Information Fusion 52 (2019) 268–280

Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

Information fusion in visual question answering: A Survey

Dongxiang Zhang[a], Rui Cao[b], Sai Wu[a,*]

[a] College of Computer Science and Technology, Zhejiang University, China
[b] School of Computer Science and Engineering, University of Electronic Science and Technology of China, China

ARTICLE INFO

Keywords:
Information fusion
Visual question answering
Survey

ABSTRACT

Visual question answering automatically answers natural language questions according to the content of an image or video. The task is challenging because it requires the understanding of semantic information in the textual and visual channels, as well as their interplay. A typical solver is composed of three components: feature extraction from singular modality, feature fusion between visual and textual channels, and answer prediction based on the learnt joint representation. Among them, information fusion plays a key role in enhancing the overall accuracy and various types of approaches have been proposed, such as simple vector operators, deep neural networks, bilinear pooling, attention mechanisms, and memory networks. The primary objective of this survey is to provide a clear organization and comprehensive review on the ever-proposed fusion techniques in the domain of visual question answering. We propose an abstract fusion framework that can fit the majority of existing VQA models, making it convenient for readers to quickly understand their key contributions. Finally, we summarize the effective fusion strategies that have been widely adopted so as to benefit readers in their model design.

1. Introduction

Visual question answering (VQA) is an interdisciplinary research problem that requires adequate foundation on computer vision and natural language processing techniques. The problem has witnessed significant progress towards complicated visual understanding and semantic reasoning tasks, owning to the availability of large-scale benchmark datasets and rapid advancement of deep learning techniques. It allows machines to reason across language and vision like human beings and is viewed as one key measurement for machine intelligence [1]. Advances in this area can benefit applications like blind person assistance [2], automous driving [3], smart camera on food images [4] and robot tutors with the function of aotomatic math problem solver [5,6]. As pointed out in [7], VQA can be used to automatically solve trivial tasks such as spotting an empty picnic table at the park, or locating the restroom at the other end of the room without having to ask. Finally, VQA can also inspire other relevant computer vision problems that require fine-grained understanding and reasoning on the semantic structures of images or videos.

According to the content in visual modality, VQA can be split into two branches: image question answering (image QA) and video question answering (video QA), both of which have attracted intensive attention in the past few years. In these two sub-domains, a considerable number of datasets have been crafted and published to benefit the community, either through crowdsourcing manner on Amazon Mechanical Turk [8] or weak-supervised algorithms [9,10]. To make the datasets diverse and challenging, they collect images or videos that cover a number of different scenes and the text questions are designed to inspect the understanding of fine-grained visual contents. In image QA, the popular benchmark datasets that are online available include COCO-QA [11], Visual7w [12], VQA 1.0 [13], and VQA 2.0 [14]. Their questions are devised with multiple types of answers, such as open-ended, yes/no, number counting, and multi-choices, to access the capability of a learnt model in terms of understanding the visual contents and textual semantics. Video QA is more difficult to solve than image QA, because its visual modality extends from one image to a long sequence of images and its questions can be designed to examine the content in one frame or multiple frames. It requires not only the visual comprehension within each image, but also the temporal relationship among multiple frames that have certain connection. Hence, we can view image QA as a special case of video QA. Similar to image QA, a number of datasets have been constructed and published from multiple sources, such as MovieQA [15], YouTube2Text [16], MSR-VTT [17], TGIF-QA [18], MarioQA [19] and ActivityNet-QA [20]. These datasets contain different types of video clips (e.g., game videos, movie clips, cartoons, etc) and questions are specially tailored to fit the contents of the videos.

An intelligent agent designed for VQA is required to fully understand the textual information in the question and visual information in the image or video, and leverage an effective fusion mechanism to find the semantic connection in these two modalities so as to return the correct answer. An end-to-end framework normally consists of three stages: 1) feature extraction, 2) feature fusion and 3) answer generation. More

* Corresponding author.
E-mail addresses: zhangdongxiang37@gmail.com (D. Zhang), caorui0503@gmail.com (R. Cao), wusai@zju.edu.cn (S. Wu).

https://doi.org/10.1016/j.inffus.2019.03.005
Received 21 December 2018; Received in revised form 5 March 2019; Accepted 11 March 2019
Available online 14 March 2019

specifically, given an image/video and a textual question, the first step is to extract condense and expressive features from the two information channels. In practice, traditional textual and visual feature extraction techniques are adopted, such as recurrent neural networks for textual data and convolutional neural networks for visual data. After that, the multi-modal features are fused to obtain the joint representation which captures key textual and visual component as well as their semantic connection. With the joint representation, the answer generation is converted into a classification or sequence generation problem, depending on the type of the question. If the question can be converted into a classification problem, a fully-connected layer with softmax function is normally applied. If it is open-ended, a decoder based on recurrent neural network is trained.

Generally speaking, information fusion from multiple modalities plays the vital role in enhancing the overall accuracy of VQA and often serves as the primary contribution in many previous publications. The objective of this survey is to provide a clear organization and comprehensive review on the ever-proposed fusion techniques in VQA. Since image QA is a topic studied earlier than video QA and quickly received significant attention, various efforts have been devoted to develop effective models. The proposed fusion techniques in image QA include simple vector operation, deep neural networks, bilinear pooling and various attention mechanisms. In contrast, video QA is still an emerging topic and researchers can take advantage of the lessons learnt from technical development in image QA. The fusion models proposed for video QA tend to reach an agreement, i.e., heavily relying on spatial and temporal attention mechanisms. An attention function uses features from one modality as the context and identifies important clues in the other modality, such as question terms, image regions or video frames. We also observed that a noticeable number of VQA models leverage memory networks, which is essentially a larger external memory, to facilitate the interaction of features from the two modalities. In this survey, we will take a close investigation to each type of fusion strategy. We propose an abstract fusion framework that allows us to clearly organize most of the existing models. More specially, we examine the fusion of a model in terms of which attention strategies are proposed and how the (attended) visual and textual are fused into a joint representation. Such organization makes it convenient for readers to quickly understand their key contributions. Finally, we summarize the popular fusion strategies that have been been shown effective. Readers can build their new models on top of the useful guidelines distilled from previous works.

It is worth noting that there have existed several other surveys on the topic of visual question answering. Most of them [21–23] are targeted at providing an overview on the dimensions of datasets and overall models. Alternatively, [24] is another survey that focuses only on the recent advancement in VQA. It explicitly explains the techniques for a selective number of recent works. In contrast, the scope of this survey is narrowed down to information fusion in VQA and we are the first paper to provide in-depth analysis of fusion techniques in VQA.

The remaining of the survey is organized as following. In Section 2, we present the tasks of image QA and video QA. In Section 3, we overview the feature extraction techniques from questions, images and videos. We present our abstract fusion framework in Section 4 to organize the previous models. We first review the various attention mechanisms, either one-hop or multiple-hops, between the visual and question channels. In Sections 6, we present fusion between two feature channels, based on simple vector operation, neural networks and bilinear pooling, respectively. Miscellaneous fusion techniques that cannot fit the abstract framework are discussed in Section 7. Finally, we provide several useful guidelines and conclude the paper in Section 8.

2. VQA task description

Given an image and a question, the task of image QA requires a machine to answer the question based on the visual clues in the image. The questions can be designed in different formats to inspect the generality and robustness of a learnt model. As shown in Fig. 1, we select four representative examples to exhibit the diversity. There is one open-ended question in which the answer "*Blue, white*", an object counting problem where the answer is a number, a multi-choice problem with four options, and a yes/no problem with two choices. To answer open-ended questions, a common practice is to train a decoder based on recurrent neural network from the text answers in the training dataset. The problems of object counting, multi-choice and yes/no can be viewed as instances of classification. They are associated with different answer vocabularies and can be solved by training a multi-class classifier. The candidate with the highest matching probability is considered as the answer.

Video QA extends the visual modality from an image to a video, which can be seen as a long sequence of images. Hence, there is an additional temporal dimension and we may need to consider the visual contents in multiple frames in order to answer a question. The datasets collected for video QA are also diversified in terms of video categories and question types. Fig. 2 illustrates several video QA examples from different datasets. In TGIF-QA [25], three types of reasoning tasks are defined: 1) count the number of repetitions of a given action; 2) detect a repeating action given its count; and 3) identify state transitions. It also includes frame QA tasks that can be answered from one of frames. The question in Fig. 2 is an example of state transition to ask "what does the model do after lower coat?". MarioQA [19] is constructed from gameplay videos "Super Mario Bros" with the goal of understanding temporal relationships between video events. The question would be like "*how many coin blocks were hit by Mario before a spiky appeared*". MovieQA [15] and TVQA [26] are harvested from movies and TV shows, respectively. These two datasets contain additional information sources such as subtitle-based dialogues, and their designed models will take advantage of the subtitles for information fusion.

There exist other types of visual understanding tasks that require the interaction between vision and language. For example, cross modal retrieval [27–29] considers similarity retrieval among the modalities of text, images and videos. A widely adopted solution framework is to project the textual and visual contents into the same latent space. It does not require to consider the fine-grained connection between a particular pair of question and image/video. Another example is image/video captioning [30–33] that intends to generate a text description for visual content. The training datasets contain both images/videos and their descriptions. Similar to VQA, we need to extract discriminative textual and visual features. However, the textual feature is normally used to train the decoder to generate the output description, similar to how VQA treats text answers in open-ended questions. Fusion between visual and textual features is not an important component worth particular attention in image/video captioning as well. Therefore, we can consider fusion as a crucial and distinctive component in VQA that deserves an intensive and close investigation. In the following, we first briefly summarize the feature extraction techniques and then focus on the various fusion mechanisms that have been ever proposed.

3. Feature extraction

Feature extraction from questions, images and videos is a fundamental component in VQA. In recent years, deep learning models have been shown to be very effective in extracting complex and discriminative features, and they have become the dominating approaches in the domain of VQA as well. In the following, we will review how deep learning models have been applied to extract visual and textual features.

