Divide and Conquer: Question-Guided Spatio-Temporal Contextual Attention for Video Question Answering

Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao,* Yue Gao*
BNRist, KLISS, School of Software, Tsinghua University, China
jianwen.alan@gmail.com, {czq18, linhj18}@mails.tsinghua.edu.cn, {zxb, gaoyue}@tsinghua.edu.cn

Abstract

Understanding questions and finding clues for answers are the key for video question answering. Compared with image question answering, video question answering (Video QA) requires to find the clues accurately on both spatial and temporal dimension simultaneously, and thus is more challenging. However, the relationship between spatio-temporal information and question still has not been well utilized in most existing methods for Video QA. To tackle this problem, we propose a Question-Guided Spatio-Temporal Contextual Attention Network (QueST) method. In QueST, we divide the semantic features generated from question into two separate parts: the spatial part and the temporal part, respectively guiding the process of constructing the contextual attention on spatial and temporal dimension. Under the guidance of the corresponding contextual attention, visual features can be better exploited on both spatial and temporal dimensions. To evaluate the effectiveness of the proposed method, experiments are conducted on TGIF-QA dataset, MSRVTT-QA dataset and MSVD-QA dataset. Experimental results and comparisons with the state-of-the-art methods have shown that our method can achieve superior performance.

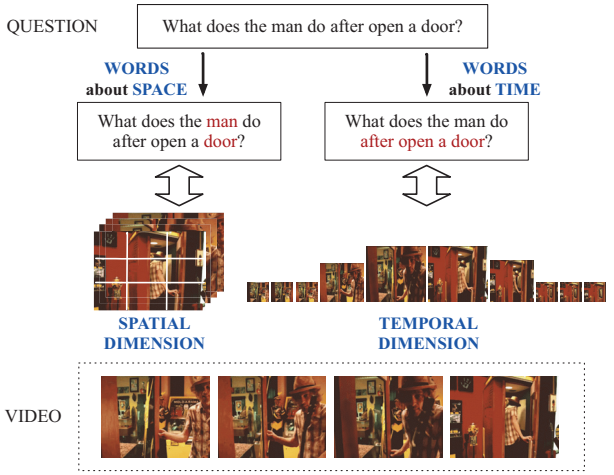

Figure 1: The key to Video QA is understanding question and finding clues of answer from both spatial and temporal dimensions in a video under guidance of question. For Video QA, information from video and question can both be divided into two dimensions, time and space.

Introduction

Recently, the visual question answering (VQA) task (Antol et al. 2015) has captured great attention due to the wide use in different areas, such as education, robot and intelligent assistant. The VQA task can be mainly divided into Image Question Answering (Image QA) and Video Question Answering (Video QA), targeting on answering a natural language question related to different visual material.

Generally, understanding the question and finding clues of the answer to the question in the given visual material is the key to VQA. For Image QA, there have been extensive efforts concentrated on (Yang et al. 2016; Xu and Saenko 2016; Wu et al. 2016; Kazemi and Elqursh 2017; Anderson et al. 2018; Gao et al. 2016; Fukui et al. 2016; Ben-Younes et al. 2017) in the last decade. For example, CBP (Gao et al. 2016), MCB (Fukui et al. 2016) and MUTAN (Ben-Younes et al. 2017) focus on finding better methods on fusing visual features with language features, which help the network to understand the question and visual material accurately. Attention mechanisms (Xu et al. 2015; Singh, Ying, and Nutkiewicz 2018; Anderson et al. 2018) have also been used to inform neural network "where are the clues of answer".

