Computers, Materials & Continua
DOI: 10.32604/cmc.2022.027097
Article

Tech Science Press

Triple Multimodal Cyclic Fusion and Self-Adaptive Balancing for Video Q&A Systems

Xiliang Zhang[1], Jin Liu[1,*], Yue Li[1], Zhongdai Wu[2,3] and Y. Ken Wang[4]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai, China
[2]Shanghai Ship and Shipping Research Institute, Shanghai, China
[3]COSCO Shipping Technology Co., LTD, Shanghai, China
[4]Division of Management and Education, University of Pittsburgh, Bradford, USA
*Corresponding Author: Jin Liu. Email: jinliu@shmtu.edu.cn
Received: 10 January 2022; Accepted: 14 April 2022

Abstract: Performance of Video Question and Answer (VQA) systems relies on capturing key information of both visual images and natural language in the context to generate relevant questions' answers. However, traditional linear combinations of multimodal features focus only on shallow feature interactions, fall far short of the need of deep feature fusion. Attention mechanisms were used to perform deep fusion, but most of them can only process weight assignment of single-modal information, leading to attention imbalance for different modalities. To address above problems, we propose a novel VQA model based on Triple Multimodal feature Cyclic Fusion (TMCF) and Self-Adaptive Multimodal Balancing Mechanism (SAMB). Our model is designed to enhance complex feature interactions among multimodal features with cross-modal information balancing. In addition, TMCF and SAMB can be used as an extensible plug-in for exploring new feature combinations in the visual image domain. Extensive experiments were conducted on MSVD-QA and MSRVTT-QA datasets. The results confirm the advantages of our approach in handling multimodal tasks. Besides, we also provide analyses for ablation studies to verify the effectiveness of each proposed component.

Keywords: Video question and answer systems; feature fusion; scaling matrix; attention mechanism

1 Introduction

Cross-modal information interaction scenarios, for instance, images with captions or videos with subtitles, can often convey richer information than ones with single model information. Nevertheless, since visual content and verbal content contain information with certain modal differences, it is very difficult to well exploit them in applications. Thus, research topics on cross-modal feature interaction such as Video Question and Answer (VQA) have become increasingly important in multi-media information processing.

VQA is a relatively new task in which a video and a natural language question are provided. Correspondingly, the model needs to give the right answer based on the multimodal information. Traditional approach directly fuses the global features of an image with question text, and then classifies them to get a predicted answer. However, these algorithms simply perform a linear fusion of multimodal features such as summation or cascade operations, do not pay attention to the "balancing" problem between modal information. To improve performance of the model on multimodal tasks, Zadeh et al. [1] proposed Multimodal Tensor Fusion Network (TNF), which fuses the features of multiple modalities by utilizing simple matrix operations. In addition, the tensor outer product is used to represent the correlation between the modalities. Low-rank Multimodal Fusion method (LMF), proposed by Liu et al. [2], performs low-rank matrix decomposition of weight parameters with excessive dimensionality based on the TNF method. Although the fusion operation of these early methods is simple in terms of computation, the dimensionality after fusion is not controllable. Especially in the case of excessively long features, it not only tends to cause parameter explosion during transforming, but also makes it difficult to obtain a model.

The development of attention mechanisms has not only driven advances in natural language processing and computer vision, but is also improving multimodal processing tasks such as video question-and-answer. The sequential video attention and temporal question attention models are proposed by Xue et al. [3]. And these two models apply attention mechanism on videos and questions, while preserving the sequential and temporal structures of the guides. Kim et al. [4] proposed a bimodal attentional memory mechanism applied to the question-and-answer structure of video stories, using the proposed questions to provide secondary attention to potential concepts. However, most of the studies focus only on improvements in attention mechanisms, still lack strong inference in VQA scenarios where feature information needs to be deeply understood.

In this paper, to solve the above problems, we propose a novel VQA model based on triple multimodal feature cyclic fusion and self-adaptive multimodal balancing mechanism. This model can fuse three feature information without generating high-dimensional vectors. In the meantime, it can solve the cross-modal information balancing problem. Specifically, we introduce a scaling matrix to dynamically adjust the weight coefficients between the refined multimodal feature vectors through a scaling transformation, then obtain the corresponding features in equilibrium. Information in three modalities is fed into a feature fusion module to obtain the fused features after interaction. Finally, these three features are fed into an answer generation module to make the prediction. Experimental results on the MSVD-QA and MSRVTT-QA datasets demonstrate the strong robustness of our proposed method in deep inference. To summarize, the main contributions of this work are as follows:

(1) We design a cyclic feature fusion module that can be applied to three different modal data for enhancing complex feature interactions among multimodal features in VQA systems.
(2) A scaling matrix is introduced to dynamically adjust the weight coefficients between features by scaling the refined multimodal feature vector through a scaling transformation.
(3) Our method achieved state-of-the-art performance on two real datasets.

2 Related Work

2.1 Video Question and Answer

Video Question and Answer (VQA) is a multi-disciplinary artificial intelligence research topic that has gained interest of mainly two communities, computer vision and Natural Language Processing (NLP). Significant progress has been made in many vision-language tasks, including image-text matching [5], visual captioning [6], visual grounding [7] and VQA [5,8]. Most existing VQA approaches make use of the relationship between visual and language features. Bilinear feature fusion approaches [9] focus on capturing the higher order relations between language and visual through outer product of features. Co-attention or bilinear attention-based approaches [4] learn the inter-modality relations between word-region pairs to identify key pairs for question answering. Besides, some computer vision and natural language processing algorithms focusing on learning intra-modality relations. Liu et al. [10] proposed a novel end-to-end multi-level semantic representation enhancement network (MLSREN) that can extract deeper phrasal semantic information using fewer network parameters. Hu et al. [11] proposed to explore intra-modality object-to-object relations to boost object detection accuracy. Zang et al. [12] proposed a Bayesian inference-based pervasive semantic knowledge mining algorithm for improving the efficiency of knowledge discovery. Yao et al. [13] modeled intra-modality object-to-object relations for improving image captioning performance. In recently proposed Bidirectional Encoder Representations from Transformers (BERT) algorithm [14] for natural language processing, word relations within modalities can be learned through a self-attentive mechanism. However, VQA is more challenging than other multimodal tasks. It requires not only an accurate understanding of the semantics of images and questions, but also an effective reasoning to get correct answer in the form of natural language [15]. Chen et al. [16] proposed a Multimodal Encoder-Decoder Attention Networks (MEDAN). This network can capture rich and reasonable question features and image features by associating keywords in question with important object regions in image. Processing text features and video features in parallel by different models in two major directions is a reasonable solution. Zhang et al. [17] proposed a new word vector training method based on parts of speech (POS) features for distinguishing the same words under different discourses and improving the quality of problematic text translation. Zhang et al. [18] proposed a motion-blurred image restoration method based on joint invertibility of Point Spread Functions (PSFs) to solve the iterative restoration problem by the joint solution of multiple images in spatial domain. Li et al. [19] use a pre-trained captioner to generate general captions and attributes with a fixed annotator, then use them to predict answers. Common problem of these approaches is not adequately combining question text features with video features.