3.1. Feature extraction in image QA

In the evolving line of convolutional neural networks (CNNs) proposed for image feature extraction, LeNet [34], AlexNet [35], GoogLeNet [36], VGG-Net [37] and ResNet [38] are representative and widely adopted networks. LeNet is a simple network structure with two

Question: What is the color of the comforter?
Answer: Blue, white

(a) Open-ended problem

Question: How many slices of pizza are there?
Answer: Six

(b) Counting problem

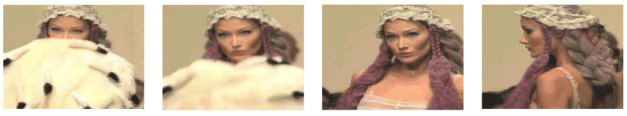
Question: Who is under the umbrella? Answer: A
A. Two women B. A child
C. An old man D. A husband and a wife

(c) Multi-choice problem

Question: Has the pizza been baked?
Answer: Yes

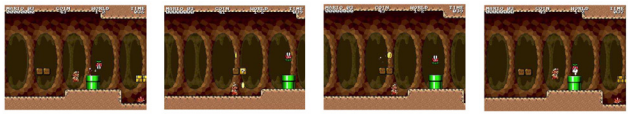
(d) Yes/no problem

Fig. 1. Examples of image QA problems.



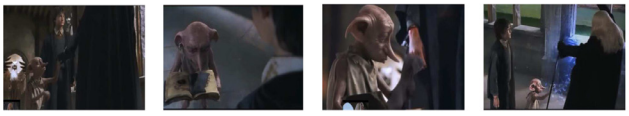
Question: What does the model do after lower coat?
Answer: Pivot around

(a) TGIF-QA



Question: How many coin blocks were hit by Mario before a spiky appeared?
Answer: 2

(b) MarioQA



Question: What does Harry trick Lucius into doing?
Answer: Freeing Dobby

(c) MovieQA

Question: What does Castle show Beckett after he turns around?
Answer: Vampire fangs

(d) TVQA

Fig. 2. Examples of video QA problems.

convolutional layers and used to solved simple tasks such as handwritten digit recognition. AlexNet, a deeper network with 5 convolutional layers, was the first deep network to boost the classification accuracy by a significant stride and won the championship of ILSVRC2012 (ImageNet Large Scale Visual Recognition Challenge). It used ReLu (Rectified Linear Unit) as the activation function which can be trained much faster than sigmoid or tanh. Dropout [39] was also adopted to avoid overfitting. In ILSVRC2014, GoogLeNet and VGG-Net were two of the best models for image classification, and consequently, they have become the new mainstream approaches to automatically extracting effective image features. Both of them deepen the convolutional network with more layers to capture more informative features. Moreover, GoogLeNet also increases the width of the network. Instead of concatenating convolved features directly, it applies 1*1 convolution to reduce dimension and accelerate the training speed. Since VQA was proposed in ICCV 2015 and has attracted a significant number of followers in a short time, there is no surprise to observe that almost of all of these early works in VQA adopted either GoogLeNet or VGG-Net for image feature extraction. As shown in Table 1, we summarize the selection of CNNs in various models proposed for VQA. We can see that VGG-Net and GoogLeNet have been very widely adopted until the appearance of ResNet [38]. ResNet [38], also called residual network, is a milestone work in the history of CNNs

Table 1
Visual feature extraction in image QA.

VGG-Net [37]	NMN+LSTM [42], LSTM Q+I [13], ABC-CNN [43], MCB [44], MRN [45], Attributes-CNN+LSTM [46], QRU [47], HieCoAtt [48], DPPnet [49], VIS+LSTM [50], VIS+LSTM-2 [50], High-Order [51], Word+Region Sel[52]., FVQA [53], SAN [54], LSTM+Att.(Visual7W) [12], NMN+Caprtion Information [55], MAN-VGG [56], DMN [57]
GoogLeNet [36]	mQA [58], Neural-Image-QA [59], SMem-VQA [60]
ResNet [38]	MUTAN [61], MCB [44], FDA [62], Shwo-Ask-Attend [63], MLB [64], MRN [45], HieCoAtt [48], Dual-MFA [65], DRAU [66], Explicit Knowledge-Base [67], High-Order [51], MLAN [68], MFB [69], MFH [70], QGHC [71], VQA-AA [62], CMF [72], FVTA [73], MAN-ResNet [56], DCN [74], CATL-QTA-M [75], VKMN [76], MF+SIG+VG [77]
Faster R-CNN [40]	Up-Down [41], Dual-MFA [65], DRAU [66], Explicit Knowledge-Base [67], Attention-on-Attention [78], Tips-Tricks [79],CVA [80], MLB+DA-NTN [81], BAN [82], CATL-QTA-M [75], Attention-on-Attention [78]

Table 2
Textual feature extraction in image QA.

LSTM [83]	NMN+LSTM [42], LSTM Q+I [13], ABC-CNN [43], OD+LSTM+RN [84], MCB [44], mQA [58], FDA [62], Shwo-Ask-Attend [63], HieCoAtt [48], Neural-Image-QA [59], VIS+LSTM [50], CNN+LSTM+RN [85], High-Order [51], Explicit Knowledge-Base [67], SMem-VQA [60], SAN [54], LSTM+Att.(Visual7W) [12], MFB [69], MFH [70], Word+Region Sel[52]., NMN+Caprtion Information [55], QGHC [71], VQA-AA [62], CMF [72], MAN [56], CATL-QTA-M [75], VKMN [76]
Bi-LSTM	ABC-CNN [43], DRAU [66], 2-VIS+BLSTM [50], FVTA [73], DCN [74]
GRU [86]	Up-Down [41], MUTAN [61], MRN [45], MLB [64], QRU [47], Dual-MFA [65], DPPnet [49], Attention-on-Attention [78], Tips-Tricks [79], DMN [57], MLAN [68], CVA [80], MLB+DA-NTN [81], BAN [82], Attention-on-Attention [78], MF+SIG+VG [77]
CNN	SAN [54], DAN-ECCV [87], CNN-QA [88], MAVQA [89]

because it effectively solves the training challenge brought by deep neural networks. As the network goes deeper, the problem of gradient exploding or vanishing becomes more severe. Its core idea lies in the concept of "identity shortcut connections" that allows to skip one or more layers. In the meanwhile, the whole network can still be trained by standard SGD. ResNet is able to train hundreds of layers and demonstrates superior performance over VGG-Net. We can observe in Table 1 that a big family of VQA works adopt ResNet to extract visual features.

A common practice in the aforementioned works is that they apply CNN model on the whole image to identify the relationship between global scene attributes and questions. There exists a branch of solutions that builds the semantic connection from a finer-grained perspective. More specifically, they first apply object recognition techniques such as Faster R-CNN [40] to identify important objects in the image. We can view the step as another form of attention mechanism such that the attended regions are restricted to pre-specified detection boxes. Such question-related regions could be more effective for answering questions about foreground objects [41]. The features in the detected boxes are extracted by ResNet or other CNNs to facilitate the identification of object attributes, quantity and categories and make accurate inference. The methods following the strategy are summarized in Table 1.

As to textual feature extraction, Table 2 summarizes the deep learning models applied in the past literature. We can see that the family of recurrent neural networks (RNNs), including LSTM, GRU and Bi-LSTM, were preferred choices as they have been shown to be remarkably effective when processing sequential data. In LSTM, there is a memory unit to control the flow of information. Its gate mechanism, such as forget gate, input gate and output gate, is developed to remember useful information and resolve the issue of vanishing gradient. Bi-LSTM is an extension of LSTM to capture the long-term dependency in both directions. A GRU only has two gates, a reset gate and an update gate. It has fewer parameters than LSTM and is computationally more efficient. Besides RNNs, convolutional neural networks (CNNs) were occasionally applied for feature extraction in VQA models, such as SAN [54], CNN-QA [88] and MAVQA [89]. As pointed out in [90], a pioneering work of applying CNN for sentence classification, CNN model is simple, easy to train and can even achieve promising performance. In the works of CNN-QA [88] and MAVQA [89], CNN was also empirically compared with RNN in terms of textual feature extraction and demonstrated superior performance. Alternatively, we can also observe in several NLP tasks, RNN models were reported to be better than CNN [91]. There also exist works, such as CRAN [92], which attempted to leverage the advantages of CNN and RNN and apply both of them for textual feature extraction. Therefore, it is difficult to conclude the definite superiority of CNN or RNN models. We suggest that the selection of textual extraction models could be dependent on the designed VQA models.

3.2. Feature extraction in video QA

We summarize the visual and textual feature extraction models in Tables 3 and 4, respectively. There are two types of visual feature extraction models. One is functioned on each video frame and contains the popular image feature extraction models such as VGG-Net, GoogLeNet and ResNet that have been widely adopted in image QA. The other is unique for video data as it takes into account the temporal attribute. For example, the two-dimensional CNN model is extended to 3-dimensional, with small $3\times3\times3$ convolution kernels, for spatio-temporal feature learning. The model is often called C3D [93] and has been adopted by multiple video QA solvers [8,19,25,94]. To capture motion information, optical flow [95] is extracted as a unique type of video feature in [96].

In video QA, the textual extraction models for questions are the same as those proposed for image QA. LSTM and GRU are preferred choices to encode semantic information in the textual dimension. It is worth noting that there are several video QA datasets associated with abundant textual descriptions to facilitate question answering. We noticed that a considerable number of models, as summarized in Table 4 adopt SkipThoughts to encode the descriptions instead of using weighted sum on the word embeddings. The reason is that SkipThoughts can better capture the semantic context in sentence level. It maps sentences that share semantic and syntactic properties to similar vector representations and is robust to handle unseen vocabulary (Tables 5 and 6).

4. Overview of fusion strategies

In this section, we present an abstract framework for the majority of end-to-end models designed for VQA. As shown in Fig. 3, the input of image/video QA contains a textual question and a visual image/video. We need to first apply the feature extraction models, which have been examined in Section 3, to extract the textual and visual features. After that, a significant number of VQA solvers will apply attention mechanism to assign higher weight for the important question words, image regions or video frames. We consider attention mechanism as one type of information fusion between visual and textual channels, because it requires the other modality as the context to determine the weight for the current modality. In certain works, the attention module may contain multiple layers, i.e., there would be multiple hops of interaction between the two channels. The details of attention mechanism used in VQA will be presented in Section 5. It is worth noting that attention is an optional component in model design and may not be adopted in all

Table 3
Visual feature extraction in video QA.