Compared to Image QA, Video QA is more challenging. In Image QA, most of the clues for the answer can be found from the spatial feature and the question is primarily based on the appearance feature of an image. In Video QA, however, a correct answer requires accurately locating the clues in not only spatial but also temporal dimension at the same time. Besides, there are more complex scene changes in video material, leading to high requirements on reasoning ability from both spatial dimension and temporal dimension. Most existing Video QA methods (Jang et al. 2017; Xu et al. 2017; Yu, Kim, and Kim 2018; Gao et al. 2018;

*Corresponding authors

Li et al. 2019; Fan et al. 2019) first generate a question embedding and then fuse it with video features to generate a joint representation for question answering. (Jang et al. 2017) utilize LSTM to encode question into single embedding for generating both spatial and temporal attention for video. (Gao et al. 2018) focus on jointly modeling motion and appearance information and building motion-appearance co-memory network. (Li et al. 2019) employ attention mechanism, instead of LSTM or GRU, to generate question embedding and video feature embedding.

Although Video QA methods have been investigated recently, it is still challenging. Question reasoning and locating answer clues in visual features are the key to VQA. For locating answer clues, different from images, videos have two dimensions in visual level, *i.e.*, space and time, leading to the inconsistency of the keypoint to the clues on both spatial and temporal dimension. As shown in Figure 1, given the question "What does the man do after open the door", the keypoints in spatial dimension are "man" and "door" while the keypoint in temporal dimension is "after open the door". Therefore, exploiting information in both spatial and temporal dimensions appropriately is significant in Video QA in order to excavate clues from videos. A group of existing methods (Gao et al. 2018; Li et al. 2019) focus on utilizing the temporal context information in video, but fail to exploit the feature on spatial dimension. Other methods (Jang et al. 2017) try to apply spatio-temporal attention on videos, but show even worse performance compared to temporal-only attention probably due to the lack of the guidance of question from both spatial and temporal dimensions. Under such circumstance, how to effectively exploit the spatio-temporal information in videos is important for Video QA.

To tackle the problems mentioned above, in this paper, we propose a Question-Guided Spatio-Temporal Contextual Attention Network (QueST) for Video QA, which divides the information of video and question into two separate parts: the spatial part and the temporal part. Then, it learns the relationship between video and question in each part.

In QueST, Video-Guided Question Attention (VGQA) encodes the question into two different question embeddings (spatial question embedding and temporal question embedding) first. Then, Question-Guided Contextual Attention Blocks (CABs) are introduced in the spatial and temporal dimension of the video sequentially to model context-aware visual features and excavate visual clues related to answer in the specific dimension of video under the guidance of corresponding question feature.

The contributions of this paper are as follows:

- We propose the Video-Guided Question Attention Block (VGQA), which introduces visual information to co-model the question information from both spatial and temporal dimensions.
- We introduce the Contextual Attention Block (CAB), which excavates the key information related to the answer in the context-aware visual feature. Based on CAB, Spatial CAB (SCAB) and Multi-Scale Temporal CAB (MS-TCAB) are designed for learning Video QA better by utilizing the interaction between visual features and corresponding question embedding on spatial and temporal dimensions respectively.
- Based on VGQA and CAB, we propose Question-Guided Spatio-Temporal Contextual Attention Network (QueST). We have conducted experiments on three Video QA datasets, *i.e.*, TGIF-QA dataset, MSRVTT-QA dataset and MSVD-QA dataset. The experimental results demonstrate the superior performance of our proposed QueST on the Video QA task.

Related Work

In this section, we briefly review some recent works related to Video QA. Given a question and an image or video regarding the question, Visual Question Answering is a task of providing accurate natural language answers. Multi-modal reasoning over visual and textual data is essential for solving VQA problem. For Image QA, most early approaches focus on fusing textual and visual information, which are extracted by a LSTM network and a CNN based model respectively. (Yang et al. 2016) proposes a Stacked Attention Networks(SAN) that queries an image multiple times to infer the answer progressively. (Fukui et al. 2016) presents Multimodal Compact Bilinear pooling(MCB) method to capture high-level interaction between visual and textual information. Dynamic Memory Network(DMN) (Kumar et al. 2016) has exhibited certain reasoning capabilities on many language tasks. (Xiong, Merity, and Socher 2016) applies DMN to image question answering by proposing improvements to the memory and input modules of DMN. To calculate attention at the levels of objects and other salient image regions, (Anderson et al. 2018) utilizes Faster R-CNN (Ren et al. 2015) features and designs a bottom-up and top-down attention mechanism.