2.2 Multimodal Fusion

Feature fusion method is the most important module in a VQA system. Lu et al. [20] focused on extracting image and problem information features to reduce the variability between different modal data by combining important information from them. The Multimodal Compact Bilinear (MCB) fusion method proposed by Fukui A et al. [21] speeds up model training process and reduces the complexity by compressing the parameters of the fusion matrix. In addition, fusion methods such as Multimodal Factorized Bilinear (MFB) [22] and Multimodal Residual Learning (MRN) [23] have also emerged to drive development of VQA systems. Kim et al. [24] introduced additional supervised information in a selective VQA task and proposed a multi-task ratio adjustment method to prioritize the learning tasks. Jiang et al. [25] made feasible optimizations to the input video stream, save additional overhead for subsequent model extraction while ensuring the reliability and security of its transmission. Zhao et al. [26] proposed a spatio-temporal attention network based on video keyframes to learn joint representation, which achieves progressive joint representation learning and improves the performance of open VQA. Gong et al. [27] proposed a deep model Knowledge Context Fusion Network (KCF-NET), which encodes the intrinsic spatial relationships of knowledge graph (KG) triples by using a capsule network, using the knowledge graph representation with context as the basis for predicting answers. Xu et al. [28] proposed an end-to-end open VQA model, and allocates attention rationally by progressively refining the attention to appearance features and motion features. Gao et al. [29] proposed a motion-appearance joint memory network, which used a temporal convolution-deconvolution architecture, built multi-level contextual information, and computes attention jointly with motion-appearance features to improve the recognition accuracy of a selective-open VQA model. However, these VQA models still have shortcomings, such as high feature dimensionality and insufficient feature fusion after multimodal fusion.

2.3 Attention Mechanism

Instead of directly using the entire image embedding from the fully connected layer of a deep convolution neural network (CNN), attention-based models have been widely used to select the most relevant image regions in VQA systems. The attention mechanism [30] usually consists of a linguistic parser and CNN feature vectors representing spatially distributed regions. Liu et al. [31] proposed a novel deep neural network model called Attention-Based BiGRU-CNN network (ABBC) to extract the features of Chinese questions effectively and learnt the context information of words. A novel multi-headed attention mapping network was proposed by Zhang et al. [32] to extract deeper overall relationships. Yang et al. [17] performed image noticing multiple times in a stacked fashion and gradually inferred the answers. Xu et al. [33] used a multi-hop image attention mechanism to capture fine-grained information from the question text. Chang et al. [34] proposed a multi-lane capsule network with strict-squash method for solving the problem of failing to extract corresponding features due to severe background interference in visual images. Shih et al. [35] proposed a method that maps textual queries and visual features from different regions into a shared space and learns to answer visual questions by selecting image regions that relevant to text-based queries. Xiong et al. [36] proposed an attention-based Gate Recurrent Unit (GRU) to help answer retrieval. A new multi-scale convolutional model based on multiple attentions was proposed by Yang et al. [37], which introduced an attention mechanism into the structure of Res2-block to better guide the feature representation. In addition to visual attention, recent work [20] has proposed a mechanism for simultaneous attention problems. In summary, attention mechanism is recently used to assign weights to different important information. Similarly, we balance the weights of different features by constructing a scaling transformation matrix, which is mainly used to solve the balancing problem between cross-modal features.

2.4 Self-Attention-Based Methods

Attention mechanisms was proposed to mimic the way of human vision. By automatically ignoring irrelevant information, the neural network can selectively focus on important features. The self-attention mechanism is a variant of the attention mechanism that is less dependent on external information and better at capturing the internal relevance of data or features. Liu et al. [38] proposed a method that combines a self-attentive mechanism with Bi-directional Long Short-Term Memory (Bi-LSTM) to improve the performance of the parser on long texts, alleviating the problem brought by unknown words. Relational networks [11] learn the relationship between object suggestions by employing self-attentive mechanisms. The in-place module can boost Faster Region-Based Convolutional Networks (RCNN) [39] and Non-Maximum-Suppression (NMS) performance. Unlike these methods, our proposed SAMB mechanism not only enhances long-range information, but also adjusts cross-modal feature imbalance weights by matrix scaling transformation.

3 Method

3.1 Problem Formalization

In the multimodal fusion task of the VQA system, video V and questions $Q = [q_1, q_2, \ldots, q_U]$ are used as inputs, where U is the number of question words. Video V always contains many redundant frames per second, which need to be sampled evenly to be a compressed representation of the entire video. After the feature extraction layer, number of frames are chosen to be compatible with the feature extractor and we can get the appearance features $F_a = \left[f_1^a, f_2^a, \ldots, f_N^a\right]$, where N is the number of frames, and the motion features $F_m = \left[f_1^m, f_2^m, \ldots, f_C^m\right]$, where C is the number of extracted video clips, respectively. The overall architecture of out VQA model is shown in Fig. 1.

