [58] | 2018 | 65.90 | 65.89 | – | – |
| Nguyen et al. [65] | 2018 | 67.02 | 66.89 | – | – |
| Su et al. [80] | 2018 | 66.10 | 66.00 | 69.10 | 69.10 |
| Wu et al. [145] | 2018 | 67.97 | 67.83 | 72.32 | 72.28 |
| Lu et al. [147] | 2018 | 66.09 | 66.01 | 69.97 | 70.04 |
| Chandu et al. [149] | 2018 | – | 57.10 | – | – |
| Osman et al. [32] | 2019 | 67.16 | 66.86 | – | – |
| Wu et al. [146] | 2019 | 68.48 | 68.62 | 73.05 | 73.31 |
| Li et al. [196] | 2019 | 60.60 | – | – | – |
| Zhang et al. [88] | 2020 | 67.33 | 67.37 | – | – |
| Liu et al. [95] | 2020 | 69.05 | 69.35 | 73.24 | 73.42 |
| Liu et al. [96] | 2020 | 69.32 | 69.52 | 73.67 | 73.61 |
| Gao et al. [97] | 2020 | 63.12 | 63.13 | – | – |

### 8.2. Comparison of the SOTA methods on VQA 2.0 dataset

VQA 2.0 dataset is a modified version of VQA 1.0 dataset. It consists of more samples when compared with VQA 1.0, containing 443,757 training questions, 214,354 validation questions, 447,793 test questions. Also, VQA 2.0 is more balanced dataset in terms of language bias. Table 17 shows the performance of SOTA VQA models on VQA 2.0 dataset. Accuracy metric provided by [10] is used to evaluate the accuracy of the generated answer given by Eq. 16. MDFNet model (Zhang et al. 2020) outperforms the SOTA models on both test-std (71.32%) and test-dev sets (71.19%). Test-dev set is utilized for developing the model and test-std is employed as a measure to evaluate the final performance. Another worth noting model is SelRes + SelMask + bbox (Hong et al. 2020) which attains an accuracy of 71.30% and 71.00% on test-std and test-dev sets respectively.

### 8.3. Comparison of the SOTA methods on COCO-QA and DAQAUAR datasets

Tables 18 and 19 show the performance of SOTA VQA models on COCO-QA and DAQUAR dataset. Accuracy metric is used to evaluate the performance of models o both the datasets. Also, WU–Palmer Similarity (WUPS) metric is used to analyze the performance of models. WUPS computes the resemblance score between predicted answer and ground truth answer depending on their common subsequence in semantic tree. Further, 0.0 and 0.9 are set as the threshold to construct metrics WUPS@0.0 and WUPS@0.9 for performance evaluation, respectively. It can be observed that ALSA model (Liu et al. 2020) attained the highest accuracy of 69.97%. on COCO-QA dataset The second best performance is attained by the IASSM model (Liu et al. 2020) with an accuracy of 69.92%. Also, both models (ALSA and IASSM) attain best results on the metrics of WUPS@0.0 and WUPS@0.9. On DAQUAR-Reduced dataset, Attributes-CNN + LSTM model (Wu et al. 2016) attains highest accuracy of 46.13% and WUPS@0.9 score of 51.83%. For WUPS@0.0, Deep Walk based VQA model (Li et al. 2019) attains best performance of 88.82%.

**Table 17**
Comparison of different SOTA methods on VQA 2.0 dataset.

| Model | Year | Accuracy on test-std | Accuracy on test-dev |
|---|---|---|---|
| Ilievski et al. [70] | 2017 | – | – |
| Goyal et al. [115] (Prior) | 2017 | 25.98 | – |
| Goyal et al. [115] (Language only) | 2017 | 44.26 | – |
| Goyal et al. [115] (LSTM + CNN) | 2017 | 54.22 | – |
| Ilievski et al. [167] | 2017 | 64.41 | – |
| Anderson et al. [26] | 2018 | 70.34 | – |
| Ma et al. [33] | 2018 | 62.10 | – |
| Yu et al. [49] | 2018 | 68.02 | – |
| Nguyen et al. [65] | 2018 | 66.97 | 66.87 |
| Teney et al. [66] | 2018 | 65.67 | 65.32 |
| Kim et al. [71] | 2018 | 70.35 | 70.04 |
| Su et al. [80] | 2018 | – | 66.67 |
| Singh et al. [156] | 2018 | – | – |
| Bai et al. [161] | 2018 | 67.94 | 67.56 |
| Li et al. [162] | 2018 | 63.94 | – |
| Norcliffe-Brown et al. [168] | 2018 | 65.77 | – |
| Yang et al. [169] | 2018 | 66.17 | 65.94 |
| Zhang et al. [170] | 2018 | 68.41 | 68.09 |
| Lin et al. [175] | 2018 | 68.02 | 67.60 |
| Osman et al. [32] | 2019 | 66.85 | 66.45 |
| Shah et al. [69] | 2019 | – | 69.87 |
| Cadene et al. [78] | 2019 | 68.40 | 68.00 |
| Yu et al. [157] | 2019 | 70.90 | 70.63 |
| Shrestha et al. [158] | 2019 | – | 65.96 |
| Fang et al. [159] | 2019 | – | 66.92 |
| Peng et al. [160] | 2019 | 66.50 | – |
| Gao et al. [164] | 2019 | 70.30 | 70.20 |
| Li et al. [165] | 2019 | 70.58 | 70.27 |
| Lu et al. [166] | 2019 | 70.34 | 70.55 |
| Zhang et al. [88] | 2020 | 67.34 | 67.20 |
| Sun et al. [89] | 2020 | 68.02 | 67.60 |
| Zhu et al. [90] | 2020 | 66.87 | 66.67 |
| Zhang et al. [94] | 2020 | 71.32 | 71.19 |
| Liu et al. [96] | 2020 | – | 69.21 |
| Gao et al. [97] | 2020 | 59.50 | – |
| Hong et al. [98] | 2020 | 71.30 | 71.00 |
| Vu et al. [99] | 2020 | – | 51.35 |