VGG-Net [37]	E-MN [8], E-VQA [8], E-SA [8], E-SS [8], refine-Att [94], r-STAN [97], Unified [98], LMN+(V) [99], MHACN [100], AHN [101]
GoogLeNet [36]	Cosine SkipThought [15], SSCB SkipThought [15], Cosine Word2Vec [15], SSCB Word2Vec [15], Uncovering-Temporal-Context [102], LMN+(G) [99]
ResNet [38]	r-ANL [9], ST-VQA-Sp [25], ST-VQA-Tp [25], ST-VQA-Sp.Tp [25], RWMN [103], DEMN [104], Co-memory [96], TVQA [26], JSFusion [105]
C3D [93]	E-MN [8], E-VQA [8], E-SA [8], E-SS [8], refine-Att [94], ST-VQA-Sp [25], ST-VQA-Tp [25], ST-VQA-Sp.Tp [25], MarioQA [19]
Optical flow [95]	Co-memory [96]

Table 4
Textual feature extraction in video QA.

SkipThought [106]	Cosine SkipThought [15], SSCB SkipThought [15], E-MN [8], E-VQA [8], E-SA [8], E-SS [8], DEMN [104]
Word2Vec	Cosine Word2Vec [15], SSCB Word2Vec [15], r-ANL [9], RWMN [103], r-STAN [97], LMN [99], MHACN [100], AHN [101]
LSTM [83]	refine-Att [94], ST-VQA-Sp [25], ST-VQA-Tp [25], ST-VQA-Sp.Tp [25], Unified [98], TVQA [26]
GRU [86]	MarioQA [19], Uncovering-Temporal-Context [102], Co-memory [96]

Table 5
Single-hop attention for visual QA.

Attention in Visual Channel	Image QA	ABC-CNN [43], CVA [80], LSTM-Att [12], Word+Region Sel[52]., Up-Down [41], Attention-on-Attention [78], FDA [62], SAN [54], MRN [45], Show-Ask-Attend [63], MLAN [68], VQA-AA [2], Dual-MFA [65], MTA [114], QGHC [71], VKMN [76]
	Video QA	ST-VQA-Sp [25], ST-VQA-Tp [25], ST-VQA-Sp.Tp [25], refine-Att [94]
Co-Attention in Both Channels	Image QA	MFB [69], High-Order [51], DRAU [66], MFH [70], MAN [56], CMF [72], FVTA [73]
	Video QA	Unified [98], TVQA [26]

Table 6
Multi-layer attention for visual QA.

Attention in Visual Channel	Image QA	SAN [54], MRN [45], SMem-VQA [60], MF+SIG+VG [77]
	Video QA	r-ANL [9], r-STAN [97], MHACN [100], GC [19]
Attention in Textual Channel	Image QA	QRU [47]
	Video QA	
Co-Attention in Both Channels	Image QA	DAN [115], DCN [74]
	Video QA	

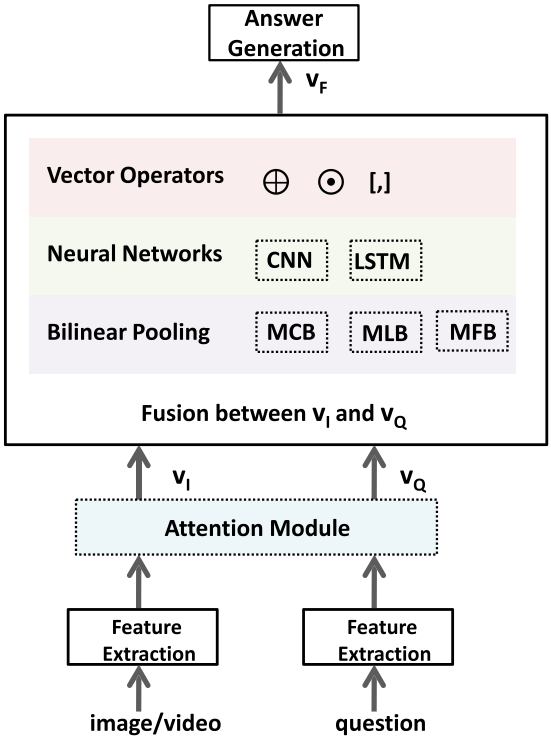


Fig. 3. Abstract representation of an end-to-end framework for VQA.

the previous works, even though it normally plays a positive effect in improving the accuracy. With the visual and textual features $\mathbf{v_I}$ and $\mathbf{v_Q}$, no matter they are attended or not, they can be fused, by simple vector operation, neural networks or bilinear pooling, to generate the joint representation $\mathbf{v_F}$. The answer generation module, which is a classifier or RNN-based decoder, accepts $\mathbf{v_F}$ and outputs the answer.

In the following sections, we focus on summarizing the works that follow the abstract model in Fig. 3. The exceptional models will be discussed in Section 7 as miscellaneous fusion techniques. For ease of presentation, we also make the following simplification to omit minor issues and focus on the important stems.

1. The features in the visual channel could be extracted from the modality of images or videos. They can also represent the global feature from the whole image or local features extracted from visual regions. For simplicity, we do not differentiate the visual sources and use $\mathbf{v_I}$ as the uniform representation.
2. In image QA, we only review the models proposed for datasets of VQA 1.0 and VQA 2.0. The models proposed for vertical domains of image QA, such as CLEVR dataset [77,84,85], TDIUC [75,107], MemexQA [73,108],will be omitted in this survey. CLEVR focuses on the reasoning on spatial objects and MemexQA is constructed to automatically answer questions that help users recover their memory about events captured in the collection. They are not as general or diversified as VQA 1.0 and VQA 2.0.
3. The embedding functions to align the dimension of $\mathbf{v_I}$ and $\mathbf{v_Q}$ will be omitted. We simply assume that $\mathbf{v_I}$ and $\mathbf{v_Q}$ are always with the same dimension. The bias term will also be ignored when we present the math equations.
4. The activation function on $\mathbf{v_I}$, $\mathbf{v_Q}$ or the fused feature $\mathbf{v_F}$ as a post-processing step will be omitted as we focus on the fusion component. For instances, [62] applies *Tanh* on the question representation and *ReLU* on the image representation. [58] and [43] use activation function, called scaled hyperbolic tangent function [109], for the fused feature $\mathbf{v_F}$.
5. For region-based image feature extraction, multiple visual features $\mathbf{v_i}$ are generated for the regions $\mathbf{r}_i$. If the local $\mathbf{v_i}$ are aggregated into a global $\mathbf{v_I}$ before interacting with $\mathbf{v_Q}$, we consider that it follows the framework in Fig. 3. Otherwise, each $\mathbf{v_i}$ may inter-

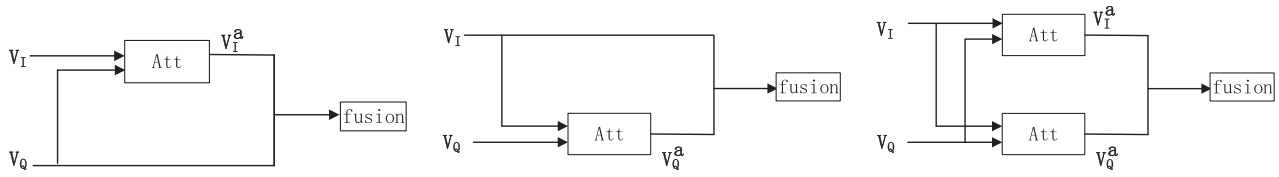


Fig. 4. Different types of single-layer attention strategies.

act with $\mathbf{v_Q}$ independently before the aggregation of local features. We consider it a violation of Fig. 3 and will be introduced in Section 7.

5. Attention mechanisms

Attention mechanism has been shown effective to enhance the interaction between the visual and textual channels. The underlying motivation is to pay more attention to the important terms in the question, regions in the query image or frames in the video, that can facilitate answer generation. Methods without attention can be viewed as coarse joint-embedding models with global features, which may contain noise and are unable to answer fine-grained questions [69]. According to the number of attention layers, we classify the existing approaches into single-hop attention and multi-hop attention. An *m*-hop attention can be viewed as applying the single-hop attention iteratively for *m* times. In the following, we elaborate the various single-hop attention mechanisms that have been proposed. After that, we present how to cascade them into multiple layers. Hence, only visual attention and co-attention strategies will be reviewed for single-hop attention.

5.1. Single-hop attention

The basic principle of attention used in VQA is to derive the weight vector for the features in one modality, by using the information from the other modality as the context or guidance. According to the modality in which attention weight is derived, we can classify single-hop attention into *visual attention, textual attention* and *co-attention*. An abstract representation of each type of attention is shown in Fig. 4. The attention module is represented by a black-box whose details will be explained in the following. To the best of our knowledge, we did not find any VQA models that simply rely on textual attention and ignores visual attention.

5.2. Visual attention

In models with visual attention, the overall idea is to use $\mathbf{v_Q}$ as the context to achieve an attention weight vector $\boldsymbol{\alpha}$, which is then applied on the visual features to derive $\mathbf{v_I^a}$, i.e., $\mathbf{v_I^a} = \boldsymbol{\alpha}\mathbf{v_I}$. The most widely adopted approach by to obtain the attention weight $\mathbf{v_I^a}$ is to first capture the correlation or similarity between $\mathbf{v_I}$ and $\mathbf{v_Q}$ and normalize it through softmax function:

$$\mathbf{z} = \mathbf{W_3}^T \tanh(\mathbf{W_1 v_I} + \mathbf{W_2 v_Q}) \tag{1}$$

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{z}) \tag{2}$$

We observe that the works [12,41,68,80] have adopted such an attention mechanism. Alternatively, there are other ways to achieve the attention weight vector. In [52], vector $\mathbf{z}$ is simply defined as the product product between $\mathbf{v_I}$ and $\mathbf{v_Q}$ and softmaxed to get $\boldsymbol{\alpha}$. In [43], $\mathbf{z}$ is defined as $\mathbf{z} = \text{sigmod}(\mathbf{W v_Q}) * \mathbf{v_I}$, where $*$ represents a convolutional operator.

The aforementioned visual attention mechanisms share the same strategy. They first calculate the correlation between the two channels, though in different ways, and normalize it through the softmax function. There exist other approaches exceptional from the framework. For example, [62] first calculates the similarity between the labels and question word in the embedding space [110]. The objects with similarity higher than 0.5 are retained. These visual objects are ordered by the sequence of their matching words and fused in a LSTM network to derive the visual feature. In [2], Ilievski et al. argued that existing visual attention mechanisms are mainly based on linear transformation and not able to fully capture the complex interaction between visual and textual channels. They proposed a novel attention mechanism based on generative adversarial networks (GAN) [111] to generate the visual attention map. Another type of novel visual attention is called differential attention [112]. It relies on additional supporting and opposing examples to obtain a differential attention region which is expected to be closer to human attention. For interested readers, up to 13 types of visual attention strategies for image QA were empirically evaluated in [78].