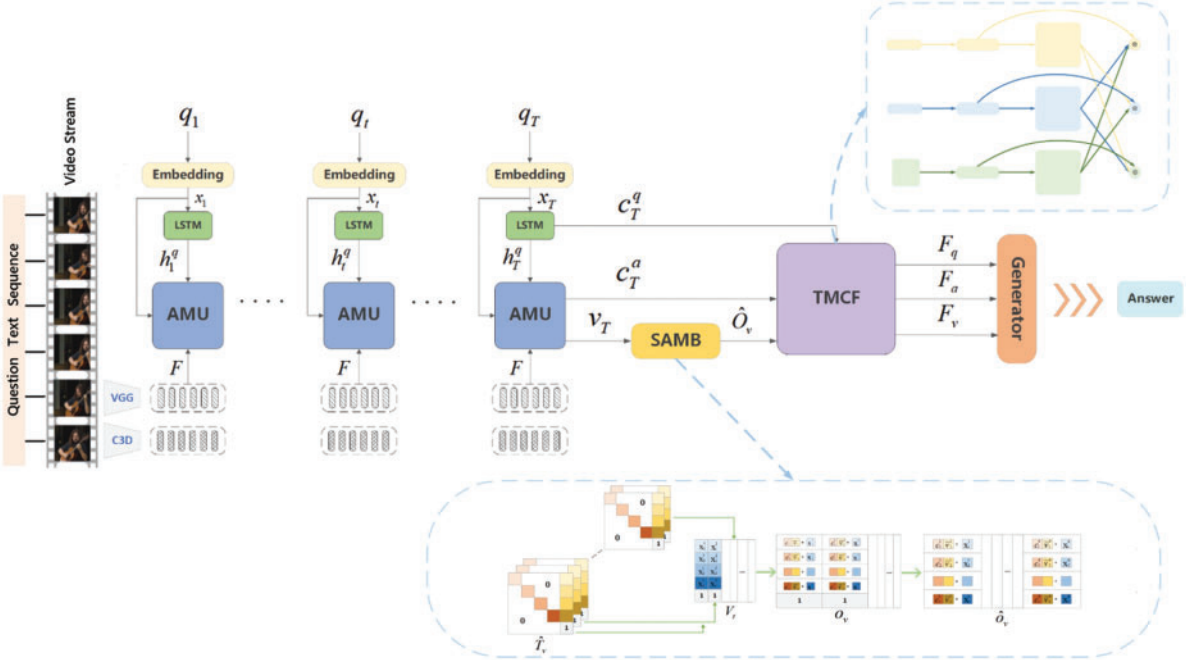


Figure 1: The overall structure of video question and answer system based on triple multimodal feature cycle fusion and self-adaptive balancing mechanism**.** The video stream V and the question text sequence Q are used as inputs to the model, and the output is the predicted answer. Specifically, Visual Geometry Group Network (VGG) and Convolutional 3D Network (C3D) act on video stream V to extract the appearance features F_a and motion features F_m, respectively. The semantic information x_t and hidden layer state h_t^q are obtained from the question text sequence Q by word2vec embedding and Long Short-Term Memory Network (LSTM) transformation, respectively. Then they are fused with video feature information $F = \{F_a, F_m\}$ and sent to Attention Memory Unit (AMU) module to obtain three feature vectors (c_T^q, c_T^a,v_T) after refinement. The v_T is adjusted by our SAMB module to obtain the characteristic $\hat{O}_v$ in the equilibrium mode. Next, $\hat{O}_v$, c_T^a and c_T^q are fed into our TMCF module to obtain the feature vectors F_q, F_a and $F_{\hat{o}}$ after interaction. Finally, we get predicted answers in Generator

3.2 Feature Extraction and Refinement Layer

Video frames contain still images of different objects that occupy a large portion of the information in the video. Thanks to its well-designed network structure, the Visual Geometry Group Network (VGG) [40] is widely used in various image-related tasks, especially in capturing visual features. Naturally, we use it as a frame-level appearance F_a. Except for static objects, motion features in video are another channel of information that distinguish video from image. The Convolutional 3D Network (C3D) [41] is often used for motion recognition tasks and has excellent ability to capture video motion information. Thus, we use it as a clip-level motion feature extractor for extracting motion features F_m. For problem feature extracting, we use word2vec [42] embedding technique to capture high-dimensional semantics of words and get the semantic information x_t after word transformation at present moment t.

Some redundant or irrelevant features usually are extracted in generated description. To reduce the impact of these features, the appearance features F_a, motion features F_m and semantic information x_t are input into Attention Memory Unit (AMU) [28], which generates an intermediate feature representation v_t of the video at present moment t after each time step. v_t contains the attention weight values generated by the appearance and motion features at present moment for the current word. After processing by the AMU module, three types of modal information are obtained after the last time step$T$, which are the problem Long Short-Term Memory (LSTM) [43] vector c_T^q, the memory vector c_T^a of the AMU module, and the intermediate result of visual features v_T respectively.

3.3 Self-Adaption Multimodal Balance Mechanism

To solve the problem that ordinary attention mechanism cannot focus on the multimodal information balance, we propose a Self-Adaption Multimodal Balance Mechanism. Specifically, the weights between features are dynamically adjusted by introducing a scaling matrix to achieve a balance between multimodal information.

First, the intermediate result of visual features $v_T = \left[v_1^t, v_2^t, \ldots, v_d^t\right]$, $v_T \in \mathbb{R}^{c\times d}$, where each entry $v_n^t = \left[\hat{v}_1^n, \hat{v}_2^n, \ldots, \hat{v}_c^n\right]^T$, $n \in \{1, 2, \ldots, d\}$. We add an element with value 1 after a vector v_n^t to obtain the transformed vector $\hat{v}_n^t$, as shown in Eq. (1).

$$\hat{v}_n^t = \left[v_n^t, 1\right]^T = \left[\hat{v}_1^n, \hat{v}_2^n, \ldots, \hat{v}_c^n, 1\right]^T \tag{1}$$

Among them, $\hat{v}_n^t \in \mathbb{R}^{(c+1)\times 1}$ is the n-th transformed vector.

Second, we construct scaling matrix T_v for the scaling transformation of the eigenvectors. Scaling matrix T_v is constructed in a unique way where diagonal elements and last column elements are non-zero. In addition, the element at last position in the lower right corner of matrix is a constant, i.e., an identity with element 1. Specifically, the constructed scaling matrix $T_v \in \mathbb{R}^{(c+1)\times(c+1)}$ consists of $c + 1$ vectors $t_n^v \in \mathbb{R}^{(c+1)}$, the last column vector is represented by $X_v \in \mathbb{R}^{1\times(c+1)}$, the equations for vectors t_v and X_v are shown as below.

$$t_n^v = \left[C_1^n, C_2^n, \ldots, C_c^n, X_n\right] \tag{2}$$

$$C_i^n, = \begin{cases} C_i^n, i = n \\ 0, i \neq n \end{cases}, i \in \{1, \ldots, c\} \tag{3}$$

$$X_v = [X_1, X_2, \ldots, X_c, 1]^T \tag{4}$$

Among them, $C_i^n$and X_i are parameters that need to be pre-trained to adjust the parameters of the elements in the feature vector $\hat{v}_n^t$, and $X_n, n \in \{1, 2, \ldots, c\}$ is the n-th element in vector X_v. When $i = n$, variable C_i^n has values that are not 0; when $i \neq n$, variable $C_i^n = 0$.

The scaling matrix T_v is then multiplied with the transformed eigenvector $\hat{v}_n^t$ to obtain the scaled and adjusted eigenvector $o_n^v \in \mathbb{R}^{(c+1)}$. The specific calculation formula is as follows:

$$\begin{aligned} o_n^v &= T_v \cdot \hat{v}_n^t \\ &= \left[t_1^v \cdot \hat{v}_n^t,\ \ t_2^v \cdot \hat{v}_n^t, \ldots,\ \ t_{c+1}^v \cdot \hat{v}_n^t\right]^T \\ &= \begin{bmatrix} C_1^n \cdot \tilde{v}_1^n + X_1 \\ C_2^n \cdot \tilde{v}_2^n + X_2 \\ \vdots \\ C_c^n \cdot \tilde{v}_c^n + X_c \\ 1 \end{bmatrix} \end{aligned} \tag{5}$$

The elements in vector t_n^v are dotted with each element in vector $\hat{v}_n^t$. Result of the operation is a polynomial form $Ax + B$, and the element in the last row is fixed to be 1.

Finally, each vector in the visual intermediate feature v_T is sequentially transformed to do the scaling matrix transformation. Here we construct a scaled feature matrix $\hat{T}_v$ of dimension $(c + 1) \times (c + 1) \times d$ for the overall scaling transformation of v_T. The computation procedure of the transformed feature matrix $O_v \in \mathbb{R}^{(c+1)\times d}$ is shown in Eq. (6).