Different from Image QA, Video QA methods (Jang et al. 2017; Xu et al. 2017; Gao et al. 2018; Li et al. 2019; Yu et al. 2019; Fan et al. 2019) introduce extra temporal reasoning modules to answer questions which involve changing on temporal dimension like action or state transition. (Jang et al. 2017) presents a dual-LSTM based approach with spatio-temporal contextual attention, which utilizes a sentence-level question embedding generated by LSTM to apply spatio-temporal attentions to video. (Xu et al. 2017) introduces an end-to-end model that gradually refines temporal attention on appearance and motion features via question guidance. (Gao et al. 2018) proposes a motion-appearance co-memory network that utilizes co-memory attention mechanism to capture temporal information both from motion and appearance. Compared with temporal information, spatial information are not well utilized in most recent methods. In (Jang et al. 2017), spatio-temporal attention model performs even worse than temporal attention only model. Therefore, both spatial and temporal information have not been sufficiently exploited in the existing Video QA methods.

Approach

In this section, we give a detailed introduction to the proposed Question-Guided Spatio-Temporal Contextual Atten-

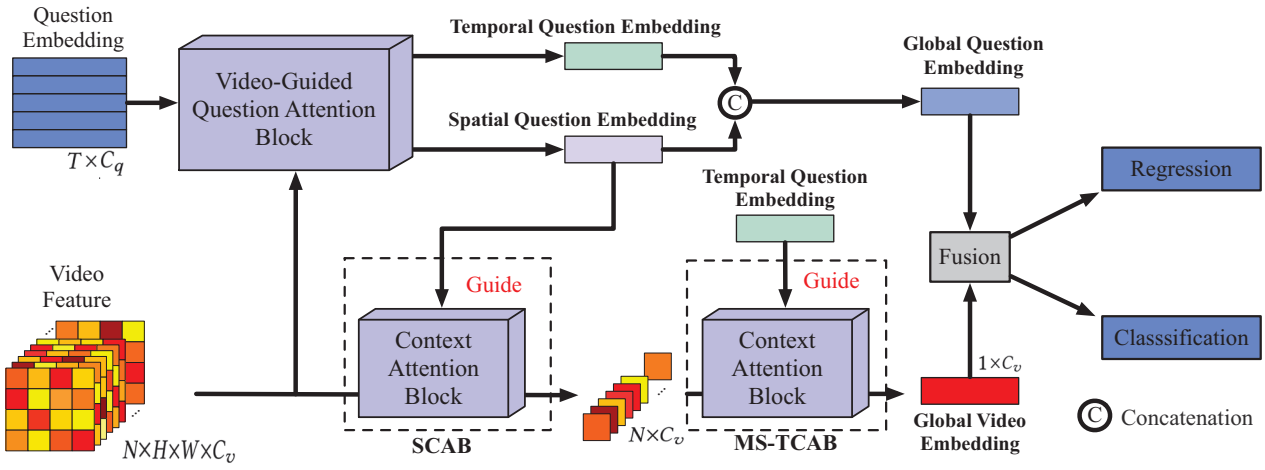

Figure 2: Overview of the proposed QueST model. First, information in question embedding are divided into spatial and temporal dimension in VGQA. Then under the guidance of the corresponding question embedding generated in VGQA, SCAB and MS-TCAB respectively excavate the spatial and temporal visual clues of answers inside the visual features to generated global video embedding. After fusing the global video embedding and question embedding, answers are finally generated.

tion Network (QueST) for Video QA. The input to QueST contains two modalities: video and question in natural language.