**Table 18**
Comparison of different SOTA methods on COCO-QA.

| Model | Year | Accuracy | WUPS 0.9 | WUPS 0.0 |
|---|---|---|---|---|
| Chen et al. [30] | 2015 | 58.10 | 68.44 | 89.85 |
| Ren et al. [50] | 2015 | 55.09 | 67.90 | 89.52 |
| Ren et al. [51] (Guess) | 2015 | 6.65 | 17.42 | 73.44 |
| Ren et al. [51] (VIS + LSTM)) | 2015 | 53.31 | 63.91 | 88.25 |
| Ren et al. [51] (2-VIS + BiLSTM) | 2015 | 55.09 | 65.34 | 88,64 |
| Ren et al. [51] (VIS + BOW) | 2015 | 55.92 | 66.78 | 88.99 |
| Yang et al. [34] | 2016 | 61.60 | 71.60 | 90.90 |
| Li et al. [35] | 2016 | 62.50 | 72.58 | 91.62 |
| Kafle et al. [41] | 2016 | 63.18 | 73.14 | 91.32 |
| Lu et al. [63] | 2016 | 65.40 | 75.10 | 92.00 |
| Wu et al. [84] | 2016 | 69.73 | 77.14 | 92.50 |
| Wu et al. [152] | 2016 | 70.98 | 78.35 | 92.87 |
| Wu et al. [154] | 2016 | 61.38 | 71.15 | 91.58 |
| Ma et al. [53] | 2016 | 58.40 | 68.50 | 89.67 |
| Teney et al. [174] | 2016 | 60.32 | 69.68 | 90.13 |
| Noh et al. [54] | 2016 | 61.19 | 70.84 | 90.61 |
| Gu et al. [173] | 2017 | 64.70 | 74.10 | 91.70 |
| Song et al. [28] | 2018 | 67.51 | 76.70 | 92.41 |
| Wu et al. [145] | 2018 | 69.33 | 78.29 | 93.02 |
| Lu et al. [147] | 2018 | 66.49 | 76.15 | 92.29 |
| Wu et al. [146] | 2019 | 69.36 | 78.35 | 93.19 |
| Li et al. [196] | 2019 | 60.51 | 70.14 | 91.61 |
| Sun et al. [89] (concatenation attention) | 2020 | 66.90 | 76.26 | 92.19 |
| Sun et al. [89] (projection attention) | 2020 | 67.60 | 76.67 | 92.35 |
| Xi et al. [91] | 2020 | 69.67 | 77.08 | 92.14 |
| Liu et al. [95] | 2020 | 69.92 | 79.04 | 93.74 |
| Liu et al. [96] | 2020 | 69.97 | 79.43 | 94.15 |
| Gao et al. [97] | 2020 | 66.84 | 75.88 | 92.06 |

### 8.4. Comparison of the SOTA methods on Visual7W and CLEVR datasets

Table 20 and Table 21 shows the performance of SOTA models on Visual7W and CLEVR datasets respectively. The boldface numbers show the highest performance on the respective dataset. On Visual7W dataset, BAN model (Kim et al. 2018) attains the highest accuracy. On CLEVR dataset, MAC model (Hudson et al. 2018) attains best accuracy. Accuracy metric is used to evaluate the performance of models on both the datasets.

**Table 19**
Comparison of different SOTA methods on DAQUAR.

| Model | Year | Accuracy | WUPS 0.9 | WUPS 0.0 |
|---|---|---|---|---|
| Chen et al. [30] | 2015 | 42.76 | 47.62 | 83.04 |
| Malinowski et al. [42] | 2015 | 34.68 | 40.76 | 79.54 |
| Ren et al. [51] (Guess) | 2015 | 18.24 | 29.65 | 77.59 |
| Ren et al. [51] (VIS + LSTM)) | 2015 | 34.41 | 46.05 | 82.23 |
| Ren et al. [51] (2-VIS + BiLSTM) | 2015 | 35.78 | 46.83 | 82.15 |
| Ren et al. [51] (VIS + BOW) | 2015 | 34.17 | 44.99 | 81.48 |
| Yang et al. [34] | 2016 | 45.50 | 50.20 | 83.60 |
| Kafle et al. [41] | 2016 | 45.17 | 49.74 | 85.13 |
| Noh et al. [54] | 2016 | 44.48 | 49.56 | 83.95 |
| Xu et al. [61] | 2016 | 40.07 | – | – |
| Wu et al. [84] | 2016 | 45.79 | 51.53 | 83.91 |
| Wu et al. [152] | 2016 | 46.13 | 51.83 | 83.95 |
| Wu et al. [154] | 2016 | 40.07 | 45.43 | 82.67 |
| Ma et al. [53] | 2016 | 39.66 | 44.86 | 83.06 |
| Li et al. [196] | 2019 | 40.56 | 50.01 | 88.82 |
| Xi et al. [91] | 2020 | 45.77 | 51.56 | 83.90 |
| Hosseinabad et al. [92] | 2020 | 40.19 | – | – |

**Table 20**
Comparison of different SOTA methods on Visual7W.

| Model | Year | Accuracy on val |
|---|---|---|
| Zhu et al. [12] | 2016 | 55.60 |
| Fukui et al. [45] | 2016 | 62.20 |
| Yu et al. [27] | 2017 | 62.40 |
| Li et al. [186] | 2017 | 66.00 |
| Song et al. [28] | 2018 | 63.80 |
| Ma et al. [33] | 2018 | 62.80 |
| Kim et al. [71] | 2018 | **71.10** |
| Liu et al. [194] | 2020 | 67.60 |

Bold represents highest accuracy

**Table 21**
Comparison of different SOTA methods on CLEVR.

| Model | Year | Accuracy |
|---|---|---|
| Santoro et al. [37] | 2017 | 95.50 |
| Zhu et al. [151] | 2017 | 78.04 |
| Johnson et al. [176] (Q-type baseline) | 2017 | 41.80 |
| Johnson et al. [176] (CNN + LSTM) | 2017 | 52.30 |
| Johnson et al. [176] (PG + EE) | 2017 | 96.90 |
| Johnson et al. [177] (CNN + LSTM + SA + MLP) | 2017 | 73.20 |
| Hu et al. [178] | 2017 | 83.70 |
| Parez et al. [180] | 2017 | 97.60 |
| Gao et al. [58] | 2018 | 86.30 |
| Desta et al. [172] | 2018 | 94.50 |
| Hudson et al. [179] | 2018 | **98.90** |
| Zhang et al. [88] | 2020 | 96.10 |
| Hong et al. [98] (without expert layout) | 2020 | 94.50 |
| Hong et al. [98] (with expert layout) | 2020 | 96.60 |

Bold represents highest accuracy

### 8.5. Comparison of the SOTA methods on FVQA, Visual7W + KB and OK-VQA datasets

Tables 22–24 shows the performance of SOTA models on FVQA, Visual7W + KB and OK-VQA datasets respectively. For each model, the top-1 and top-3 accuracy is computed. Also, the overall accuracy is the average of accuracy 5 test splits. In FVQA, the top visual concepts extracted from all images are used to construct the knowledge base. These visual concepts are further queried to three knowledge bases, including DBpedia [81], ConceptNet [116] and WebChild [117]. The GRUC model (Yu et al. 2020) achieves highest top-1 and top-3 accuracy among the SOTA models. In Visual7W + KB, knowledge-based questions based on the test images in Visual7W are collected by Li et al. [186]. Visual7W + KB contain 16,850 open- domain QA pairs depending on 8425 images in Visual7W test split. Also, Visual7W + KB use ConceptNet for guiding question generation. The GRUC model (Yu et al. 2020) achieves highest top-1 and top-3 accuracy among the SOTA models on Visua7W + KB dataset. OK-VQA includes questions from different domains like history, science, arts and sports. Again the GRUC model (Yu et al. 2020) attains highest top- and top-3 accuracy on OK-VQA dataset.