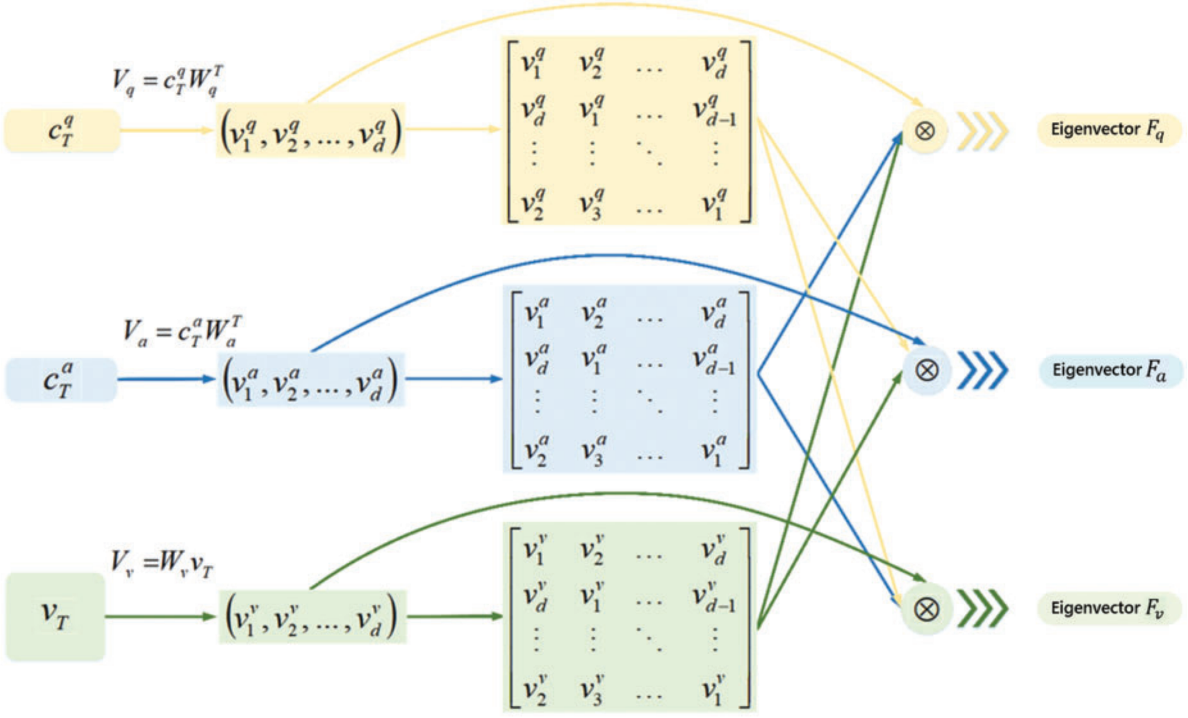
$$\begin{aligned} O_v &= \left[o_1^v, o_2^v, \ldots, o_d^v\right] \\ &= \left[T_1^v \cdot \hat{v}_1^t,\ T_2^v \cdot \hat{v}_2^t, \ldots, T_d^v \cdot \hat{v}_d^t\right]^T \\ &= \begin{bmatrix} C_1^1 \cdot \tilde{v}_1^1 + X_1^1 & C_1^2 \cdot \tilde{v}_1^2 + X_1^2 & \ldots & C_1^2 \cdot \tilde{v}_1^2 + X_1^2 & \ldots \\ C_2^1 \cdot \tilde{v}_2^1 + X_2^1 & C_2^2 \cdot \tilde{v}_2^2 + X_2^2 & \ldots & C_2^2 \cdot \tilde{v}_2^2 + X_2^2 & \ldots \\ \vdots & \vdots & & \vdots & \\ C_c^1 \cdot \tilde{v}_c^1 + X_c^1 & C_c^2 \cdot \tilde{v}_c^2 + X_c^2 & \ldots & C_c^2 \cdot \tilde{v}_c^2 + X_c^2 & \ldots \\ 1 & 1 & & 1 & \end{bmatrix} \end{aligned} \tag{6}$$

Among them, $T_n^v, n \in \{1, 2, \ldots, d\}$is the feature matrix of the n-th layer in $\hat{T}_v$. We remove all the vectors with element values of 1 in the last row of the scaled transformed feature matrixO_v, so that the dimensionality of the feature matrix $\hat{O}_v \in \mathbb{R}^{c\times d}$ remains the same as the original middle visual feature matrix $v_T \in \mathbb{R}^{c\times d}$. The scaling transformation of v_T is completed to achieve dynamic balance between multimodal features. The feature matrix $\hat{O}_v$ is shown in Eq. (7).

$$\hat{O}_v = \begin{bmatrix} C_1^1 \cdot \tilde{v}_1^1 + X_1^1 & C_1^2 \cdot \tilde{v}_1^2 + X_1^2 & \dots & C_1^d \cdot \tilde{v}_1^d + X_1^d \\ C_2^1 \cdot \tilde{v}_2^1 + X_2^1 & C_2^2 \cdot \tilde{v}_2^2 + X_2^2 & \dots & C_2^d \cdot \tilde{v}_2^d + X_2^d \\ \vdots & \vdots & & \vdots \\ C_c^1 \cdot \tilde{v}_c^1 + X_c^1 & C_c^2 \cdot \tilde{v}_c^2 + X_c^2 & \dots & C_c^d \cdot \tilde{v}_c^d + X_c^d \end{bmatrix} \tag{7}$$

3.4 Triple Multimodal Cyclic Fusion Layer

The TMRF module explores all interactions between different modal vectors by reconstructing the original features into a cyclic matrix. The loop fusion involves some simple mathematical operations, avoiding the introduction of additional parameters that increase the computational cost. Flow of the method is shown in Fig. 2.



$V_q = c_T^q W_q^T$
c_T^q
$(v_1^q, v_2^q, \dots, v_d^q)$
$\begin{bmatrix} v_1^q & v_2^q & \dots & v_d^q \\ v_d^q & v_1^q & \dots & v_{d-1}^q \\ \vdots & \vdots & \ddots & \vdots \\ v_2^q & v_3^q & \dots & v_1^q \end{bmatrix}$
⊗
Eigenvector F_q
$V_a = c_T^a W_a^T$
c_T^a
$(v_1^a, v_2^a, \dots, v_d^a)$
$\begin{bmatrix} v_1^a & v_2^a & \dots & v_d^a \\ v_d^a & v_1^a & \dots & v_{d-1}^a \\ \vdots & \vdots & \ddots & \vdots \\ v_2^a & v_3^a & \dots & v_1^a \end{bmatrix}$
⊗
Eigenvector F_a
$V_v = W_v v_T$
v_T
$(v_1^v, v_2^v, \dots, v_d^v)$
$\begin{bmatrix} v_1^v & v_2^v & \dots & v_d^v \\ v_d^v & v_1^v & \dots & v_{d-1}^v \\ \vdots & \vdots & \ddots & \vdots \\ v_2^v & v_3^v & \dots & v_1^v \end{bmatrix}$
⊗
Eigenvector F_v

Figure 2: Flow chart of triple multi-modal cyclic fusion

After above adjustments, three kinds of modal information can be obtained, which are the problem LSTM memory vector $c_T^q \in \mathbb{R}^a$, the memory vector $c_T^a \in \mathbb{R}^b$ for AMU module, $\hat{O}_v \in \mathbb{R}^{c\times d}$. Naturally, we project three features into same low-dimensional space, here with $\hat{O}_v$ as the reference for the projection, and the equation is as follows:

$$V_q = c_T^q W_q^T \tag{8}$$

$$V_a = c_T^a W_a^T \tag{9}$$

$$V_{\hat{o}} = W_{\hat{o}} v_T \tag{10}$$

Among them, superscript T of matrix $W_i,\ i \in \{1,2,3\}$ represents the transpose operation. The matrices of three modes are $W_q \in \mathbb{R}^{d\times a}$, $W_a \in \mathbb{R}^{d\times b}$ and$W_{\hat{o}} \in \mathbb{R}^{c}$, and mapped feature vectors are $V_q \in \mathbb{R}^d$, $V_a \in \mathbb{R}^d$ and $V_{\hat{o}} \in \mathbb{R}^d$, respectively. At this point, feature vectors V_q, V_a and $V_{\hat{o}}$ are all in the same dimension, for which different circular matrices are constructed separately.