The main difference in the attention mechanism between image QA and video QA is that there is an additional temporal dimension for video QA. It is often required to consider the relevance between a particular frame and the question. In [25], two types of attentions are considered. One is called spatial attention to learn which regions in a frame to attend for each question word. This part is similar to the aforementioned visual attention of local regions in image QA. The other is temporal dimension to learn which frames in a video to attend to. In this work, both spatial and temporal attentions are computed with the same function f_{att} by feeding different inputs. [94] considers both appearance and motion as the visual channels and processes the question word by word. Attention Memory Unit (AMU) was proposed to generate and refine the attention over appearance and motion features of the video at each timestamp. The visual attention is defined as

$$\mathbf{z} = \tanh(\mathbf{W_1 v_I}) + \tanh(\mathbf{W_2 v_{Q_t}}) \tag{3}$$

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{z}) \tag{4}$$

where $\mathbf{v_I}$ could be motion or appearance features and $\mathbf{v_{Q_t}}$ is the feature for the *t*th word.

5.2.1. Co-attention

Co-attention implies that two attention weight vectors will be derived in both visual and question channels. In the fusion models with bilinear pooling, as will be introduced in Section 6.3, the co-attention mechanism is often designed as symmetric, i.e., the two attention vectors in the visual and textual channels are computed in the same fashion and can be conducted in parallel. In [51] with one-hop co-attention, a similarity matrix that captures high-order correlations between the two modalities is learnt. The (*i, j*)th entry in the matrix represents the correlation (inner-product) of the *i*-column and *j*th row of embedded visual and textual features, i.e., the *i*th word in a question and *j*th patch in an image. The visual and textual attention is obtained by convolving a row or a column in the matrix. The Dual Recurrent Attention Unit (DRAU) proposed in [66] is also symmetric. Its attention weight is calculated as

$$\mathbf{z} = \text{PReLU}(\mathbf{W_2}\ \text{LSTM}(\text{PReLU}(\mathbf{W_1 v}))) \tag{5}$$

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{z}) \tag{6}$$

where PReLU activation [113] is used and $\mathbf{v}$ could be textual or visual features. In [56], vector $\mathbf{z}$ has a different representation:

$$\mathbf{z} = \tanh(\mathbf{W_1 v}) + \tanh(\mathbf{W_2 v_I} \odot \mathbf{v_Q}) \tag{7}$$

To calculate the visual or question attention weight, we simply need to replace $\mathbf{v}$ with $\mathbf{v_I}$ or $\mathbf{v_Q}$.

The co-attention mechanism proposed by Yu et al. in [69] and [70] is not symmetric. Instead, their image and question modules are loosely coupled such that the image features are not exploited when learning the question attention module, as they expect the network to be able to infer the key words of the question without seeing the image. The attended question feature is merged with each of the 196 (14×14 grids) image features using MCB, followed by some feature transformations (e.g., 1×1 convolution and ReLU activation) and softmax normalization to predict the attention weight for each grid location.

In video QA, taking into account term attention is not a common strategy. We observed that [98] interprets the token attention as the degree to which the network attends to a particular frame in the question sentence when watching the videos. The weighted tokens are then accumulated to derive the final representation for the question sentence.

5.3. Multi-hop attention

Similar to one-hop attention, we also organize the models according to the modality in which the attention weight is generated and classify them into *visual attention, textual attention* and *co-attention*. The difference is that we observed an instance of model that purely relies on textual attention.

5.3.1. Visual attention

In [54], Yang et al. argued that visual question answering often requires multiple steps of reasoning. Given a query like "*what are sitting in the basket on a bicycle*", we need to locate the objects (e.g., *basket* and *bicycle*) and concepts (e.g., *sitting in*) mentioned in the question, then filter the irrelevant objects and pinpoint the region that we can derive the final answer. They proposed SAN (Stacked Attention Network) that extends the attention mechanism from one-layer to multiple layers. The objective is to attend to image regions iteratively according to the query vector. For ease of presentation, we use $\mathbf{v_F^{(1)}}$ to represent the fused feature from the single-layer attention and only plot the SAN with two additional layers. The output $\mathbf{v_F^{(3)}}$ can be considered as the fused output that would be passed to the classifier for answer generation. It is worth noting that the number of attention layers is a customizable parameter. Formally, we present the equations of SAN to help readers understand the implementation details of attention function and the vector-based operator:

$$\mathbf{z_v^{(k)}} = \mathbf{W_3}^T \tanh(\mathbf{W_1^{(k)} v_I} + \mathbf{W_2^{(k)} v_F^{(k-1)}}) \tag{8}$$

$$\boldsymbol{\alpha}^{(\mathbf{k})} = \mathrm{softmax}(\mathbf{z_v^{(k)}}) \tag{9}$$

$$\mathbf{v_I^{(k)}} = \boldsymbol{\alpha}^{(\mathbf{k})} \mathbf{v_I} \tag{10}$$

$$\mathbf{v_F^{(k)}} = \mathbf{v_I^{(k)}} + \mathbf{v_F^{(k-1)}} \tag{11}$$

where $\mathbf{k}$ refers to the $\mathbf{k}$th iteration or layer and $\boldsymbol{\alpha}^{(\mathbf{k})}\mathbf{v_I}$. The vector operator is element-wise addition to fuse $\boldsymbol{\alpha}^{(\mathbf{k})}\mathbf{v_I}$ and $\mathbf{v_F^{(k-1)}}$.

The accuracy of SAN is not promising in the VQA 1.0 dataset. MRN [45] improves SAN from an interesting perspective. In order to apply deep residual learning, whose effectiveness has been well acknowledged, MRN converts the attention function into a non-linear module:

$$\mathbf{v_I^{(k)}} = \tanh(\mathbf{W_q^{(k)} v_F^{(k-1)}}) \odot \tanh(\mathbf{W_2^{(k)}} \tanh(\mathbf{W_I^{(k)} v_I})) \tag{12}$$

It directly uses the global visual feature for the element-wise multiplication, rather than the weighted visual features from multiple regions. In this way, MRN can be seen as an implicit attention model without explicit attention parameters. By converting Eq. (11) into $\mathbf{v_I^{(k)}} = \mathbf{v_F^{(k)}} - \mathbf{v_F^{(k-1)}}$, we know that the parameters in the non-linear module are optimized to fit $\mathbf{v_F^{(k)}} - \mathbf{v_F^{(k-1)}}$. This is very similar to the deep residual learning. Show-Ask-Attend [63] also used the same stacked attention network as proposed in [54]. However, the authors reported that using stacked attentions only lead to marginal improvement. They also observed that a two-layer deep classifier can be significantly better than a single layer. A two-hop attention machanism was proposed in [60]. Similar to [51], its visual attention is derived from the correlation matrix between visual features and question word vectors, i.e., $\mathbf{C} = \mathbf{v_I}(\mathbf{W v_Q})$. The visual attention weights are calculated by taking the maximum over the word dimension in the correlation matrix $\mathbf{C}$.

As to video QA, we found that the models [9] and [19] adopt the same multi-hop attention strategy to recursively update the temporal attention for multiple times. In the models proposed by Zhao et al. [97,100], both spatial and temporal attention are considered in the network and they are extended to multi-hop to facilitate the multi-step reasoning process.

5.3.2. Textual attention

QRU (Question Representation Update) [47] is another type of multi-layer attention network. Its main difference with SAN is that the attention weight vector in SAN is functioned on $\mathbf{v_I}$, whereas the similar attention vector in QRU is functioned on $\mathbf{v_Q}$. The motivation is that the resulting question representation from RNN or CNN lacks detailed information from the given image. The multi-layer reasoning in the textual channel is targeted at making the question to be more specific. The model adopts the same strategy as in [84] and [85] and uses MLP to generate the correlation between the question and image features.

5.3.3. Co-Attention

In image QA, DAN (Dual Attention Network) [115] has complicated attention mechanism as it extends co-attention mechanism into multiple hops. It derives two attention vectors in a symmetric manner, one for textual features and the other for image features. Their outputs are aggregated through two vector-based operators:

$$\mathbf{v_F^{(k)}} = \mathbf{v_F^{(k-1)}} + \mathbf{v_I^{(k)}} \odot \mathbf{v_Q^{(k)}} \tag{13}$$

where $\mathbf{v_I^{(k)}}$ and $\mathbf{v_Q^{(k)}}$ are the attended visual and textual features in the kth step. The number of attention steps is set to 2 which empirically shows the best performance. In [74], a symmetric attention mechanism was prposed to enable dense and bi-directional interactions between visual and textual channels. The attention weight is derived by first building an affinity matrix $\mathbf{A}$ between $\mathbf{v_I}$ and $\mathbf{v_Q}$. The matrix $\mathbf{A}$ (or its transpose) then softmaxed to obtain the attention weight for question words (or image regions). The attention module can be stacked into multiple layers for multi-step interaction. Inspired by the recurrent residual framework [116], the Cross-modal Multistep Fusion (CMF) [72] is also symmetric. Its CMF unit can produce attended question and image features. Multiple CMF units can be cascaded for multi-hop co-attention.

6. Fusion between two feature channels

6.1. Fusion based on simple vector operation

With visual feature $\mathbf{v_I}$ and textual feature $\mathbf{v_Q}$, the most straightforward and common approaches that directly utilize simple vector-based operation to generate the joint representation. As summarized in Table 7, three types of operators including vector concatenation, element-wise addition and element-wise multiplication, have been adopted by previous methods. Vector concatenation puts $\mathbf{v_I}$ and $\mathbf{v_Q}$ together to form a $(m+n)$-dimensional vector $\mathbf{v_F}$ as follows:

$$\mathbf{v_F} = [\mathbf{v_I}, \mathbf{v_Q}] \quad \text{(vector concatenation)} \tag{14}$$

Table 7
VQA: fusion based on simple vector operation.