**Table 22**
Comparison of different SOTA methods on FVQA.

| Model | Year | Overall Accuracy | |
|---|---|---|---|
| | | top-1 | top-3 |
| Wang et al. [83] (Human) | 2017 | 77.99 | |
| Wang et al. [83] (LSTM-Q + Image + Pre-VQA) | 2017 | 24.98 | 40.40 |
| Wang et al. [83] (Hie-Q + Image + Pre-VQA) | 2017 | 43.14 | 59.44 |
| Wang et al. [83] (top-3-QQmaping) | 2017 | 56.91 | 64.65 |
| Wang et al. [83] (Ensemble) | 2017 | 58.76 | |
| Narasimhan et al. [183] | 2018 | 69.35 | 80.25 |
| Narasimhan et al. [184] | 2018 | 62.20 | 75.60 |
| Li et al. [185] | 2019 | 62.96 | 70.08 |
| Yu et al. [87] | 2020 | 79.63 | 91.20 |

**Table 23**
Comparison of different SOTA methods Visual7W + KB.

| Model | Year | Overall Accuracy | |
|---|---|---|---|
| | | top-1 | top-3 |
| Li et al. [186] (w/o external knowledge) | 2017 | 45.10 | |
| Li et al. [186] (attention-based knowledge incorporation) | 2017 | 51.90 | |
| Li et al. [186] (dynamic memory network based knowledge incorporation) | 2017 | 57.90 | |
| Li et al. [186] (Ensemble) | 2017 | 60.90 | |
| Yu et al. [87] | 2020 | 69.03 | 88.12 |
| Narasimhan et al. [183] | 2018 | 57.32 | 71.61 |

**Table 24**
Comparison of different SOTA methods OK-VQA.

| Model | Year | Overall Accuracy | |
|---|---|---|---|
| | | top-1 | top-3 |
| MUTAN [148] | 2017 | 26.41 | – |
| BAN [71] | 2018 | 25.17 | – |
| Q-only | 2019 | 14.93 | – |
| MLP | 2019 | 20.67 | – |
| ArticleNet [187] | 2019 | 5.28 | – |
| BAN + ArticleNet [187] | 2019 | 25.61 | – |
| MUTAN + ArticleNet [187] | 2019 | 27.84 | – |
| BAN/ArticleNet oracle [187] | 2019 | 27.59 | – |
| MUTAN/ArticleNet oracle [187] | 2019 | 28.47 | – |
| Yu et al. [87] | 2020 | 29.87 | 32.65 |

### 8.6. Comparison of the SOTA methods on GQA dataset

Table 25 shows the performance of SOTA models on GQA dataset. Apart from the standard accuracy metric, GQA dataset supports several new metrics such as consistency, plausibility, validity, and distribution. For consistency, validity, plausibility, higher score is considered better and for distribution metric, lower score indicates better performance. The MDFNet model (Zhang et al. 2020) achieves best performance among SOTA VQA models. The boldface numbers show the highest performance for a given evaluation metric.

### 8.7. Comparison of the scene-text reading VQA models on TextVQA, ST-VQA and OCR-VQA datasets

Table 26–28 shows the performance of recent SOTA VQA models which are capable to answer questions related to the text present in images on TextVQA, ST-VQA and OCR-VQA datasets. LaAP-Net (Han et al. 2020) achieves highest accuracy among the SOTA VQA models on TextVQA dataset. M4C model (Hu et al. 2020) and LoRRA (Singh et al. 2019) are other notable models on TextVQA dataset. On ST-VQA dataset, again LaAP-Net (Han et al. 2020) attains highest accuracy and ANLS score. On OCR-VQA, CRN model (Liu et al. 2020) achieves the highest accuracy among the recent SOTA models.

## 9. Discussions and future directions

### 9.1. Visual feature representation

The performance of the VQA models can be improved by fine-grained and semantic feature extraction methods. The models that use local visual features attain better results as compared to those using global features only as they are not capable to provide fine-grained inference. Also, Reference [66] recommended extracting region-based features for visual objects instead of grid-based features from a CNN.

**Table 25**
Comparison of different SOTA methods on GQA.

| Model | Year | Accuracy | Open | Binary | Validity | Plausibility | Consistency |
|---|---|---|---|---|---|---|---|
| Kim et al. [71] | 2018 | 56.19 | 41.13 | 73.31 | 96.77 | 85.58 | 84.64 |
| Hudson et al. [86] (Human) | 2019 | **89.30** | **87.40** | **91.20** | **98.90** | **97.20** | **98.40** |
| Hudson et al. [86] (CNN + LSTM) | 2019 | 46.55 | 31.80 | 63.26 | 96.02 | 84.25 | 74.57 |
| Hudson et al. [86] (Bottom-Up) | 2019 | 49.74 | 34.83 | 66.64 | 96.18 | 84.57 | 78.71 |
| Hudson et al. [86] (MAC) | 2019 | 54.06 | 38.91 | 71.23 | 96.16 | 84.48 | 81.59 |
| Hu et al. [181] | 2019 | 56.10 | – | – | – | – | – |
| Yang et al. [182] | 2019 | 54.56 | 40.63 | 70.33 | 95.90 | 84.23 | 83.49 |
| Zhang et al. [94] | 2020 | **57.05** | 41.86 | 73.98 | 97.62 | 84.87 | 86.98 |

Bold represents highest accuracy and other evaluation metrics.

**Table 26**
Comparison of different SOTA methods on TextVQA.

| Model | Year | Accuracy on val | Accuracy on test |
|---|---|---|---|
| Kim et al. [71] | 2018 | 12.30 | – |
| Singh et al. [190] | 2018 | 13.04 | 14.00 |
| Singh et al. [105] | 2019 | 26.56 | 27.63 |
| Lin et al. [188] | 2019 | 31.48 | 31.44 |
| MSFT_VT1 [189] | 2019 | 32.92 | 32.46 |
| Hu et al. [108] | 2020 | 39.40 | 39.01 |
| Gao et al. [192] | 2020 | 39.58 | 40.29 |
| Han et al. [193] | 2020 | **40.48** | **40.54** |

Bold represents highest accuracy and other evaluation metrics.

**Table 27**
Comparison of different SOTA methods on ST-VQA.

| Model | Year | Accuracy on val | ANLS on val | ANLS on test |
|---|---|---|---|---|
| SAN [34] + STR [191] | 2016 | – | – | 0.14 |
| SAAA [163] + STR [191] | 2017 | – | – | 0.09 |
| VTA [106] | 2019 | – | – | 0.28 |
| M4C [108] | 2020 | 38.05 | 0.47 | 0.46 |
| SMA [192] | 2020 | – | – | 0.47 |
| LaAP-Net [193] | 2020 | **39.74** | **0.50** | **0.49** |
| CRN [194] | 2020 | – | – | 0.48 |

Bold represents highest accuracy and other evaluation metrics.

**Table 28**
Comparison of different SOTA methods on OCR-VQA.

| Model | Year | Accuracy on val | Accuracy on test |
|---|---|---|---|
| BLOCK [107] | 2019 | – | 42.00 |
| CNN [107] | 2019 | – | 14.30 |
| BLOCK + CNN [107] | 2019 | – | 41.50 |
| BLOCK + CNN + W2V [107] | 2019 | – | 48.30 |
| M4C [108] | 2020 | 63.50 | 63.90 |
| LaAP-Net [193] | 2020 | 63.80 | 64.10 |
| CRN [194] | 2020 | **64.09** | **64.48** |

Bold represents highest accuracy and other evaluation metrics.

In future, the researchers must explore the successors of R-CNN, fast R-CNN [126], faster R-CNN [5,50,51] and the latest Mask R-CNN [127] for image featurization.

### 9.2. Textual feature representation

In order to generate word embeddings for out-of-vocabulary (OOV), FastText [129] can be used which includes character n-grams. It can be useful for goal-oriented VQA datasets like Viz-Wiz [122] and VQA-Med [128] which will be used more in next few years. ELMo (Embeddings from Language Models) [130] and BERT (Bidirectional Encoder Representations from Transformers) [131] are two effective algorithms for deep learning based VQA models. Both are contextualized word embedding algorithms and also able to resolve OOV problem. They are able to produce distinct word embeddings for a word by considering the context.