In QueST, the raw inputs are first fed into **Feature Extraction** module to obtain video embedding $\boldsymbol{V}$ and question embedding $\boldsymbol{Q}$. Next, as Figure 2 shows, the **Video-Guided Question Attention Block** (VGQA) is employed in question embedding to generate spatial question embedding and temporal question embedding. The two new embeddings focus on the information from corresponding dimension in the initial question embedding. Based on the new question embeddings, we introduce **Question-Guided Contextual Attention Block** (CAB) to obtain visual features related to the question from spatial and temporal dimensions. Then, we combine both spatial and temporal question embeddings with the video features generated by CABs to obtain a **Joint Representation** of the video and the question. At last, the answer is generated in **Answer Module** by the classification or regression branch.

Detailed descriptions of the aforementioned 5 procedures are provided as follows.

Feature Extraction

In the subsection, we describe the feature extraction method in QueST, which converts the input raw video and raw natural language data into feature embeddings.

Video. Most recent methods in Visual QA (Anderson et al. 2018; Jang et al. 2017; Gao et al. 2018; Li et al. 2019; Fan et al. 2019) adopt CNN methods, such as Faster R-CNN (Ren et al. 2015), ResNet (He et al. 2016), C3D (Tran et al. 2015), flow CNN, as the visual feature extractor. In QueST, for a given video, the features generated by these CNN methods are denoted as $\boldsymbol{V} = [v_1, v_2, ..., v_N]$, where N is the number of frames sampled in the video. $v \in \mathbb{R}^{H \times W \times C_v}$ is the feature of each frame, where H, W and C_v are height, width and channel dimension of v's feature map, respectively.

Question. A question can be represented as a sequence of words. We split the question with delimiter to acquire words. Then, a pre-trained GloVe (Pennington, Socher, and Manning 2014) is used to convert each word into a 300-D feature vector. In order to utilize the relation among words, the word embeddings are fed into LSTM (Hochreiter and Schmidhuber 1997) and the hidden state of each time step is used as the new word embedding and is collected to obtain the question embedding $\boldsymbol{Q} = [q_1, q_2, ..., q_T]$, where T is the number of words and $q_i \in \mathbb{R}^{C_q}$ is the feature of each word.

Video-Guided Question Attention Block

In this subsection, we introduce the Video-Guided Question Attention Block (VGQA) for understanding the question considering the video structure. Figure 3 demonstrates the VGQA module. We first fuse the video features and question embeddings to generate **Word Attention** to highlight the information related to video in the question. Then, we introduce self-attention mechanism and Diversity Loss(Liu, Jiang, and Wang 2019) to generate spatial question embedding and temporal question embeddings.

Fusion of Video Features and Question Embeddings. The inputs of VGQA are the video feature $\boldsymbol{V}$ and the initial question embedding $\boldsymbol{Q}$. The procedure of VGQA is illustrated in Figure 3. We first average video features along spatial and temporal dimensions to obtain a global video feature v^g. Then, we fuse the video features with the initial question embedding in order to acquire a joint feature for attention generation. A linear layer is used to project the C_v channel video feature v^g to a C^q_{inter} channel vector and each word embedding in Q is also projected to a C^q_{inter} channel vector by linear layers. Next, element-wise multiplication is adopted to generate a joint feature of the global video feature and each word embedding. This process for the i-th word can be formulated as:

$$J^q_i = \text{Fuse}(v^g, q_i) = W^q_v v^g \odot W^q_q q^i \quad (1)$$

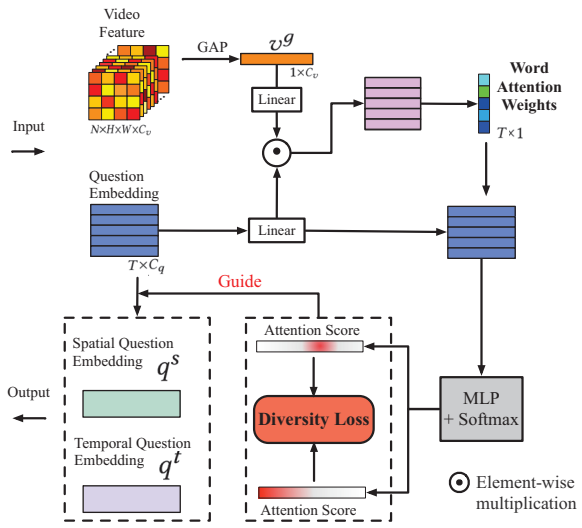

Figure 3: Video-Guided Question Attention Block (VGQA). It takes the video feature and initial question embedding as input. The video feature is fused with question embedding to guide the word attention on question embedding. Then by self-attention and Diversity Loss, the enhanced question embedding generate two new question embedding: spatial question embedding and temporal question embedding.