Element-wise Addition	Image QA	SMem-VQA [60], CVA [80], SAN [54], mQA [58], ABC-CNN [43], NMN + Caption Information [55], MTA [114]
	Video QA	r-ANL [9], ST-VQA-Tp [25], r-STAN [97], MHACN [100], GC [15], ST [8], RWMN [103], LMN [99]
Element-wise Multiplication	Image QA	LSTM Q+I [13], LSTM-Att [12], Up-Down [41], Attention-on-Attention [78], FDA [62], VQA-AA [2], VKMN [76]
	Video QA	CG [19]
Vector Concatenation	Image QA	DCN [74], FVTA [73], MAN [56], CATL-QTA-M [75]
	Video QA	MDAM [117]

As to element-wise addition and multiplication, they require $\mathbf{v_I}$ and $\mathbf{v_Q}$ to be associated with the same dimension. If not, linear projection is required before we can apply these two vector-based operators, as appeared in [115]:

$$\mathbf{v_I} = \mathbf{W_v}\mathbf{v_I} \tag{15}$$

$$\mathbf{v_Q} = \mathbf{W_q}\mathbf{v_Q} \tag{16}$$

With $\mathbf{W_v}$ and $\mathbf{W_q}$, the two vectors are embedded to the same space, say $\mathbb{R}^n$. Then, we can derive the fused vector $\mathbf{v_F}$:

$$\mathbf{v_F} = \mathbf{v_I} + \mathbf{v_Q} \quad \text{(element-wise addition)} \tag{17}$$

$$\mathbf{v_F} = \mathbf{v_I} \odot \mathbf{v_Q} \quad \text{(element-wise multiplication)} \tag{18}$$

Directly use these simple operators to fuse the two channels is not effective and only very few early works adopt the strategy. For example, the pioneering work of VQA [13] presented the dataset of VQA 1.0. A baseline that directly uses element-wise multiplication to fuse the question and image encodings was also proposed. In practice, most models apply the vector operators on the attended features of $\mathbf{v_I}$ and $\mathbf{v_Q}$. From the categorization in Table 7, element-wise addition is the most favourable fusion strategy between attended visual and textual features. In contrast, vector concatenation was not frequently used and only appears in [74] and [117]. The reason could be that vector concatenation will increase the dimension of joint feature space.

6.2. Fusion based on neural networks

In the preceding subsection, we have introduced the fusion of attended visual and textual features through simple vector-based operation. Here, we review another type of fusion mechanism that adopts the non-linear neural networks to learn the interaction between the visual and textual features.

6.2.1. LSTM-based fusion

Since LSTM is a popular recurrent neural network adopted by VQA models to handle sequential data in the question, a natural and simple idea is to embed the image feature $\mathbf{v_I}$ into the same dimension as the word embedding and treat $\mathbf{v_I}$ as one of the "words" in the question. Ren et al. adopt this strategy in [50] to support image question answering in the datasets of DAQUAR and COCO-QA, in which the answers consist of one or multiple words. The image feature $\mathbf{v_I}$ extracted from CNN is projected to the word embedding space and treated as the first term in the question. In this way, the input to the LSTM model is represented as $(\mathbf{v_I}, \mathbf{v_{Q_1}}, \mathbf{v_{Q_2}}, \ldots, \mathbf{v_{Q_m}})$, where $\mathbf{v_{Q_i}}$ refers to the embedding of the i-th word in the question. Since the output is a sequence with one or multiple words, the LSTM network is viewed as a sequence-to-sequence model and can directly generate the answer. To enhance the interaction between $\mathbf{v_I}$ and $\mathbf{v_{Q_i}}$, [59] presents another fusion strategy with LSTM. Instead of treating $\mathbf{v_I}$ as the first word, it is concatenated with all the word embeddings. In other words, the input $\mathbf{v_Q}$ is augmented into $([\mathbf{v_I}, \mathbf{v_{Q_1}}], [\mathbf{v_I}, \mathbf{v_{Q_2}}], \ldots, [\mathbf{v_I}, \mathbf{v_{Q_m}}])$. Other types of improvement include placing $\mathbf{v_I}$ in both front and tail positions of $\mathbf{v_Q}$ [11], and the input becomes $(\mathbf{v_I}, \mathbf{v_{Q_1}}, \mathbf{v_{Q_2}}, \ldots, \mathbf{v_{Q_m}}, \mathbf{v_I})$. Moreover, an additional reverse LSTM layer can be added optionally in [11] to make the RNN become bi-directional. In [56], memory-augmented neural network [118] was applied to maintain a relatively long memory. The standard LSTM network was exploited to control the interaction between input data and external memory module using a number of read and write heads.

In video QA, [98] presents a two-stacked single-layer LSTM to fuse the information from the two channels. The visual features are used as the input for the bottom-layer LSTM. Its outputs, together with the textual features, are fused in the next layer of LSTM, whose output is a sequence of tokens as the answer for an open-ended question.

6.2.2. CNN-based fusion

It was argued in [88] that treating the image representation as an individual word may not well exploit the complicate relation between the image and the high-level semantic representations because the effect of image will vanish at each time step of LSTM. To deal with the problem, CNN model was proposed as an alternative type of neural network based fusion approachin [88]. Given features $\mathbf{v_I}$ and $\mathbf{v_T}$, both extracted by CNN, are fused by a multi-modal convolution process to adequately exploit their interactions. In [49], Noh et al. proposed an intersting fusion mechanisim based on CNN. They adopt VGG-16 [37] for image feature extraction. However, they remove the last layer in VGG-16 and attach three fully-connected (FC) layers. The tricky part is that they replace the second FC layer with dynamic parameters, which are derived from the output of GRU for textual feature extraction. The mapping between the output of GRU and the dynamic weight layer in VGG-16 is implemented via hashing mechanism. In [71], Gao et al. proposed language-guided hybrid convolution for feature fusion. The model performs convolution on visual feature maps and the convolution kernels are predicted by the guidance of question features.

In video QA, [105] uses CNN to composes pairwise joint representation of language and video sequences into a 3D tensor, using a soft-attention mechanism. The attention matrix is expected to figure out which pairs of video frames and question words are more closely related. A convolutional hierarchical decoder was proposed to learn the hierarchical relation patterns between the two sequences.

6.3. Fusion based on bi-linear models

Bilinear pooling computes the outer product between two vectors and provides multiplicative interaction between all elements of both vectors. Tenenbaum et al. [119] proposed the idea of bilinear pooling models to solve two-factor tasks, such as separating style and content, to fully capture the interaction between data sources. However, directly applying bilinear pooling in VQA is not effective because it takes into account of the interaction of each pair of visual and textual element. As pointed out in [44], if the dimensions of $\mathbf{v_I}$ and $\mathbf{v_Q}$ are both set to 2048 and there are 3000 classes for VQA task, the naive bilinear model can generate 12.5 billion parameters, leading to tremendous memory consumption and computation cost.

There have been several types of improvements proposed to tackle the issue of huge parameter space. Multimodal Compact Bilinear Pooling (MCB) was proposed in [44] to compress bilinear models. The image and text features are randomly projected to a common space by count sketch

function [120]. It has been proven in [121] that the count sketch of the outer product of two vectors can be expressed as convolution of both count sketches: $\Psi(\mathbf{v_I} \otimes \mathbf{v_Q}, h, s) = \Psi(\mathbf{v_I}, h, s) * \Psi(\mathbf{v_Q}, h, s)$, where $*$ is the convolution operator and, h and s are randomly sampled parameters by the algorithm. In this way, we can avoid computing the outer product explicitly. Furthermore, Fast Fourier Transformation (FFT) is adopted to replace the convolution operator with more efficient element-wise multiplication. In [51], MCB was also adopted as the fusion strategy, except that it was functioned on the attended visual and textual features. Yu et al. [69,70] also applied MCB on the attended features, but with a different attention mechanism.

The drawback of MCB is that it still requires a high dimensional feature space to achieve satisfactory performance. The dimension in the projection space is set to 16,000 in the experimental study and the total number of parameters is around 48 million. To further reduce the number of parameters, Kim et al. proposed a multimodal low-rank bilinear pooling method (MLB) [64]. The idea is to rewrite the weight matrix into the multiplication of two small matrices [122], i.e., $\mathbf{W} = \mathbf{UV}$, where $\mathbf{W} \in \mathbb{R}^{m \times n}$, $\mathbf{U} \in \mathbb{R}^{m \times d}$, $\mathbf{V} \in \mathbb{R}^{d \times n}$ and $d \leq \min(m, n)$. In this way, the projected feature can be written as

$$f_i = \mathbf{v_I}^T \mathbf{W}_i \mathbf{v_Q} = \mathbf{1}^T (\mathbf{U}_i^T \mathbf{v_I} \odot \mathbf{V}_i^T \mathbf{v_Q}). \tag{19}$$

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

$$f = \mathbf{P}^T (\mathbf{U}^T \mathbf{v_I} \odot \mathbf{V}^T \mathbf{v_Q}). \tag{20}$$

The evaluation in [66] shows that MLB can achieve more accurate results than MCB. However, MLB is slow in terms of convergence and sensitive to hyper-parameters.

MFB (Multimodal Factorized Bilinear Model) [69] can be viewed as an enhanced version of MCB. It preserves the compact output representation of MCB, but with better training stability. In MCB, we have $z = \mathbf{U}^T \mathbf{v_I} \odot \mathbf{V}^T \mathbf{v_Q}$. In MFB, the features $\mathbf{v_I}$ and $\mathbf{v_Q}$ are first expanded to high-dimensional space and then squeezed by sum pooling into compact output feature.

$$z = \text{SumPooling}(\tilde{\mathbf{U}}^T \mathbf{v_I} \odot \tilde{\mathbf{V}}^T \mathbf{v_Q}, k) \tag{21}$$

Since the magnitude of the output neurons may vary dramatically as element-wise production is used, power normalization and L_2 normalization are appended after MFB to improve training stability. MFB requires more parameters than MLB and is able to learn more powerful features. MLB can be viewed as a special case of MFB when $k = 1$. The subsequent work by Yu et al. [70], namely MFH, cascades multiple MFB blocks to further enhance the representation capacity of fused features. Element-wise multiplication is performed between the output of the current MFB and cascaded features by previous blocks:

$$\mathbf{z_{exp}}^i = \mathbf{z_{exp}}^{i-1} \odot \text{Dropout}(\tilde{\mathbf{U}}^T \mathbf{v_I} \odot \tilde{\mathbf{V}}^T \mathbf{v_Q}) \tag{22}$$

where i refers to the i-th MFB block. All the outputs of MFB blocks are squeezed by sum pooling to generate the final fused feature:

$$\mathbf{z} = \text{SumPool}(\mathbf{z_{exp}}) \tag{23}$$

Alternatively, MUTAN [61] applied tensor-based Tucker decomposition to its full bilinear model for VQA and showed that MCB and MLB are its spacial cases. It decomposes the weighting tensor in bilinear model into three factor matrices and a core tensor:

$$\mathbf{W} = \mathcal{T}_\mathbf{c} \times \mathbf{W_q} \times \mathbf{W_v} \times \mathbf{W_o} \tag{24}$$

The number of parameters can be controlled by imposing constraints on the core tensor. [81] constructs a bilinear tensor to model the pairwise interaction. A slice-wise attention module was also proposed to select the most discriminative reasoning process for inference.