Another promising direction can be the use of sentence embeddings in place of word embeddings. These sentence embeddings must encapsulate the rich semantics and sensitive to ordering of words, context and tense. In future, sentence embedding models like Quick-thought vectors [132], InferSent [133] and Google's Universal Sentence Encoder [134] can be used for question representation. For short length questions, we can use character level embeddings in place of contextual embeddings. For short questions, little contextual information is available which is not sufficient for these contextual embeddings to perform well. VLGrammar [217] can be used for jointly learning vision and language in common space, which employs complex probabilistic context-free grammars to provoke the sentence grammar and the image grammar concurrently.

Another suggestion can be the use of Capsule Networks (CapsNet) as they are found to be effective for textual question answering. To reduce the number of parameters utilized in word embeddings, [135] used CapsNet effectively. They can be directly applied in VQA task by providing input as triples (image, question, answer).

### 9.3. Datasets

An important limitation that VQA models face is due to the unavailability of goal-oriented datasets. Still the benchmark datasets like DAQUAR and VQA are used by researchers to perform the experiments. The models fail to handle real-world applications like helping blind people, supporting a data analyst to provide a useful content from huge amount of data, guiding children learning a concept on smart devices and communicating with a robot. VizWiz and VQA-Med are only two goal-oriented datasets which publically available till date. Only few papers are available in literature which has used VizWiz dataset for VQA task. Among them, [122] attained the best accuracy of 50% only. Similarly, few papers have addresses VQA-Med dataset, with a best BLEU score of 0.162 according to [128]. So, there is a need of building publically available goal-oriented datasets to enhance the VQA performance.

### 9.4. Evaluation metrics

We have seen that a large set of VQA models in literature were evaluated on accuracy metric which is good for multiple-choice type of questions. In future research, new evaluation metrics must be explored to measure the open-ended questions. We can move in this direction by using automatic Machine Translation (MT) evaluation metrics with VQA model metric, as concluded from two latest VQA challenges, VizWiz and VQA-Med. They have used BLEU [139] and METEOR [140] metrics for classical machine translation task. It can be noticed that BLEU metric generally fails with short sentences. On the other hand, generally short answers exist in VQA systems. In future, we can use NEVA (Ngram EVAluation) [141] as it is suitable for short sentences.

### 9.5. Others

Another future direction can be the use of adjective phrases to the generated answer as discussed by [136]. It will certainly help the visually impaired users by eliminating their accessibility barriers. [137] discusses fuzzy systems with sentimental analysis with motivating results. So the combination of fuzzy mood detector to VQA network will certainly improve the accuracy of affective question answering systems.

Another important future direction can be the use of external knowledge-bases for the category of questions that require world-knowledge or common-sense to predict answer as discussed by Wang et al. [82], Wu et al. [84] and Wang et al. [83] with certain limitations. So, there is a need to build new VQA models that can answer any form of question and have good reasoning capabilities with improved performance.

Attention methods have been used efficiently in capturing the relationship between visual and textual features. Also, it is observed that models that do not use attention mechanism show lesser performance as compared to those models that use attention mechanisms. In addition, VQA models using multi-hop attention attains higher performance compared with single-hop attention, but need additional tuning efforts. We have observed that the VQA models attain best accuracy when the number of hops is set to a small value such as 2.

VQA models that use bilinear fusion techniques are expected to attain better results as they are capable to capture the interaction between image features and question features, in depth. In future, we will try to integrate bilinear fusion techniques with an improved attention mechanism.

One of the limitations of encoder–decoder architecture is their time consuming sequential encoding steps. The use of Transformers proposed by Vaswani et al. [138] can handle this limitation. It contains layers of encoder and decoder which provide parallel attention. We can use Transformers in real applications which require VQA-Med [128] and Viz-Wiz [122] datasets.

## 10. Conclusion

In this survey, we present a critically discussed the recent SOTA models, datasets and evaluation metrics. Apart from standard datasets discussed in previous surveys, we have also discussed some new datasets developed in 2019 and 2020 such as GQA, OK-VQA, TextVQA, ST-VQA, and OCR-VQA. Similarly, the evaluation metrics such as BLEU, MPT, METEOR, Average Normalized Levenshtein Similarity (ANLS), Validity, Plausibility, Distribution, Consistency, Grounding, F1-Score which are supported by these new datasets are also discussed. We have also discussed the VQA models that are capable to answer question related to text present in images. We have also discussed the attention mechanism to get fine-grained image and questions features together with different fusion techniques in detail. In the end, we have discussed the several useful directions such as creation of unbiased and goal-oriented datasets. Further, VQA algorithms must be improved in such a way that they can use the power of CNNs to extract more natural visual features by employing recent word embedding algorithms to perform improved multi-modal fusion. We hope that current and future research directions will surely benefit the VQA task as well as the common aim of understating of visual scenes.

## Declaration of Competing Interest

None.

## References


[1] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, A. van den Hengel, Visual question answering: a survey of methods and datasets, Comput. Vis. Image Underst. 163 (2017) 21–40.

[2] K. Simonyan, A. Zisserman, Deep convolutional networks for large-scale image recognition, Proceedings of the International Conference on Learning Representations, 2015.

[3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CVPR 2016, pp. 770–778.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You only look once: Unified, real-time object detection, Proceedings of the IEEE conference on computer vision and pattern recognition 2016, pp. 779–788.

[5] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, NIPS 2015, pp. 91–99.

[6] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 2625–2634.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, , 1725-1732.

[8] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Proces. Syst. 27 (2014) 568–576.

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

[10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, Dhruv Batra, C.L. Zitnick, D. Parikh, Vqa: Visual question answering, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 2425–2433.

[11] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, D. Parikh, Yin and yang: Balancing and answering binary visual questions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 5014–5022.

[12] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7w: Grounded question answering in images, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 4995–5004.

[13] D. Zhang, R. Cao, S. Wu, Information fusion in visual question answering: a survey, Information Fusion 52 (2019) 268–280.

[14] M. Sruthy, B.C. Kovoor, Visual question answering: a state-of-the-art review.“ artificial intelligence review (2020): 1-41.Manmadhan, Sruthy, and Binsu C. Kovoor. "visual question answering: a state-of-the-art review, Artif. Intell. Rev. (2020) 1–41.

[15] C. Patil, M. Patwardhan, Visual question generation: the state of the art, ACM Computing Surveys (CSUR) 53 (3) (2020) 1–22.

[16] Y. LeCun, B.E. Boser, J.S. Denker, D. Henderson, R.E. Howard, W.E. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neu- ral Comput. (1989) 541–551.

[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM (2017) 84–90.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, CVPR 2015, pp. 1–9.

[19] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput (1997) 1735–1780.

[20] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, EMNLP 2014, pp. 1724–1734.

[21] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint (2013).

[22] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Proces. Syst. 26 (2013) 3111–3119.

[23] Jeffrey Pennington, Richard Socher, Christopher D. Manning, Glove: Global vectors for word representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014, pp. 1532–1543.

[24] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, IEEE Computational Intelligence Magazine 13 (3) (2018) 55–75.