The initial eigenvector is used as the first layer of the cyclic matrix. To obtain matrix vector of the $(m+i)$ row, we shift vector element of the m row by i unit distance in positive direction. By this circular shift, the $m*n$-dimensional cyclic matrix can be calculated. In our experiments, setting $i=1$ was found to produce good results. The constructed circular matrix is shown in the following equation.

$$C_q = Ciro\left(V_q\right) \tag{11}$$

$$C_a = Ciro\left(V_a\right) \tag{12}$$

$$C_{\hat{o}} = Ciro\left(V_{\hat{o}}\right) \tag{13}$$

Among them, $C_q \in \mathbb{R}^{d\times d}$, $C_a \in \mathbb{R}^{d\times d}$ and $C_{\hat{o}} \in \mathbb{R}^{d\times d}$ are cyclic matrices of three modal characteristics, and $Ciro\left(\cdot\right)$ is the cyclic matrix balance function. Then elemental products of the different cyclic matrices are made separately to explore full interaction between three modal information, as follows Eqs. (14)–(16)

$$F_q = \frac{1}{d}\sum\nolimits_{i=1}^{d} f_i^q \odot C_a \odot C_{\hat{o}} \tag{14}$$

$$F_a = \frac{1}{d}\sum\nolimits_{i=1}^{d} f_i^a \odot C_q \odot C_{\hat{o}} \tag{15}$$

$$F_{\hat{o}} = \frac{1}{d}\sum\nolimits_{i=1}^{d} f_i^v \odot C_q \odot C_a \tag{16}$$

where f_i is the i row eigenvector of cyclic matrix corresponding to the modal features and $\odot$ denotes dot product between element A and $B.F_q \in \mathbb{R}^d$, $F_a \in \mathbb{R}^d$ and $F_{\hat{o}} \in \mathbb{R}^d$ are eigenvectors after the intersection of three modalities, respectively. Finally, the *softmax* classifier is used to generate answers after interaction of three different modal feature vectors.

$$answer = softmax\left(W_s F_q \cdot F_a \cdot F_{\hat{o}}\right) \tag{17}$$

where W_s represents the learning weight.

4 Experiment

4.1 Datasets and Experimental Setup

We choose two public datasets in this field for evaluating the performance of model. One of them is the MSVD-QA [44], a corpus of video description collected and made publicly available by Microsoft, which is often used for video captioning experiments and mainly oriented towards real scenes of slightly shorter duration. The dataset contains about 1,970 video clips and 50,505 question-answer pairs. We split it into 61% training set, 13% validation set and 26% test set. The statistics of MSVD-QA datasets are shown in Tab. 1.

Table 1: Statistics of the MSVD-QA dataset for offline evaluation

Video Question and Answer pairs			Question Type				
			What	*Who*	*How*	*When*	*Where*
Train	1,200	30,933	19,485	10,479	736	161	72
Val	250	6,415	3,995	2,168	185	51	16
Test	520	12,157	8,149	4,552	370	58	28
All	1,970	50,505	31,629	17,199	1,291	270	116

The other is MSRVTT-QA [45] based on the evolution of MSRVTT dataset, which is larger, more complex, and primarily oriented towards real scenarios of somewhat longer duration. And it contains 10,000 video clips and 243,000 question-answer pairs. We split it into 65% training set, 5% validation set and 30% test set. Statistics of MSRVTT-QA datasets are shown in Tab. 2.

Table 2: Statistics of the MSRVTT-QA dataset for offline evaluation

Video Question and Answer pairs			Question Type				
			What	*Who*	*How*	*When*	*Where*
Train	6,513	158,581	108,792	43,592	4,067	1,626	504
Val	497	12,278	8,337	3,439	344	106	52
Test	2,990	72,821	49,869	20,385	1,640	677	250
All	10,000	243,680	166,998	67,416	6,051	2,409	806

4.2 Implementation Details

We use mini-batch stochastic gradient descent to optimize the model with a default learning rate of 0.001 for an Adaptive Moment Estimation (Adam) [46] optimizer. Our training batch size is 64. The model is trained for up to 30 epochs and can be stopped early. To handle questions of different lengths efficiently, we divide questions into five buckets based on the length of questions in each of two datasets. In each bucket, the questions are populated to the length of the longest question in the bucket. Implementation of experiments is based on the server with open-source framework Tensorflow-gpu2.4.0 [47], keras2.4.3 and python3.7, CUDA11.2. The GPU used for our experiments is NVIDIA GeForce RTX 3090 with 128GB memory.

4.3 Experimental Results

The performance of TMCF and six benchmark models are represented in Tab. 3. We can observe that TMCF performs better than all baselines on evaluation criteria. In the MSVD-QA dataset, our model achieved the highest accuracy in "*what*" and "*how*" question categories, although the "*who*" type questions did not achieve the best results, they were improved compared to traditional models.

Table 3: Performance comparison of different models on MSVD-QA dataset

Methods	*what*	*who*	*how*	*when*	*where*	ALL
E-VQA [44]	0.097	0.422	0.838	0.724	**0.536**	0.233
E-SA [44]	0.150	0.451	0.838	0.655	0.322	0.276
E-MN [44]	0.129	0.465	0.803	0.707	0.500	0.267
ST-VQA [48]	0.181	**0.500**	0.838	0.724	0.286	0.313
Co-Mem [29]	0.196	0.487	0.835	**0.741**	0.317	0.317
DLAN [49]	0.212	0.460	0.832	0.724	0.500	0.318
Ours	**0.216**	0.477	**0.840**	0.724	0.536	**0.322**

Tab. 4 summarizes the performance of TMCF on the MSRVTT-QA dataset compared to six benchmarks. In the MSRVTT-QA dataset, our model obtained the best results compared with other models in question types "*what*" and "*who*", and similar results to baseline were achieved on other question types. Fig. 3 visualizes loss curve of our model during training on the MSRVTT-QA dataset.