7. Miscellaneous fusion techniques

Up to here, we have examined the VQA models that fit the abstract framework in Fig. 3. In this section, we review the remaining works with miscellaneous fusion techniques.

7.1. Attended single-modality features

There exist several works that do not apply fusion between the attended visual and textual features. Instead, they directly apply the classified on the attended features $\mathbf{v_I}$ or $\mathbf{v_Q}$ because they consider that the attention mechanism has taken into account the interaction of the two channels. Show-Ask-Attend [63] uses only the attended visual features and QRU [47] uses only the attended textual features.

7.2. Fusion among multiple feature channels

In our default fusion framework, there is one (attended) visual channel and one (attended) textual channel. In certain proposed models, there could be multiple visual or textual channels involved for fusion to capture fine-grained interaction.

The models hat use Faster R-CNN to detect multiple object regions would generate multiple visual features. Each feature captures the visual information about a local region and may directly interact with question features. [52] used dot product to derive attention weight for the local regions. Each attended visual feature is concatenated with textual feature and considered as one feature channel. These channels are aggregated into a joint global feature by weighted average. In [48], multiple channels of textual features, including word embedding, phrase embedding and question embedding, are attended with the image feature. After that, each channel of textual feature is aggregated with visual feature via element-wise addition. Finally, these three channels are fused recursively via vector concatenation followed by a multi-layer perception.

Both [68] and [65] generate two visual channels, one to capture image-level global information from the whole image and the other to capture concept-level information from local regions. To fuse these two visual channels with the textual feature, they adopt vector-based operation. In [68], each visual channel is first aggregated with the textual feature by element-wise addition. Then, these two channels are further aggregated by element-wise multiplication to generate the joint representation for classification. The sequence of the two vector-based operators are applied in a reverse order in [65].

Instead of generating multiple visual channels, [55] extends textual information into multiple channels. It utilizes existing image captioning techniques [123] to associate the query image with an additional caption, which is further attended with the question. The other textual source is to query Wikipedia or other external knowledge bases. These textual channels are fused by element-wise addition.

BAN [82] is a bilinear-pooling solution extended to handle multiple visual and textual channels and improves MLB by considering bilinear interactions from a fine-granularity perspective. It replaces the global features $\mathbf{v_I}$ and $\mathbf{v_Q}$ with two groups of input channels, one group for the terms in the question and the other for the visual regions in the image. The low-rank bilinear pooling is then used to extract the joint representations for each pair of channels.

In video QA, [94] is a work that fuses three channels. Besides the attended question and video features, the fusion also takes into account the output of AMU (Attention Memory Unit) which contains information about attention history. The three channels are fused by element-wise multiplication and its result is used for answer generation.

7.3. Memory network

Memory networks (MemN2N) [124] were originally proposed for text QA to model the complex relationship between stories, questions and answer. One key benefit of external memory is to enable a neural model to cache sequential inputs in memory slots, and explicitly utilize even far early information. Memory networks have been popularly applied as state-of-the-art approaches to many QA tasks, such as bAbI task [125], SQuAD [126] and LSMDC [127].

Memory networks were first applied to solve video QA problems in [15]. Compared with previous models, the fusion is not directly conducted between $\mathbf{v_Q}$ and $\mathbf{v_I}$. MemN2N uses an external memory to store sufficient amount of sequential information for answer generation. Similar to the attention mechanisms, it is usually associated with a similarity function to determine which memory slots are more important. Let $\mathbf{m_l}$ represent the feature for the l-th video frame stored in memory network. Its attention weight is achieved in conjunction with $\mathbf{v_Q}$ via a softmax function:

$$\boldsymbol{\alpha}_\mathbf{l} = \text{softmax}(\mathbf{v_Q}^T \mathbf{m_l}) \tag{25}$$

In [8], Zeng et al. leveraged a large volume of video descriptions from external sources to enhance the learning of video question answering. They contributed a new dataset and examined four types of basic fusion models, including basic vector operators, memory network and seq2seq models. Nevertheless, their experimental results show that their implementation of MemN2N does not achieve superior performance.

[103] presents a more powerful memory network which allows to read and write sequential memory cells as chunks. In other words, it can flexibly read and write more complex and abstract information into memory. Its write network based on CNN is able to capture high level information such as scene which may cover multiple adjacent frames. The attention vector is calculated in a similar way to MemN2N, i.e., by applying the softmax to the dot product between the question embedding and each cell of memory. In [104], the video stories, including visual scenes and textual dialogues, are stored in the long-term memory component. To answer a question, the best-matching stories are selected by a similarity scoring function. [99] raises a layered memory network to capture hierarchical representation in the frame level and clip level. The regional features in video frames are attended via inner product with the embeddings of words in a vocabulary that are stored in a static word memory. Furthermore, a dynamic subtitle memory module is put forward to represent movie clips with movie subtitles. With the features stored in the dynamic memory, the answer generation procedure is the same as MemN2N.

In [96], two visual channels, motion and appearance, are used and a co-memory network based on DMN (Dynamic Memory Network) [128] is proposed. DMN contains an episodic memory module to encode the input sources multiple cycles and an attention mechanism to focus on different contents in each cycle. The information of motion and appearance are encoded with separate memories and attended with questions and facts. Their output are fused by concatenation and sent to the answer generation module. [76] jointly embeds knowledge RDF triples and deep visual features into visual knowledge features, which are stored in a key-value memory network. The query representation $\mathbf{q}$ is a joint embedding between visual knowledge feature and question encoding. As to key addressing and value reading in the memory network, the model addresses each candidate memory slot and assigns it with a relevance probability. The values of memory slots are then read by weight averaging for answer generation.

8. Discussions and conclusions

In this survey, we have focused on the fusion strategies proposed to solve VQA problems. We proposed an abstract framework to organize the models proposed for both image QA and video QA. The model has two major components that involves the interaction between visual and textual channels: one is attention mechanisms and the other is to obtain the joint representation of (attended) visual and question features. We examine the existing VQA models and emphasize the proposed techniques w.r.t. these two fusion components. In this way, we can provide a relatively clear picture to readers and help them understand which fusion strategies are more commonly adopted. We also collect the experimental results on the benchmark datasets in the Appendix, which allow readers to quickly compare the performances of models on the same dataset, even though the improvement may not be always caused by the fusion techniques.

Since video QA is still an emerging and challenging research direction in VQA, we believe it will continue to attract more and more attention in the near future. As to the information channels, the audio embedded in the video has not been exploited by existing works. It would be interesting to examine if the fusion of audio information can be helpful for question answering. A better pre-processing framework to resolve the issues of varied video formats, lengths and resolutions could also benefit the research community. For example, the resolution of surveillance videos may not be as high as the pictures used in image QA, which may require tailored approaches for visual feature extraction. The datasets in video QA also tend to be diversified. A large and popular benchmark dataset similar to VQA 1.0/VQA 2.0 is desirable for video QA. Last but not the least, the computation issues of video QA can also be an issue worth investigation.

In the end of the survey, we present several useful guidelines that may be helpful for practitioners in VQA model design. In [79], Teney et al. have summarized a number of useful and detailed implementation tips and tricks for visual question answering, such as using gated tanh activations in all non-linear layers, using soft scores as ground truth targets and using large mini-batches and smart shuffling during stochastic gradient descent. In the following, we summarize several VQA model design guidelines from the perspective of effective fusion.

1. Fine-grained and semantic feature extraction can help improve the performance. The models using local visual features often perform better than those with only global visual features as it is difficult for them to conduct fine-grained inference. It is also suggested in [79] to use region-specific features for visual objects rather than grid-like feature maps from a CNN.
2. Augmenting the training dataset with external data sources can further boost the accuracy. In [76], external knowledge bases from DBpedia [129] and ConceptNet [130] are leveraged to extract tuples relevant to the question-answer pairs. With the augmented information, the model achieves state-of-the-art performance in VQA 1.0 dataset. It was also empirically evaluated in [79] that the use of additional questions/answers from Visual Genome [131] increased the performance on VQA 2.0 in all question types.
3. Attention mechanisms have been shown effective in capturing the interaction between visual and textual channels. Models without attention mechanisms often demonstrate much inferior performance. Multi-hop attention can further achieves higher accuracy compared with single-hop attention, but require more tuning efforts. The best performance often occurs when the number of hops is set to a small value such as 2.
4. Overall, bilinear models are expected to achieve promising results and they have received significant attention in recent publications. This is because they can fully capture the interaction between visual features and question features. A possible research direction in VQA is to integrate more effective attention mechanisms with bilinear fusion strategies.
5. Image QA has been intensively studied and it is challenging to establish new state-of-the-art in the datasets of VQA 1.0 and VQA 2.0. Possible research directions include focusing on sub-tasks of VQA. For example, multi-choice problems were particularly handled in [114]. Video QA has been an emerging research direction and still has great room for improvement. It can be expected that this research area will receive significant attention in the near future.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant no. 61602087, 61661146001 and 61872315).

Appendix

Table 8
Performance of models on VQA 1.0.

Model	Open-ended test-std	Open-ended test-dev	Multiple test-std	Multiple test-dev
NMN+LSTM [42]	55.10	54.80	–	–
LSTM Q+I [13]	–	57.75	–	62.70
MLB+DA-NTN [64]	66.89	–	–	–
MUTAN [61]	67.36	67.42	–	–
NMN+Caption Information [55]	–	57.10	–	–
ABC-CNN [43]	48.38	–	–	–
MCB [44]	66.50	66.70	70.10	70.20
QGHC [71]	65.90	65.89	–	–
VQA-AA [111]	65.90	–	–	69.80
CMF [72]	–	66.40	–	–
FDA [62]	59.54	59.24	64.18	64.01
MRN [45]	61.84	61.68	66.33	–
DMN [57]	60.40	60.30	–	–
QRU [47]	60.76	60.72	65.43	65.43
Dual-MFA [65]	66.09	66.01	69.97	70.04
DPPnet [49]	57.36	57.22	62.69	62.48
DRAU [66]	67.16	66.86	–	–
Word+Region Sel[52].	62.43	62.44	–	–
CVA [80]	66.20	65.92	70.41	70.30
MAN [56]	64.10	63.80	69.40	69.50
DCN [74]	67.02	66.89	–	–
VKMN [76]	66.10	66.00	69.10	69.10
MF+SIG+VG [77]	68.14	67.19	72.08	–
SMem-VQA [60]	58.24	57.99	–	–
SAN [54]	58.90	58.70	–	–
MLAN [68]	65.30	65.20	70.00	70.00
MFB [68]	66.60	66.90	71.40	71.30
MFH [70]	67.50	67.70	72.10	72.30
HieCoAtt [48]	62.10	61.80	66.10	65.80
High-Order [51]	–	–	69.30	69.40
Attribute CNN+LSTM [46]	59.50	–	–	–

Table 9
Performance of models on VQA2.0.