[25] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, ICCV 2017, pp. 1839–1848.

[26] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bot- tom-up and top-down attention for image captioning and VQA, CVPR (2018) 4995–5004.

[27] D. Yu, J. Fu, T. Mei, Y. Rui, Multi-level attention networks for visual question answering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 4709–4717.

[28] J. Song, P. Zeng, L. Gao, H.T. Shen, From pixels to objects: cubic visual attention for visual question answering, IJCAI 2018, pp. 906–912.

[29] K.J. Shih, S. Singh, D. Hoiem, Where to look: Focus regions for visual question answering, CVPR 2016, pp. 4613–4621.

[30] K. Chen, J. Wang, L. Chen, H. Gao, W. Xu, R. Nevatia, ABC-CNN: an attention based convolutional neural network for visual question answering, CoRR (2015).

[31] I. Schwartz, A.G. Schwing, T. Hazan, High-order attention models for visual question answering, NIPS 2017, pp. 3667–3677.

[32] A. Osman, W. Samek, Dual recurrent attention units for visual question answering, CoRR (2018).

[33] C. Ma, C. Shen, A.R. Dick, Q. Wu, P. Wang, A. van den Hengel, I.D. Reid, Visual question answering with memory-augmented networks, CVPR, IEEE Computer Society 2018, pp. 6975–6984.

[34] Z. Yang, X. He, J. Gao, L. Deng, A.J. Smola, Stacked attention networks for image question answering, CVPR 2016, pp. 21–29.

[35] R. Li, J. Jia, Visual question answering with question representation update (QRU), NIPS 2016, pp. 4655–4663.

[36] M.T. Desta, L. Chen, T. Kornuta, Object-based reasoning in VQA, 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, pp. 1814–1823.
[37] A. Santoro, D. Raposo, D.G.T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, NIPS 2017, pp. 4974–4983.
[38] H. Nam, J.-W. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, CVPR 2017, pp. 2156–2164.
[39] R. Kiros, Y. Zhu, R. Salakhutdinov, R.S. Zemel, R. Urtasun, A. Torralba, S. Fidler, Skip-thought vectors, NIPS 2015, pp. 3294–3302.
[40] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, R. Fergus, Simple baseline for visual question answering, arXiv preprint (2015).
[41] K. Kafle, C. Kanan, Answer-type prediction for visual question answering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 4976–4984.
[42] M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons: a neural-based approach to answering questions about images, ICCV 2015, pp. 1–9.
[43] K. Saito, A. Shin, Y. Ushiku, T. Harada, Dualnet: Domain-invariant network for visual question answering, 2017 IEEE International Conference on Multimedia and Expo (ICME), IEEE 2017, pp. 829–834.
[44] V. Lioutas, N. Passalis, A. Tefas, Explicit ensemble attention learning for improving visual question answering, Pattern Recogn. Lett. 111 (2018) 51–57.
[45] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, EMNLP 2016, pp. 457–468.
[46] M. Charikar, K. Chen, M. Farach-Colton, Finding frequent items in data streams, International Colloquium on Automata, Languages, and Programming, Springer, Berlin, Heidelberg 2002, pp. 693–703.
[47] J.-H. Kim, K.W. On, W. Lim, J. Kim, J. Ha, B.-T. Zhang, Hadamard product for low-rank bilinear pooling, CoRR (2016).
[48] H. Pirsiavash, D. Ramanan, C.C. Fowlkes, Bilinear classifiers for visual recognition, NIPS 2009, pp. 1482–1490.
[49] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, Dacheng Tao, Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering, IEEE transactions on neural networks and learning systems 29 (12) (2018) 5947–5959.
[50] M. Ren, R. Kiros, R. Zemel, Exploring models and data for image question answering, Adv. Neural Inf. Proces. Syst. 28 (2015) 2953–2961.
[51] M. Ren, R. Kiros, R. Zemel, Image question answering: a visual semantic embedding model and a new dataset, Proc. Advances in Neural Inf. Process. Syst 1 (2) (2015) 5.
[52] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T.P. Lillicrap, Meta-learning with memory- augmented neural networks, ICML 2016, pp. 1842–1850.
[53] L. Ma, Z. Lu, H. Li, Learning to answer questions from image using convolutional neural network, AAAI 2016, pp. 3567–3573.
[54] H. Noh, P.H. Seo, B. Han, Image question answering using convolutional neural network with dynamic parameter prediction, CVPR 2016, pp. 30–38.
[55] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, W. Xu, Are you talking to a machine? Dataset and methods for multilingual image question, Adv. Neural Inf. Proces. Syst. 28 (2015) 2296–2304.
[56] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J. Ha, B.-T. Zhang, Multimodal residual learning for visual QA, NIPS 2016, pp. 361–369.
[57] M. Lao, Y. Guo, H. Wang, X. Zhang, Cross-modal multistep fusion network with co-attention for visual question answering, IEEE Access 6 (2018) 31516–31524.
[58] P. Gao, H. Li, S. Li, P. Lu, Y. Li, S.C.H. Hoi, X. Wang, Question-guided hybrid convolution for visual question answering, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 469–485.
[59] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Neural module networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 39–48.
[60] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, International Conference on Machine Learning 2015, pp. 2048–2057.
[61] H. Xu, K. Saenko, Ask, attend and answer: exploring question-guided spatial attention for visual question answering, ECCV 2016, pp. 451–466.
[62] I. Ilievski, S. Yan, J. Feng, A focused dynamic attention model for visual question answering, CoRR (2016).
[63] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, NIPS 2016, pp. 289–297.
[64] https://web.stanford.edu/class/cs224n/reports/custom/15709925.pdf.
[65] D.-K. Nguyen, T. Okatani, Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, CVPR 2018, pp. 6087–6096.
[66] D. Teney, P. Anderson, X. He, A. van den Hengel, Tips and tricks for visual question answering: learnings from the 2017 challenge, CVPR, IEEE Computer Society 2018, pp. 4223–4232.
[67] J. Liang, L. Jiang, L. Cao, L. Li, A.G. Hauptmann, Focal visual-text attention for visual question answering, CVPR, IEEE Computer Society 2018, pp. 6135–6143.