Table 4: Performance comparison between ours and other six models on MSRVTT-QA dataset

Methods	*what*	*who*	*how*	*when*	*where*	ALL
E-VQA	0.189	0.387	0.835	0.705	0.292	0.264
E-SA	0.220	0.416	0.796	0.731	0.332	0.293
E-M	0.234	0.418	**0.837**	0.708	0.276	0.304
ST-VQA	0.245	0.412	0.780	**0.765**	0.349	0.309
Co-Mem	0.239	0.425	0.741	0.690	**0.429**	0.320
DLAN	0.254	0.428	0.810	0.721	0.312	0.320
Ours	**0.265**	**0.435**	0.796	0.725	0.292	**0.328**

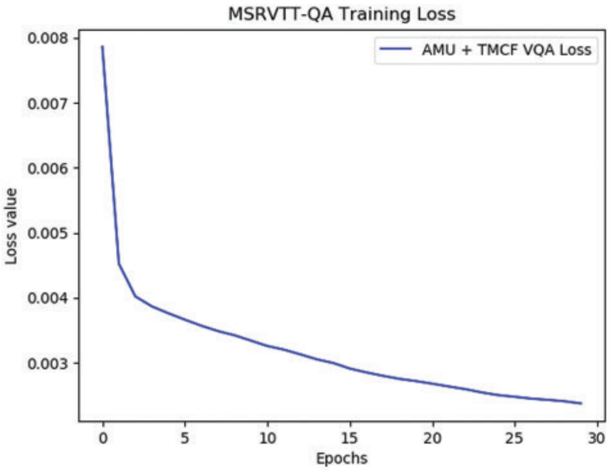


Figure 3: Loss function curve of our model on MSRVTT-QA training set

The graph of the loss curve variation shows that our model converges successfully after 30 rounds of training, which validates the fact that we can avoid introducing additional parameters to influence training process by using the TMCF module. Fig. 4 visualizes the performance of our proposed model on the MSRVTT-QA validation set.

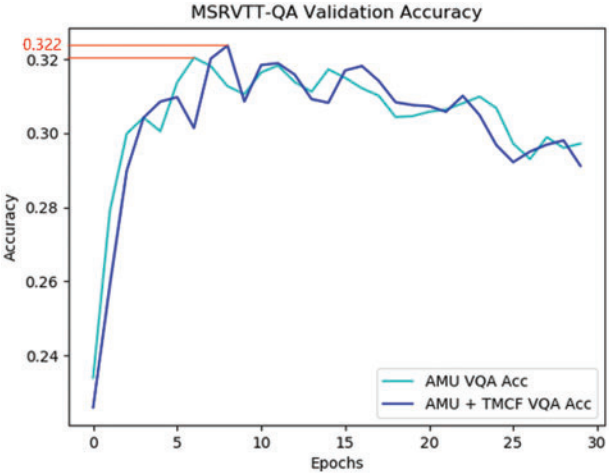


Figure 4: Accuracy comparison of models fusing different networks on MSRVTT-QA validation set

As can be seen in Fig. 4, the accuracy curve of our model on the MSRVTT-QA validation set is oscillating with an overall decreasing trend, which is due to the overfitting in training process. We set following rule to avoid the risk of overfitting without being able to adjust our model structure: if the accuracy on validation set of the N-th training round is decreasing for 10 consecutive rounds, then the result of the N-th training round is taken as our final. Meanwhile, the highest accuracy of VQA model incorporating TMCF in validation set is higher than AMU model alone, which also proves the effectiveness of our proposed TMCF module.

Thus, our proposed model has higher accuracy in answering "*what*" and "*who*" types of questions compared to other baseline models in both MSVD-QA and MSRVTT-QA datasets. The accuracy of our model is low for certain types of problems (e.g., where), mainly because the dataset for this type of problem is too small and our model tends to be overfitted when performing training. Specifically, large number of identical appearance features and motion features are repeatedly fused, resulting in a model that loses focus on capturing features, decreases generalization ability, and fails to show good performance in answering other questions. To further investigate the role of SAMB mechanism, we use the MSRVTT-QA dataset for our experiments here and compare results with some mainstream VQA models to demonstrate the effectiveness of our model. The experimental accuracy results on the MSRVTT-QA dataset are shown in Tab. 5.

As can be seen from Tab. 5, VQA model fusing TMCF and SAMB achieves the best overall accuracy on the most dominant question types "*what*" and "*who*", where overall accuracy reaches 33.2%, proving the excellent performance of our proposed model. In other question types, ours was beaten by other models, but also achieved similar results as baselines. This is because the amount of data corresponding to other types of problems is too small for our model to learn cross-modal information well. Nevertheless, by adopting an overfitting avoidance strategy, we can effectively improve the prediction accuracy rate of the model. In addition, Fig. 5 was created to provide a visual representation of the responses to each type of question.

Table 5: Performance comparison between ours and other eight models on MSRVTT-QA dataset

Methods	*what*	*who*	*how*	*when*	*where*	ALL
E-VQA	0.189	0.387	0.835	0.705	0.292	0.264
E-SA	0.220	0.416	0.796	0.731	0.332	0.293
E-MN	0.234	0.418	**0.837**	0.708	0.276	0.304
ST-VQA	0.245	0.412	0.780	**0.765**	0.349	0.309
Co-Mem	0.239	0.425	0.741	0.690	**0.429**	0.320
DLAN	0.254	0.428	0.810	0.721	0.312	0.320
AMU [28]	0.262	0.430	0.802	0.725	0.300	0.325
TMCF	0.265	0.435	0.796	0.725	0.292	0.328
TMCF + SAMB	**0.268**	**0.439**	0.780	0.728	0.292	**0.332**

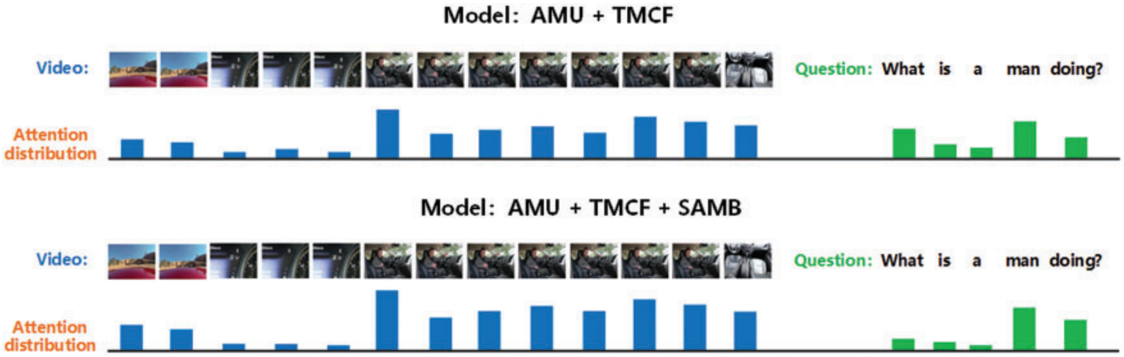


Figure 5: Attention allocation during question answering in a VQA system incorporating different modules

Fig. 5 shows the attention distribution in answering questions for AMU model based on TMCF, and the one based on fusion of TMCF and SAMB. As can be observed from the figure, the inclusion of SAMB mechanism makes the model pay less attention to irrelevant information in video, and the attention to important video information features gains a substantial increase. Specifically, attentional weight of key words is highlighted by SAMB in the face of questions, such as *man* and *doing*, while reducing the attentional allocation of other irrelevant information. Rational allocation of attentional resources is the reason why our SAMB shows strong robustness in balancing multimodal information.