Model	Accuracy on val	Accuracy on test-std	Accuracy on test-dev
Up-Down [41]	63.20	70.34	–
VQA-AA [2]	60.80	–	–
BAN [82]	–	70.35	70.04
DRAU [66]	–	66.85	66.45
MAN [56]	–	62.10	–
DCN [74]	62.94	66.97	66.87
VKMN [76]	–	–	66.67
MFH [70]	–	68.02	–
Attention-on-Attention [78]	64.78	–	–
Tips-Tricks [79]	63.37	–	–

Table 10
Performance of models on Visual7W.

Model	Accuracy
MCB [44]	62.2
CVA [80]	63.8
MAN [56]	62.8
MLAN [68]	62.4
LSTM+Att [12]	55.6

Table 11
Performance of models on COCO.

Model	Accuracy	WUPS 0.9	WUPS 0.0
ABC-CNN [43]	58.10	68.44	89.85
QRU [47]	62.50	72.58	91.62
Dual-MFA [65]	66.49	76.15	92.29
CNN-QA [88]	58.40	68.50	89.67
DPPnet [49]	61.19	70.84	90.61
2-VIS+BLSTM [50]	55.09	67.90	89.52
CVA [80]	67.51	76.70	92.41
SAN [54]	61.60	71.60	90.90
HieCoAtt [48]	65.40	75.10	92.00
Attribute CNN+LSTM [46]	70.98	78.35	92.87

Table 12
Performance of models on CLEVR.

Model	Accuracy
OD+LSTM+RN [84]	94.50
QGHC [71]	86.30
CNN+LSTM+RN [85]	95.50
MF+SIG+VG [77]	78.04

References

[1] C.L. Zitnick, A. Agrawal, S. Antol, M. Mitchell, D. Batra, D. Parikh, Measuring machine intelligence through visual question answering, AI Mag. 37 (1) (2016) 63–72.
[2] I. Ilievski, J. Feng, Generative attention model with adversarial self-learning for visual question answering, in: MM, 2017, pp. 415–423.
[3] Y. Wang, D. Zhang, Y. Liu, B. Dai, L.H. Lee, Enhancing transportation systems via deep learning: a survey, Transp. Res. Part C 99 (2019) 144–163, doi:10.1016/j.trc.2018.12.004.
[4] Y. Kawano, K. Yanai, Foodcam-256: a large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights, in: MM, 2014, pp. 761–762.
[5] L. Wang, D. Zhang, L. Gao, J. Song, L. Guo, H.T. Shen, Mathdqn: solving arithmetic word problems via deep reinforcement learning, in: AAAI, AAAI Press, 2018, pp. 5545–5552.
[6] D. Zhang, L. Wang, N. Xu, B.T. Dai, H.T. Shen, The gap of semantic parsing: a survey on automatic math word problem solvers, CoRR abs/1808.07290 (2018).
[7] J.P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R.C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, T. Yeh, Vizwiz: nearly real-time answers to visual questions, in: UIST, ACM, 2010, pp. 333–342.
[8] K. Zeng, T. Chen, C. Chuang, Y. Liao, J.C. Niebles, M. Sun, Leveraging video descriptions to learn video question answering, in: AAAI, 2017, pp. 4334–4340.
[9] Y. Ye, Z. Zhao, Y. Li, L. Chen, J. Xiao, Y. Zhuang, Video question answering via attribute-augmented attention network learning, in: SIGIR, 2017, pp. 829–832.
[10] M. Heilman, N.A. Smith, Good question! statistical ranking for question generation, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, in: HLT '10, 2010, pp. 609–617.
[11] M. Ren, R. Kiros, R.S. Zemel, Image question answering: a visual semantic embedding model and a new dataset, CoRR (2015).
[12] Y. Zhu, O. Groth, M.S. Bernstein, L. Fei-Fei, Visual7w: Grounded question answering in images, in: CVPR, 2016, pp. 4995–5004.
[13] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, VQA: visual question answering, in: ICCV, 2015, pp. 2425–2433.
[14] 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, in: CVPR, IEEE Computer Society, 2017, pp. 6325–6334.
[15] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, S. Fidler, Movieqa: understanding stories in movies through question-answering, in: CVPR, 2016, pp. 4631–4640.
[16] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R.J. Mooney, T. Darrell, K. Saenko, Youtube2text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition, in: ICCV, 2013, pp. 2712–2719.
[17] J. Xu, T. Mei, T. Yao, Y. Rui, MSR-VTT: a large video description dataset for bridging video and language, in: CVPR, 2016, pp. 5288–5296.
[18] Y. Li, Y. Song, L. Cao, J.R. Tetreault, L. Goldberg, A. Jaimes, J. Luo, TGIF: a new dataset and benchmark on animated GIF description, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 4641–4650.
[19] J. Mun, P.H. Seo, I. Jung, B. Han, Marioqa: answering questions by watching gameplay videos, in: ICCV, 2017, pp. 2886–2894.