[68] L. Jiang, J. Liang, L. Cao, Y. Kalantidis, S. Farfade, A.G. Hauptmann, Memexqa: visual memex question answering, CoRR abs/1708.01336, 2017.
[69] M. Shah, X. Chen, M. Rohrbach, D. Parikh, Cycle-consistency for robust visual question answering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 6649–6658.
[70] I. Ilievski, J. Feng, Generative attention model with adversarial self-learning for visual question answering, Proceedings of the on Thematic Workshops of ACM Multimedia, 2017, 2017, pp. 415–423.
[71] J.-H. Jin, J. Jun, B.-T. Zhang, Bilinear Attention Networks, CoRR, 2018.
[72] M.C. Marneffe, C.D. Manning, The Stanford typed dependencies representation, Coling 2008: Proceedings of the Workshop on Cross-framework and Cross-domain Parser Evaluation 2008, pp. 1–8.
[73] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Learning to compose neural networks for question answering, arXiv preprint (2016).
[74] H. Noh, B. Han, Training recurrent answering units with joint loss minimization for vqa, arXiv preprint (2016).
[75] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, R. Socher, Ask me anything: Dynamic memory networks for natural language processing, International Conference on Machine Learning 2016, pp. 1378–1387.
[76] C. Xiong, S. Merity, R. Socher, Dynamic memory networks for visual and textual question answering, International Conference on Machine Learning 2016, pp. 2397–2406.
[77] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, J. Tenenbaum, Neural-symbolic vqa: disentangling reasoning from vision and language understanding, Adv. Neural Inf. Proces. Syst. 31 (2018) 1031–1042.
[78] R. Cadene, H. Ben-Younes, M. Cord, N. Thome, Murel: multimodal relational reasoning for visual question answering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, , 1989-1998.
[79] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, From recognition to cognition: Visual commonsense reasoning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 6720–6731.
[80] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, J. Li, Learning visual knowledge memory networks for visual question answering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 7736–7745.
[81] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, et al., DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia, Semantic web 6 (2) (2015) 167–195.
[82] P. Wang, Q. Wu, C. Shen, A. van den Hengel, A. Dick, Explicit knowledge-based reasoning for visual question answering, arXiv preprint (2015).
[83] P. Wang, Q. Wu, C. Shen, A. Dick, A. Van Den Hengel, Fvqa: fact-based visual question answering, IEEE Trans. Pattern Anal. Mach. Intell. 40 (10) (2017) 2413–2427.
[84] Q. Wu, P. Wang, C. Shen, A. Dick, A. Van Den Hengel, Ask me anything: Free-form visual question answering based on knowledge from external sources, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 4622–4630.
[85] T. Tommasi, A. Mallya, B. Plummer, S. Lazebnik, A.C. Berg, T.L. Berg, Combining multiple cues for visual madlibs question answering, Int. J. Comput. Vis. 127 (1) (2019) 38–60.
[86] D.A. Hudson, C.D. Manning, Gqa: A new dataset for real-world visual reasoning and compositional question answering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 6700–6709.
[87] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, J. Tan, Cross-modal knowledge reasoning for knowledge-based visual question answering, Pattern Recogn. 108 (2020) 107563.
[88] W. Zhang, J. Yu, H. Hu, H. Hu, Z. Qin, Multimodal feature fusion by relational reasoning and attention for visual question answering, Information Fusion 55 (2020) 116–126.
[89] B. Sun, Z. Yao, Y. Zhang, L. Yu, Local relation network with multilevel attention for visual question answering, J. Vis. Commun. Image Represent. 102762 (2020).
[90] X. Zhu, Z. Mao, Z. Chen, Y. Li, Z. Wang, B. Wang, Object-difference drived graph convolutional networks for visual question answering, Multimed. Tools Appl. (2020) 1–19.
[91] Y. Xi, Y. Zhang, S. Ding, S. Wan, Visual question answering model based on visual relationship detection, Signal Process. Image Commun. 80 (2020) 115648.
[92] S.H. Hosseinabad, M. Safayani, A. Mirzaei, Multiple answers to a question: a new approach for visual question answering, Vis. Comput. (2020) 1–13.
[93] Z. Bai, Y. Li, M. Woźniak, M. Zhou, D. Li, DecomVQANet: decomposing visual question answering deep network via tensor decomposition and regression, Pattern Recogn. 110 (2020) 107538.
[94] W. Zhang, J. Yu, Y. Wang, W. Wang, Multimodal deep fusion for image question answering, Knowl.-Based Syst. 106639 (2020).
[95] Y. Liu, X. Zhang, F. Huang, Z. Zhou, Z. Zhao, Z. Li, Visual question answering via combining inferential attention and semantic space mapping, Knowl.-Based Syst. 207 (2020) 106339.
[96] Y. Liu, X. Zhang, Z. Zhao, B. Zhang, L. Cheng, Z. Li, ALSA: adversarial learning of supervised attentions for visual question answering, IEEE Transactions on Cybernetics (2020).
[97] L. Gao, L. Cao, X. Xu, J. Shao, J. Song, Question-led object attention for visual question answering, Neurocomputing 391 (2020) 227–233.
[98] J. Hong, S. Park, H. Byun, Selective residual learning for visual question answering, Neurocomputing 402 (2020) 366–374.
[99] M.H. Vu, T. Löfstedt, T. Nyholm, R. Sznitman, A question-centric model for visual question answering in medical imaging, IEEE Trans. Med. Imaging 39 (9) (2020) 2856–2868.
[100] Z. Huasong, J. Chen, C. Shen, H. Zhang, J. Huang, X.S. Hua, Self-adaptive neural module transformer for visual question answering, IEEE Transactions on Multimedia 23 (2020) 1264–1273.
[101] S. Lobry, D. Marcos, J. Murray, D. Tuia, RSVQA: visual question answering for remote sensing data, IEEE Trans. Geosci. Remote Sens. 58 (12) (2020) 8555–8566.
[102] N. Gupta, A.S. Jalal, Integration of textual cues for fine-grained image captioning using deep CNN and LSTM, Neural Comput. & Applic. (2019) 1–10.
[103] X. Bai, M. Yang, P. Lyu, Y. Xu, J. Luo, Integrating scene text and visual appearance for fine-grained image classification, IEEE Access 6 (2018) 66322–66335.