4.4 Ablation Analysis

We provide a detailed ablation study of our VQA model to investigate the relative contribution to the performance improvement. Different design choices were tested, regarding diversity of the backbone network for feature extraction, number of TMCF fusion features. The experimental results are shown in Tab. 6.

Table 6: Effect of different parameters of our model under ablation experiments

Model	Appearance Extraction Network	Motion Extraction Network	Number of TMCF Fusion Features	Accuracy
Baseline	VGG	C3D	-	0.325
Model 1	CNN	C3D	-	0.287
Model 2	VGG	LSTM	-	0.292
Model 3	CNN	LSTM	-	0.281
Model 4	CNN	C3D	2	0.290
Model 5	CNN	C3D	3	0.296
Model 6	VGG	LSTM	2	0.295
Model 7	VGG	LSTM	3	0.300
Model 8	CNN	LSTM	2	0.303
Model 9	CNN	LSTM	3	0.310
Model 10	VGG	C3D	2	0.326
Ours	VGG	C3D	3	**0.328**

We performed extensive comparison experiments using different appearance and motion feature extraction networks, and judged the merits of each feature extraction model choice by the accuracy in answering questions. The original AMU equipped with VGG appearance feature extraction network and C3D motion feature extraction network is our baseline model. We found that using CNN and LSTM as appearance and motion feature extraction networks would be worse than VGG and C3D networks compared Model 1-Model 3 respectively. Specifically, using the VGG network to extract appearance features leads to 1.1% accuracy increase in model's answer to the question, while using the C3D network as our model for motion feature extraction results in a 0.6% accuracy uptick. It is demonstrated that advanced feature extraction networks can improve the performance of VQA system to some extent. In addition, comparing Model 1 with Model 2, it is found that changes in appearance feature extraction network have a greater impact on the performance of VQA model. This is because most of the questions in the dataset are simple, i.e., they do not involve a logical reasoning part, and the machine relies more on static appearance features when answering simple questions. Besides, comparing baseline with other models shows that fusing intermediate features, then feeding them into answer generation module is more effective than feeding them directly, while three intermediate features incorporated achieves the best performance ratio improvement for VQA system.

To further investigate the impact of balance condition between multimodal information on VQA system and the role of SAMB mechanism, ablation experiments are designed on MSRVTT-QA dataset. The original AMU as the baseline model, feature fusion method and multimodal balance method are taken as variables, as shown in Tab. 7.

Table 7: Effect of different parameters of our model under ablation experiments

Models	Feature Fusion	Balance Factor	Self-Adaption Multimodal Balance	Accuracy
Baseline	ADD	-	-	0.325
Model 1	MCF	-	-	0.326
Model 2	TMCF	-	-	0.328
Model 3	TMCF	2	-	0.330
Model 4	TMCF	1/2	-	0.321
Model 5	TMCF	-	√	**0.332**

Among them, feature fusion methods are simple accumulation, Multimodal cyclic fusion (MCF) method and our TMCF fusion method. The balance factor parameter represents whether it is used to adjust the characteristic information of a mode. A balance factor of 2 means that the value of video feature is twice the original, and 1/2 means that the value of video feature is 1/2 the original. Self-Adaption Multimodal Balance field represents whether our proposed SAMB mechanism is used in model. As can be seen from the table, TMCF fusion method can better intermingle the multimodal information in VQA system, and enhance the machine's ability to understand multimodal information compared with ADD and MCF methods. In addition, the model tuned by using the balanced factor method is clearly more advantageous compared to the baseline. However, performance of model with balance factor taking the value of 1/2 is degraded, and we speculate that the accuracy of model is affected by the weight of key feature information in video. Finally, we use SAMB mechanism to find the optimal balance factor dynamically and leverage TMCF fusion method for multimodal feature fusion, so that the optimal performance can be obtained by our model.

5 Conclusion

In this paper, we propose a novel VQA model based on triple multimodal feature cyclic fusion and self-adaptive balancing mechanism. This model is not only applicable for balancing multimodal feature information, but also can be used to enhance complex feature interactions among multimodal features in VQA systems. In contrast to the most widely used bottom-up and top-down models, our model is better-designed for VQA tasks, not only is architecturally lightweight, but also enables full interaction of cross-modal features. Experimental results show that our model can substantially uplift the state of the art (SOTA) results on video question and answer tasks, especially when dealing with complex questions. Moreover, our ablation analysis corroborate that the enhancement is mainly attributed to our design choices in terms of diversity of feature fusion methods, modulation of balance factors, and adaptive multimodal balance mechanisms.

Acknowledgement: Xiliang Zhang: Conceptualization, Methodology, Writing-original draft, Software. Jin Liu: Supervision, Project administration, Formal analysis, Data curation. Yue Li: Visualization, Software, Resources. Zhongdai Wu: Resources, Validation. Y. Ken Wang: Data curation.

Funding Statement: This work was supported by the National Natural Science Foundation of China (No. 61872231), the National Key Research and Development Program of China (No. 2021YFC2801 000), and the Major Research plan of the National Social Science Foundation of China (No. 20&ZD130).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Zadeh, M. H. Chen, S. Poria, E. Cambria and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *2017 Conf. on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 1103–1114, 2017.

[2] Z. Liu, Y. Shen, V. La, P. P. Liang, A. Zadeh *et al.,* "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, vol. 1, pp. 2247–2256, 2018.

[3] H. Y. Xue, Z. Zhao and D. Cai, "Unifying the video and question attentions for open-ended video question answering," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5656–5666, 2017.

[4] K. M. Kim, S. H. Choi, J. H. Kim and B. T. Zhang, "Multi-modal dual attention memory for video story question answering," in *Proc. European Conf. on Computer Vision*, Munich, Germany, pp. 673–688, 2018.

[5] J. H. Kim, J. Jun and B. T. Zhang, "Bilinear attention networks," in *32nd Conf. on Neural Information Processing Systems*, Montréal, Canada, pp. 1571–1581, 2018.

[6] P. Anderson, X. D. He, C. Buehler, D. Teney, M. Johnson *et al.,* "Bottom-up and top-down attention for image captioning and visual question answering," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6077–6086, 2018.

[7] Z. Yu, J. Yu, C. C. Xiang, Z. Zhao, Q. Tian *et al.,* "Rethinking diversified and discriminative proposal generation for visual grounding," in *Proc, IJCAI*, Stockholm, Sweden, pp. 1114–1120, 2018.

[8] Z. Zhao, Z. Zhang, S. Xiao, Z. Yu, J. Yu *et al.,* "Open-ended long-form video question answering via adaptive hierarchical reinforced networks," in *Proc. IJCAI*, Stockholm, Sweden, pp. 3683–3689, 2018.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Proc. NIPS*, Long Beach, CA, USA, pp. 5998–6008, 2017.

[10] J. Liu, Y. H. Yang and H. H. He, "Multi-level semantic representation enhancement network for relationship extraction," *Neurocomputing*, vol. 403, pp. 282–293, 2020.