[20] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, D. Tao, Activitynet-qa: a dataset for understanding complex web videos via question answering, in: AAAI, 2019.
[21] Q. Wu, D. Teney, P. Wang, S. Chunhua, A.R. Dick, A. van den Hengel, Visual question answering: a survey of methods and datasets, Comput. Vision Image Underst. 163 (2017) 21–40.
[22] K. Kushal, C. Kanan, Visual question answering: datasets, algorithms, and future challenges, Comput. Vision Image Underst. 163 (2017) 3–20.
[23] A.K. Gupta, Survey of visual question answering: datasets and techniques, CoRR abs/1705.03865(2017).
[24] P. Supriya, S. Shagun, Survey of recent advances in visual question answering, CoRR abs/1709.08203(2017).
[25] Y. Jang, Y. Song, Y. Yu, Y. Kim, G. Kim, TGIF-QA: toward spatio-temporal reasoning in visual question answering, in: CVPR, 2017, pp. 1359–1367.
[26] J. Lei, L. Yu, M. Bansal, T.L. Berg, TVQA: localized, compositional video question answering, CoRR abs/1809.01696 (2018).
[27] W. Wang, B.C. Ooi, X. Yang, D. Zhang, Y. Zhuang, Effective multi-modal retrieval based on stacked auto-encoders, PVLDB 7 (8) (2014) 649–660.
[28] W. Wang, X. Yang, B.C. Ooi, D. Zhang, Y. Zhuang, Effective deep learning-based multi-modal retrieval, VLDB J. 25 (1) (2016) 79–101.
[29] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: MM, 2017, pp. 154–162.
[30] J.S.Y. Guo, L. Gao, X. Li, A. Hanjalic, H.T. Shen, From deterministic to generative: multimodal stochastic rnns for video captioning, IEEE Trans. Neural Netw. Learn. Syst. (2018), doi:10.1109/TNNLS.2018.2851077.
[31] J. Song, L. Gao, Z. Guo, W. Liu, D. Zhang, H.T. Shen, Hierarchical LSTM with adjusted temporal attention for video captioning, in: IJCAI, ijcai.org, 2017, pp. 2737–2743.
[32] L. Gao, Z. Guo, H. Zhang, X. Xu, H.T. Shen, Video captioning with attention-based LSTM and semantic consistency, IEEE Trans. Multimedia 19 (9) (2017) 2045–2055.
[33] Y. Pan, T. Yao, H. Li, T. Mei, Video captioning with transferred semantic attributes, in: CVPR, IEEE Computer Society, 2017, pp. 984–992.
[34] 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, Neural Comput. (1989) 541–551.
[35] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM (2017) 84–90.
[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: CVPR, 2015, pp. 1–9.
[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR (2014).
[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[39] X. Li, S. Chen, X. Hu, J. Yang, Understanding the disharmony between dropout and batch normalization by variance shift, CoRR (2018).
[40] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: NIPS, 2015, pp. 91–99.
[41] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and VQA, CVPR (2018) 4995–5004.
[42] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Deep compositional question answering with neural module networks, CoRR (2015).
[43] 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).
[44] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: EMNLP, 2016, pp. 457–468.
[45] 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, in: NIPS, 2016, pp. 361–369.
[46] 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 (2016).
[47] R. Li, J. Jia, Visual question answering with question representation update (QRU), in: NIPS, 2016, pp. 4655–4663.
[48] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: NIPS, 2016, pp. 289–297.
[49] H. Noh, P.H. Seo, B. Han, Image question answering using convolutional neural network with dynamic parameter prediction, in: CVPR, 2016, pp. 30–38.
[50] M. Ren, R. Kiros, R.S. Zemel, Exploring models and data for image question answering, in: NIPS, 2015, pp. 2953–2961.
[51] I. Schwartz, A.G. Schwing, T. Hazan, High-order attention models for visual question answering, in: NIPS, 2017, pp. 3667–3677.
[52] K.J. Shih, S. Singh, D. Hoiem, Where to look: Focus regions for visual question answering, in: CVPR, 2016, pp. 4613–4621.
[53] P. Wang, Q. Wu, C. Shen, A.R. Dick, A. van den Hengel, FVQA: fact-based visual question answering, IEEE Trans. Pattern Anal. Mach. Intell. (2018) 2413–2427.
[54] Z. Yang, X. He, J. Gao, L. Deng, A.J. Smola, Stacked attention networks for image question answering, in: CVPR, 2016, pp. 21–29.
[55] 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).
[56] 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, in: CVPR, IEEE Computer Society, 2018, pp. 6975–6984.
[57] C. Xiong, S. Merity, R. Socher, Dynamic memory networks for visual and textual question answering, in: ICML, 2016, pp. 2397–2406.
[58] 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, in: NIPS, 2015, pp. 2296–2304.
[59] M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons: a neural-based approach to answering questions about images, in: ICCV, 2015, pp. 1–9.
[60] H. Xu, K. Saenko, Ask, attend and answer: exploring question-guided spatial attention for visual question answering, in: ECCV, 2016, pp. 451–466.
[61] H. Ben-younes, R. Cadène, M. Cord, N. Thome, MUTAN: multimodal tucker fusion for visual question answering, in: ICCV, 2017, pp. 2631–2639.
[62] I. Ilievski, S. Yan, J. Feng, A focused dynamic attention model for visual question answering, CoRR (2016).
[63] V. Kazemi, A. Elqursh, Show, ask, attend, and answer: a strong baseline for visual question answering, CoRR (2017).
[64] J.-H. Kim, K.W. On, W. Lim, J. Kim, J. Ha, B.-T. Zhang, Hadamard product for low-rank bilinear pooling, CoRR (2016).
[65] 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, in: AAAI, 2018.
[66] A. Osman, W. Samek, Dual recurrent attention units for visual question answering, CoRR (2018).
[67] P. Wang, Q. Wu, C. Shen, A.R. Dick, A. van den Hengel, Explicit knowledge-based reasoning for visual question answering, in: IJCAI, 2017, pp. 1290–1296.
[68] D. Yu, J. Fu, T. Mei, Y. Rui, Multi-level attention networks for visual question answering, in: CVPR, 2017a, pp. 4187–4195.
[69] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: ICCV, 2017b, pp. 1839–1848.
[70] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering, IEEE Trans. Neural Netw. Learn. Syst. 29 (12) (2018) 5947–5959.
[71] P. Gao, P. Lu, H. Li, S. Li, Y. Li, S. Hoi, X. Wang, Question-guided hybrid convolution for visual question answering, CoRR abs/1808.02632 (2018).
[72] 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.
[73] J. Liang, L. Jiang, L. Cao, L. Li, A.G. Hauptmann, Focal visual-text attention for visual question answering, in: CVPR, IEEE Computer Society, 2018, pp. 6135–6143.
[74] D.-K. Nguyen, T. Okatani, Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, in: CVPR, 2018, pp. 6087–6096.
[75] Y. Shi, T. Furlanello, S. Zha, A. Anandkumar, Question type guided attention in visual question answering, in: ECCV, 2018, pp. 158–175.
[76] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, J. Li, Learning visual knowledge memory networks for visual question answering, in: CVPR, IEEE Computer Society, 2018, pp. 7736–7745.
[77] C. Zhu, Y. Zhao, S. Huang, K. Tu, Y. Ma, Structured attentions for visual question answering, in: ICCV, 2017, pp. 1300–1309.
[78] J. Singh, V. Ying, A. Nutkiewicz, Attention on attention: architectures for visual question answering VQA, CoRR (2018).
[79] D. Teney, P. Anderson, X. He, A. van den Hengel, Tips and tricks for visual question answering: learnings from the 2017 challenge, in: CVPR, IEEE Computer Society, 2018, pp. 4223–4232.
[80] J. Song, P. Zeng, L. Gao, H.T. Shen, From pixels to objects: cubic visual attention for visual question answering, in: IJCAI, 2018, pp. 906–912.
[81] Y. Bai, J. Fu, T. Zhao, T. Mei, Deep attention neural tensor network for visual question answering, in: ECCV, 11216, Springer, 2018, pp. 21–37.
[82] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, CoRR (2018).
[83] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. (1997) 1735–1780.
[84] M.T. Desta, L. Chen, T. Kornuta, Object-based reasoning in VQA, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, 2018, pp. 1814–1823.
[85] A. Santoro, D. Raposo, D.G.T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: NIPS, 2017, pp. 4974–4983.
[86] 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, in: EMNLP, 2014, pp. 1724–1734.
[87] H. Xu, K. Saenko, Dual attention network for visual question answering, in: ECCV 2016 2nd Workshop on Storytelling with Images and Videos (VisStory), 2016.
[88] L. Ma, Z. Lu, H. Li, Learning to answer questions from image using convolutional neural network, in: AAAI, 2016, pp. 3567–3573.
[89] N. Ruwa, Q. Mao, L. Wang, J. Gou, M. Dong, Mood-aware visual question answering, Neurocomputing 330 (2019) 305–316.
[90] Y. Kim, Convolutional neural networks for sentence classification, in: EMNLP, ACL, 2014, pp. 1746–1751.
[91] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: EMNLP, 2015, pp. 1422–1432.
[92] L. Guo, D. Zhang, L. Wang, H. Wang, B. Cui, CRAN: a hybrid CNN-RNN attention-based model for text classification, in: ER, in: Lecture Notes in Computer Science, 11157, Springer, 2018, pp. 571–585.
[93] D. Tran, L.D. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: ICCV, 2015, pp. 4489–4497.

[94] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, Y. Zhuang, Video question answering via gradually refined attention over appearance and motion, in: MM, 2017, pp. 1645–1653.
[95] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Image Analysis, 13th Scandinavian Conference, SCIA 2003, 2003, pp. 363–370.
[96] J. Gao, R. Ge, K. Chen, R. Nevatia, Motion-appearance co-memory networks for video question answering, CoRR (2018).
[97] Z. Zhao, Q. Yang, D. Cai, X. He, Y. Zhuang, Video question answering via hierarchical spatio-temporal attention networks, in: IJCAI, 2017, pp. 3518–3524.
[98] H. Xue, Z. Zhao, D. Cai, Unifying the video and question attentions for open-ended video question answering, IEEE Trans. Image Processing (2017) 5656–5666.
[99] B. Wang, Y. Xu, Y. Han, R. Hong, Movie question answering: Remembering the textual cues for layered visual contents, in: AAAI, 2018.
[100] Z. Zhao, X. Jiang, D. Cai, J. Xiao, X. He, S. Pu, Multi-turn video question answering via multi-stream hierarchical attention context network, in: IJCAI, 2018a, pp. 3690–3696.
[101] Z. Zhao, Z. Zhang, S. Xiao, Z. Yu, J. Yu, D. Cai, F. Wu, Y. Zhuang, Open-ended long-form video question answering via adaptive hierarchical reinforced networks, in: IJCAI, 2018b, pp. 3683–3689.
[102] L. Zhu, Z. Xu, Y. Yang, A.G. Hauptmann, Uncovering temporal context for video question and answering, CoRR (2015).
[103] S. Na, S. Lee, J. Kim, G. Kim, A read-write memory network for movie story understanding, in: ICCV, 2017, pp. 677–685.
[104] K.-M. Kim, M.-O. Heo, S.-H. Choi, B.-T. Zhang, Deepstory: video story QA by deep embedded memory networks, in: IJCAI, 2017, pp. 2016–2022.
[105] Y. Yu, J. Kim, G. Kim, A joint sequence fusion model for video question answering and retrieval, in: ECCV, 2018, pp. 487–503.
[106] R. Kiros, Y. Zhu, R. Salakhutdinov, R.S. Zemel, R. Urtasun, A. Torralba, S. Fidler, Skip-thought vectors, in: NIPS, 2015, pp. 3294–3302.
[107] K. Kafle, C. Kanan, An analysis of visual question answering algorithms, in: ICCV, IEEE Computer Society, 2017, pp. 1983–1991.
[108] L. Jiang, J. Liang, L. Cao, Y. Kalantidis, S. Farfade, A.G. Hauptmann, Memexqa: visual memex question answering, CoRR abs/1708.01336 (2017).
[109] Y. LeCun, L. Bottou, G.B. Orr, K. Müller, Efficient backprop, in: Neural Networks: Tricks of the Trade (2nd ed.), in: Lecture Notes in Computer Science, 7700, Springer, 2012, pp. 9–48.
[110] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: NIPS, 2013, pp. 3111–3119.
[111] 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.
[112] B.N. Patro, V.P. Namboodiri, Differential attention for visual question answering, in: CVPR, IEEE Computer Society, 2018, pp. 7680–7688.
[113] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: ICCV, 2015, pp. 1026–1034.
[114] 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, pp. 1742–1750.
[115] H. Nam, J.-W. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, in: CVPR, 2017, pp. 2156–2164.
[116] R. Kiros, R. Salakhutdinov, R.S. Zemel, Multimodal neural language models, in: ICML, 2014, pp. 595–603.
[117] K.-M. Kim, S.-H. Choi, J.-H. Kim, B.-T. Zhang, Multimodal dual attention memory for video story question answering, in: ECCV, 2018, pp. 698–713.
[118] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T.P. Lillicrap, Meta-learning with memory-augmented neural networks, in: ICML, 2016, pp. 1842–1850.
[119] J.B. Tenenbaum, W.T. Freeman, Separating style and content with bilinear models, Neural Comput (2000) 1247–1283.
[120] M. Charikar, K.C. Chen, M. Farach-Colton, Finding frequent items in data streams, in: ICALP, 2002, pp. 693–703.
[121] N. Pham, R. Pagh, Fast and scalable polynomial kernels via explicit feature maps, in: SIGKDD, 2013, pp. 239–247.
[122] H. Pirsiavash, D. Ramanan, C.C. Fowlkes, Bilinear classifiers for visual recognition, in: NIPS, 2009, pp. 1482–1490.
[123] A. Karpathy, F.-F. Li, Deep visual-semantic alignments for generating image descriptions, in: CVPR, 2015, pp. 3128–3137.
[124] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, End-to-end memory networks, in: NIPS, 2015, pp. 2440–2448.
[125] J. Weston, A. Bordes, S. Chopra, T. Mikolov, Towards ai-complete question answering: a set of prerequisite toy tasks, CoRR abs/1502.05698 (2015).
[126] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100, 000+ questions for machine comprehension of text, in: EMNLP, 2016, pp. 2383–2392.
[127] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C.J. Pal, H. Larochelle, A.C. Courville, B. Schiele, Movie description, Int. J. Comput. Vis. 123 (1) (2017) 94–120.
[128] 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, in: ICML, 2016, pp. 1378–1387.
[129] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z.G. Ives, Dbpedia: a nucleus for a web of open data, in: ISWC 2007 + ASWC 2007, in: Lecture Notes in Computer Science, 4825, Springer, 2007, pp. 722–735.
[130] H. Liu, P. Singh, Conceptnet — a practical commonsense reasoning tool-kit, BT Technol. J. 22 (4) (2004) 211–226, doi:10.1023/B:BTTJ.0000047600.45421.6d.
[131] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73.