[104] A.U. Dey, S.K. Ghosh, E. Valveny, G. Harit, Beyond visual semantics: exploring the role of scene text in image understanding, arXiv preprint 149 (2019) 164–171.
[105] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, M. Rohrbach, Towards vqa models that can read, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 8317–8326.
[106] A.F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C.V. Jawahar, D. Karatzas, Scene text visual question answering, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 4291–4301.
[107] A. Mishra, S. Shekhar, A.K. Singh, A. Chakraborty, Ocr-vqa: Visual question answering by reading text in images, 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE 2019, pp. 947–952.
[108] R. Hu, A. Singh, T. Darrell, M. Rohrbach, Iterative answer prediction with pointer-augmented multimodal transformers for textvqa, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 9992–10002.
[109] M. Malinowski, M. Fritz, A multi-world approach to question answering about real-world scenes based on uncertain input, Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14) 2014, pp. 1682–1690.
[110] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, et al., Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73.
[111] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Deep compositional question answering with neural module networks, CoRR abs/1511.02799, 2015.
[112] L. Yu, E. Park, A.C. Berg, T.L. Berg, Visual madlibs: Fill in the blank image generation and question answering, arXiv preprint (2015).
[113] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, European Conference on Computer Vision, Springer, Cham 2014, pp. 740–755.
[114] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, European Conference on Computer Vision, Springer, Berlin, Heidelberg 2012, pp. 746–760.
[115] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the V in VQA matter: Elevating the role of image understanding in visual question answering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[116] H. Liu, P. Singh, ConceptNet—a practical commonsense reasoning tool-kit, BT Technol. J. 22 (4) (2004) 211–226.
[117] N. Tandon, G. De Melo, F. Suchanek, G. Weikum, Webchild: Harvesting and organizing commonsense knowledge from the web, Proceedings of the 7th ACM International Conference on Web Search and Data Mining 2014, pp. 523–532.
[118] N. Tandon, G. Melo, G. Weikum, Acquiring Comparative Commonsense Knowledge from the Web, AAAI 2014, pp. 166–172.
[119] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L.G. Bigorda, S.R. Mestre, J. Mas, D.F. Mota, J.A. Almazan, L.P. De Las Heras, ICDAR 2013 robust reading competition, 2013 12th International Conference on Document Analysis and Recognition, IEEE 2013, pp. 1484–1493.
[120] D. Karatzas, L.G. Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, et al., ICDAR 2015 competition on robust reading, 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE 2015, pp. 1156–1160.
[121] J. Deng, W. Dong, R. Socher, L.J. Li, Kai Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2009, pp. 248–255.
[122] D. Gurari, Q. Li, A.J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, J.P. Bigham, Vizwiz grand challenge: Answering visual questions from blind people, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 3608–3617.
[123] A. Mishra, K. Alahari, C.V. Jawahar, Image retrieval using textual cues, Proceedings of the IEEE International Conference on Computer Vision 2013, pp. 3040–3047.
[124] A. Veit, T. Matera, L. Neumann, J. Matas, S. Belongie, Coco-text: Dataset and benchmark for text detection and recognition in natural images, arXiv preprint (2016).
[125] Z. Wu, M. Palmer, Verbs semantics and lexical selection, Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics, Las Cruces, New Mexico, 1994.
[126] R. Girshick, Fast r-cnn, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 1440–1448.
[127] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 2961–2969.
[128] S.A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, M.P. Lungren, Overview of ImageCLEF 2018 Medical domain visual question answering task. In CLEF (Working Notes), 2018.
[129] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.
[130] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, arXiv preprint arXiv (2018) 180205365.
[131] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint (2018).
[132] L. Logeswaran, H. Lee, An efficient framework for learning sentence representations, arXiv preprint (2018).
[133] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised learning of universal sentence representations from natural language inference data, arXiv preprint (2017).
[134] D. Cer, Y. Yang, S.V. Kong, N. Hua, N. Limtiaco, R. St John, N. Constant, et al., Universal sentence encoder, arXiv preprint (2018).
[135] H. Ren, H. Lu, Compositional coding capsule network with k-means routing for text classification, arXiv preprint (2018).
[136] N. Ruwa, Q. Mao, L. Wang, M. Dong, Affective visual question answering network, 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE 2018, pp. 170–173.
[137] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recogn. Lett. 125 (2019) 264–270.
[138] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhinz, Attention is all you need, Advances in Neural Information Processing Systems 2017, pp. 5998–6008.
[139] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, BLEU: a method for automatic evaluation of machine translation, Proceedings of the 40th annual meeting of the Association for Computational Linguistics 2002, pp. 311–318.
[140] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, Proceedings of the Ninth Workshop on Statistical Machine Translation 2014, pp. 376–380.
[141] E. Forsbom, Training a super model look-alike: Featuring edit distance, n-gram occurrence, and one reference translation, Proceedings of the Workshop on Machine Translation Evaluation: Towards Systemizing MT Evaluation 2003, pp. 29–36.
[145] C. Wu, J. Liu, X. Wang, X. Dong, Object-difference attention: A simple relational attention for visual question answering, Proceedings of the ACM Multimedia Conference, ACM 2018, pp. 519–527.
[146] C. Wu, J. Liu, X. Wang, R. Li, Differential networks for visual question answering, The Thirty-Third AAAI Conference on Artificial Intelligence 2019, pp. 8997–9004.
[147] P. Lu, H. Li, W. Zhang, J. Wang, X. Wang, Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering, AAAI, 2018.
[148] H. Ben-younes, R. Cadène, M. Cord, N. Thome, MUTAN: multimodal tucker fusion for visual question answering, ICCV 2017, pp. 2631–2639.
[149] K.R. Chandu, M.A. Pyreddy, M. Felix, N.N. Joshi, Textually enriched neural module networks for visual question answering, CoRR abs/1809.08697, 2018.
[150] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.),NIPS 2014, pp. 2672–2680.
[151] C. Zhu, Y. Zhao, S. Huang, K. Tu, Y. Ma, Structured attentions for visual question answering, ICCV 2017, pp. 1300–1309.
[152] Q. Wu, C. Shen, A. van den Hengel, P. Wang, A.R. Dick, Image captioning and visual question answering based on attributes and their related external knowledge, CoRR 40 (6) (2017) 1367–1381.
[153] A. Jiang, F. Wang, F. Porikli, Y. Li, Compositional memory for visual question answering, arXiv preprint (2015).
[154] Q. Wu, C. Shen, L. Liu, A. Dick, A. Van Den Hengel, What value do explicit high level concepts have in vision to language problems? Proceedings of the IEEE conference on computer vision and pattern recognition 2016, pp. 203–212.
[155] A. Jabri, A. Joulin, L. Van Der Maaten, Revisiting visual question answering baselines, European conference on computer vision, Springer, Cham 2016, pp. 727–739.
[156] J. Singh, V. Ying, A. Nutkiewicz, Attention on attention: Architectures for visual question answering (vqa), arXiv preprint (2018).
[157] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, Proceedings of the IEEE conference on computer vision and pattern recognition 2019, pp. 6281–6290.
[158] R. Shrestha, K. Kafle, C. Kanan, Answer them all! toward universal visual question answering models, Proceedings of the IEEE conference on computer vision and pattern recognition 2019, pp. 10472–10481.
[159] F. Zhiwei, J. Liu, Y. Li, Y. Qiao, H. Lu, Improving visual question answering using dropout and enhanced question encoder, Pattern Recogn. 90 (2019) 404–414.
[160] L. Peng, Y. Yang, Y. Bin, N. Xie, F. Shen, Y. Ji, X. Xu, Word-to-region attention network for visual question answering, Multimed. Tools Appl. 78 (3) (2019) 3843–3858.
[161] Y. Bai, J. Fu, T. Zhao, T. Mei, Deep attention neural tensor network for visual question answering, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 20–35.
[162] M. Li, L. Gu, Y. Ji, C. Liu, Text-guided dual-branch attention network for visual question answering, Pacific Rim Conference on Multimedia, Springer, Cham 2018, pp. 750–760.
[163] V. Kazemi, A. Elqursh, Show, ask, attend, and answer: A strong baseline for visual question answering, 2017 (arXiv: 1704.03162v2).
[164] P. Gao, H. Li, H. You, Z. Jiang, P. Lu, S. Hoi, X. Wang, Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering, CVPR, 2019.
[165] L. Li, Z. Gan, Y. Cheng, J. Liu, Relation-aware graph attention network for visual question answering, ICCV 2019, pp. 10312–10321.
[166] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, NIPS 2019, pp. 13–23.
[167] I. Ilievski, J. Feng, Multimodal learning and reasoning for visual question answering, Advances in neural information processing systems 2017, pp. 551–562.
[168] W. Norcliffe-Brown, S. Vafeias, S. Parisot, Learning conditioned graph structures for interpretable visual question answering, Advances in neural information processing systems 2018, pp. 8334–8343.
[169] Z. Yang, J. Yu, C. Yang, Z. Qin, Y. Hu, Multi-modal learning with prior visual relation reasoning, arXiv preprint 3 (7) (2018).
[170] Y. Zhang, J. Hare, A. Prügel-Bennett, Learning to count objects in natural images for visual question answering, arXiv preprint (2018).
[171] B.K. Iwana, S.T.R. Rizvi, S. Ahmed, A. Dengel, S. Uchida, Judging a book by its cover, arXiv preprint (2016).
[172] M.T. Desta, L. Chen, T. Kornuta, Object-based reasoning in VQA, 2018 IEEE winter conference on applications of computer vision, 2018, WACV 2018, pp. 1814–1823.