[11] H. Hu, J. Y. Gu, Z. Zhang, J. F. Dai and Y. C. Wei, "Relation networks for object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, vol. 2, pp. 3588–3597, 2018.

[12] Y. Zang, T. Hu, and T. Zhou, "An automated penetration semantic knowledge mining algorithm based on Bayesian inference," *Computers, Materials & Continua*, vol. 66, no. 3, pp. 2573–2585, 2021.

[13] T. Yao, Y. W. Pan, Y. H. Li and T. Mei, "Exploring visual relationship for image captioning," in *Proc. European Conf. on Computer Vision*, Munich, Germany, pp. 684–699, 2018.

[14] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. NAACL-HLT*, Minneapolis, Minnesota, pp. 4171–4186, 2019.

[15] C. Deng, G. Zeng, Z. Cai and X. Xiao, "A survey of knowledge-based question answering with deep learning," *Journal on Artificial Intelligence*, vol. 2, no. 4, pp. 157–166, 2020.

[16] C. Chen, D. Z. Han and J. Wang, "Multimodal encoder-decoder attention networks for visual question answering," *IEEE Access*, vol. 8, pp. 35662–35671, 2020.

[17] J. Y. M. Zhang, J. Liu and X. Y. Lin, "Improve neural machine translation by building word vector with part of speech," *Journal on Artificial Intelligence*, vol. 2, no. 2, pp. 79–88, 2020.

[18] Y. Zhang, J. Huang, J. Liu and H. A. Chohan, "Motion-blurred image restoration based on joint invertibility of psfs," *Computer Systems Science and Engineering*, vol. 36, no. 2, pp. 407–416, 2021.

[19] Q. Li, J. Fu, D. Yu, T. Mei and J. Luo, “Tell-and-answer: Towards explainable visual question answering using attributes and captions,” in *Conf. on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 1338–1346, 2018.

[20] J. Lu, J. W. Yang, D. Batra and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” in *30th Int. Conf. on Neural Information Processing Systems*, Barcelona, Spain, pp. 289–297, 2016.

[21] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell *et al.,* “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in *Proc. Conf. on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, pp. 457–468, 2016.

[22] Z. Yu, J. Yu, J. P. Fan and D. Tao, “Multi-modal factorized bilinear pooling with co-attention learning for visual question answering,” in *IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 1821–1830, 2017.

[23] J. H. Kim, S. W. Lee, D. Kwak, M. O. Heo, J. Kim *et al.,* “Multimodal residual learning for visual qa,” in *30th Int. Conf. Neural Information Processing Systems*, Barcelona, Spain, pp. 361–369, 2016.

[24] J. Kim, M. Ma, K. Kim, S. Kim and C. D. Yoo, “Gaining extra supervision via multi-task learning for multi-modal video question answering,” in *Proc. IJCNN*, Budapest, Hungary, pp. 1–8, 2019.

[25] X. Jiang, F. R. Yu, T. Song and V. C. M. Leung, “Resource allocation of video streaming over vehicular networks: A survey, some research issues and challenges,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2021. http://dx.doi.org/10.1109/TITS.2021.3065209.

[26] Z. Zhao, Q. Yang, D. Cai, X. F. He and Y. Zhuang, “Video question answering via hierarchical spatio-temporal attention networks,” in *Proc. IJCAI*, Melbourne, Australia, pp. 3518–3524, 2017.

[27] P. Z. Gong, J. Liu, Y. H. Yang and H. H. He, “Towards knowledge enhanced language model for machine reading comprehension,” *IEEE Access*, vol. 8, pp. 224837–224851, 2020.

[28] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang *et al.,* “Video question answering via gradually refined attention over appearance and motion,” in *ACM Int. Conf. on Multimedia*, New York, NY, USA, pp. 1645–1653, 2017.

[29] J. Gao, R. Ge and K. Chen, “Motion-appearance co-memory networks for video question answering,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6576–6585, 2018.

[30] J. Andreas, M. Rohrbach, T. Darrell and D. Klein, “Neural module networks,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 39–48, 2016.

[31] J. Liu, Y. H. Yang, S. Q. Lv, J. Wang and H. Chen, “Attention-based BiGRU-CNN for Chinese question classification,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 2, pp. 1–12, 2019.

[32] B. Zhang, H. Ling, P. Li, Q. Wang, Y. Shi *et al.,* “Multi-head attention graph network for few shot learning,” *Computers, Materials & Continua*, vol. 68, no. 2, pp. 1505–1517, 2021.

[33] H. Xu and K. Saenko, “Ask, attend and answer: Exploring question-guided spatial attention for visual question answering,” in *European Conf. on Computer Vision*, Springer, Cham, pp. 451–466, 2016.

[34] S. Chang and J. Liu, “Multi-lane capsule network for classifying images with complex background,” *IEEE Access*, vol. 8, pp. 79876–79886, 2020.

[35] K. J. Shih, S. Singh and D. Hoiem, “Where to look: Focus regions for visual question answering,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 4613–4621, 2016.

[36] C. Xiong, S. Merity and R. Socher, “Dynamic memory networks for visual and textual question answering,” in *Proc. Int. Conf. on Machine Learning*, New York, USA, pp. 2397–2406, 2016.

[37] Y. Yang, C. Xu, F. Dong and X. Wang, “A new multi-scale convolutional model based on multiple attention for image classification,” *Applied Sciences*, vol. 10, no. 1, pp. 101, 2020.

[38] D. Liu, L. Zhang, Y. Shao and J. Sun, “Leverage external knowledge and self-attention for Chinese semantic dependency graph parsing,” *Intelligent Automation & Soft Computing*, vol. 28, no. 2, pp. 447–458, 2021.

[39] S. Ren, K. He, R. Girshick and J. Sun, “Faster rcnn: Towards real-time object detection with region proposal networks,” in *Proc. NIPS*, Montreal, Quebec, Canada, pp. 91–99, 2015.

[40] K. Simon and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. ICLR*, San Diego, CA, USA, pp. 1409–1556, 2015.

[41] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 4489–4497, 2015.

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proc. Int. Conf. on Neural Information Processing Systems*, Red Hook, NY, USA, vol. 2, pp. 3111–3119, 2013.

[43] Hochreiter Sepp and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[44] D. L. Chen and W. B. Dolan, “Collecting highly parallel data for paraphrase evaluation,” in *Proc. ACL*, Portland, OR, USA, pp. 190–200, 2011.

[45] J. Xu, T. Mei and T. Yao, “MSR-VTT: A large video description dataset for bridging video and language,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 5288–5296, 2016.

[46] D. P. Kingma and L. J. Ba, “Adam: A method for stochastic optimization,” in *Proc. ICLR*, San Diego, CA, USA, pp. 13, 2015.

[47] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis *et al.,* “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” in *12th USENIX Conf. on Operating Systems Design and Implementation*, Savannah, GA, USA, pp. 265–283, 2016.

[48] Y. Jang, Y. Song, Y. Yu, Y. Kim and G. Kim, “TGIF-QA: Toward spatio-temporal reasoning in visual question answering,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2758–2766, 2017.

[49] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan *et al.,* “Long-term recurrent convolutional networks for visual recognition and description,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 2625–2634, 2015.