[173] G. Gu, S.T. Kim, Y.M. Ro, Adaptive attention fusion network for visual question answering, Proceedings of the IEEE International Conference on Multimedia and Expo 2017, pp. 997–1002.
[174] D. Teney, A. van den Hengel, Zero-shot visual question answering, arXiv preprint (2016).
[175] Y. Lin, Z. Pang, D. Wang, Y. Zhuang, Feature Enhancement in Attention for Visual Question Answering, IJCAI 2018, pp. 4216–4222.
[176] H.B. Johnson, L. Maaten, F.-F. Li, Inferring and executing programs for visual reasoning, ICCV 2017, pp. 3008–3017.
[177] H.B. Johnson, L. Maaten, F.-F. Li, Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, CVPR 2017, pp. 1988–1997.
[178] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, K. Saenko, Learning to reason: End-to-end module networks for visual question answering, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 804–813.
[179] D. Husdon, C. Manning, Compositional attention networks for machine reasoning, ICLR, 2018.
[180] E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville, Film: Visual reasoning with a general conditioning layer, arXiv preprint 32 (1) (2018) 3942–3951.
[181] R. Hu, A. Rohrbach, T. Darrell, K. Saenko, Language-conditioned graph networks for relational reasoning, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 10294–10303.
[182] Z. Yang, Z. Qin, J. Yu, Y. Hu, Scene graph reasoning with prior visual relationship for visual question answering, arXiv preprint arX560iv:1812.09681v2, 2019.
[183] M. Narasimhan, S. Lazebnik, A. Schwing, Out of the box: reasoning with graph convolution nets for factual visual question answering, NeurIPS 2018, pp. 2654–2665.
[184] M. Narasimhan, A.G. Schwing, Straight to the facts: learning knowledge base retrieval for factual visual question answering, ECCV 2018, pp. 451–468.
[185] H. Li, P. Wang, C. Shen, A. van den Hengel, Visual question answering as reading comprehension, CVPR 2019, pp. 6319–6328.
[186] G. Li, H. Su, W. Zhu, Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks, arXiv preprint arXiv 1712 (2017) 00733.
[187] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, OK-VQA: a visual question answering benchmark requiring external knowledge, CVPR 2019, pp. 3195–3204.
[188] Y. Lin, H. Zhao, Y. Li, D. Wang, DCD ZJU, TextVQA Challenge 2019 Winner, 2014.
[189] Anonymous submission, MSFT VTI, TextVQA Challenge 2019 top entry (post-challenge), https://evalai.cloudcv.org/web/challenges/challenge-page/244/.
[190] A. Singh, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, D. Parikh, Pythia-a platform for vision & language research, SysML Workshop, NeurIPS, 2018.
[191] L. Gomez, A. Mafla, M. Rusinol, D. Karatzas, Single shot scene text retrieval, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 700–715.
[192] C. Gao, Q. Zhu, P. Wang, H. Li, Y. Liu, A. van den Hengel, Q. Wu, Structured multimodal attentions for textvqa, arXiv preprint (2020).
[193] W. Han, H. Huang, T. Han, Finding the Evidence: Localization-aware Answer Prediction for Text Visual Question Answering, arXiv preprint (2020).
[194] F. Liu, G. Xu, Q. Wu, Q. Du, W. Jia, M. Tan, Cascade Reasoning Network for Text-based Visual Question Answering, Proceedings of the 28th ACM International Conference on Multimedia 2020, pp. 4060–4069.
[195] H. Sharma, A.S. Jalal, Visual question answering model based on graph neural network and contextual attention, Image and Vision Computing 110 (2021) 104165.
[196] Q. Li, F. Xiao, L. An, X. Long, X. Sun (Eds.),Semantic Concept Network and Deep Walk- based Visual Question Answering. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15, 2019, pp. 1–19 , no. 2s.
[197] Y. Shi, T. Furlanello, S. Zha, A. Anandkumar, Question type guided attention in visual question answering, ECCV 2018, pp. 158–175.
[198] A.S. Toor, H. Wechsler, M. Nappi, Question action relevance and editing for visual question answering, Multimed. Tools Appl. 78 (3) (2019) 2921–2935.
[199] D. Teney, A. van den Hengel, Visual question answering as a meta learning task, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 219–235.
[200] D. Yu, X. Gao, H. Xiong, Structured semantic representation for visual question answering, 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE 2018, pp. 2286–2290.
[201] M. Malinowski, C. Doersch, A. Santoro, P. Battaglia, Learning visual question answering by bootstrapping hard attention, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 3–20.
[202] X. Lin, D. Parikh, Leveraging visual question answering for image-caption ranking, European Conference on Computer Vision, Springer, Cham 2016, pp. 261–277.
[203] L. Cao, L. Gao, J. Song, X. Xu, H.T. Shen, Jointly learning attentions with semantic cross-modal correlation for visual question answering, Australasian Database Conference, Springer, Cham 2017, pp. 248–260.
[204] H. Xu, K. Saenko, Dual attention network for visual question answering, ECCV 2016 2nd Workshop on Storytelling with Images and Videos (VisStory), 2016.
[205] N. Ruwa, Q. Mao, L. Wang, J. Gou, M. Dong, Mood-aware visual question answering, Neurocomputing 330 (2019) 305–316.
[206] L. Gao, P. Zeng, J. Song, X. Liu, H.T. Shen, Examine before you Answer: Multi-Task Learning with Adaptive-Attentions for Multiple-Choice VQA, in: MM, 2018 1742–1750.
[207] H.O. Lancaster, E. Seneta, Chi-square distribution, Encyclopedia of biostatistics 2 (2005).
[208] S.E. Kahou, V. Michalski, A. Atkinson, Á. Kádár, A. Trischler, Y. Bengio, Figureqa: An annotated figure dataset for visual reasoning, arXiv preprint (2017).
[209] K. Kafle, B. Price, S. Cohen, C. Kanan, Dvqa: Understanding data visualizations via question answering, Proceedings of the IEEE conference on computer vision and pattern recognition 2018, pp. 5648–5656.
[210] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, A. Farhadi, A diagram is worth a dozen images, European Conference on Computer Vision, Springer, Cham 2016, pp. 235–251.
[211] K. Kushal, C. Kanan, An analysis of visual question answering algorithms, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 1965–1973.
[212] H. Sharma, A.S. Jalal, Image captioning improved visual question answering, Multimedia Tools and Applications, 2021https://doi.org/10.1007/s11042-021-11276-2.
[213] X. Chen, M. Jiang, Q. Zhao, Predicting Human Scanpaths in Visual Question Answering, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 10876–10885.
[214] S. Whitehead, H. Wu, H. Ji, R. Feris, K. Saenko, Separating Skills and Concepts for Novel Visual Question Answering, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 5632–5641.
[215] A. Urooj, H. Kuehne, K. Duarte, C. Gan, N. Lobo, M. Shah, Found a Reason for me? Weakly- supervised Grounded Visual Question Answering using Capsules, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 8465–8474.
[216] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, J. Gao, Vinvl: Revisiting visual representations in vision-language models, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 5579–5588.
[217] Y. Hong, Q. Li, S.C. Zhu, S. Huang, VLGrammar: Grounded Grammar Induction of Vision and Language, arXiv preprint (2021).