where W_v^q, W_q^q are learnable weights in $\mathbb{R}^{C_{inter}^q \times C_v}$, $\mathbb{R}^{C_{inter}^q \times C_q}$ respectively.

Word Attention on Question (WA). Based on the joint embedding J^q obtained as (1), we adopt a convolution layer to generate a weight score s^{qw} for each word. The weight score is applied in each word embedding to re-weight it by the significance of the word in the relation between question and video. To avoid the loss of important features during the attention operation, we add a residual connection to the output of WA to obtain enhanced question embedding Q^w.

$$\begin{aligned} s_i^{qw} &= \frac{exp(W_w^q J_i^q)}{\sum_{i=1}^{T} exp(W_w^q J_i^q)} \\ q_i^w &= s_i^{qw} q_i + q_i \\ Q^w &= [q_1^w, q_2^w, ..., q_T^w] \end{aligned} \tag{2}$$

where s_i^{qw} denotes the importance score assigned to each word and W_w^q is learnable weights.

Spatial and Temporal Question Embedding. After the above attention blocks, the video-related parts in question embedding have been enhanced in Q^w. Then self-attention mechanism is employed on Q^w here to generate two different question attention masks, denoted as $Mask^s$ and $Mask^t$. The $Mask^s$ and $Mask^t$ are designed for highlight the different parts of the question feature and then used to generate the spatial question embedding and temporal question embedding respectively. For instance, the spatial question embedding q^s is generated by:

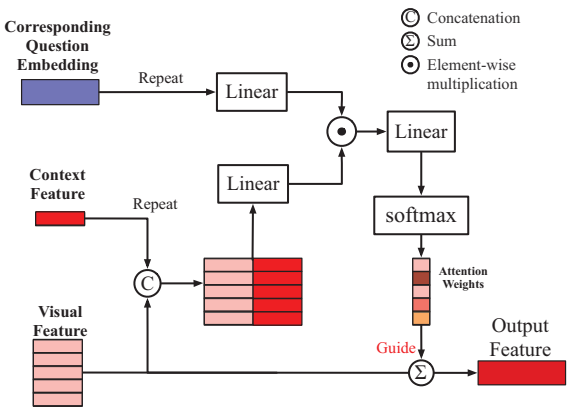

Figure 4: Question-Guided Contextual Attention Block.

$$\begin{aligned} mask_i^s &= \frac{exp(W_s^q q_i^w)}{\sum_{i=1}^{T} exp(W_s^q q_i^w)} \\ Mask^s &= [mask_1^s, mask_2^s, ..., mask_T^s] \\ q^s &= \sum_{i=1}^{T} mask_i^s q_i^w \end{aligned} \tag{3}$$

where W_s^q is the learnable weights.

In order to avoid the spatial question embedding and temporal question embedding focusing on the same parts of the question feature, We introduce Diversity Loss here. Diversity Loss maximizes the cosine similarity distance of $Mask^s$ and $Mask^t$ and help them concentrate on different parts of the question feature.

$$\mathcal{L}_{div} = \frac{Mask^s \cdot Mask^t}{\|Mask^s\| \|Mask^t\|} \tag{4}$$

Question-Guided Spatio-Temporal Contextual Attention Block

In this subsection, we describe the Question-Guided Spatio-Temporal Contextual Attention Block (CAB). As Figure 2 shows, CABs are employed to attend to spatial visual features and temporal visual features related to answer sequentially under the guidance of the corresponding question embedding generated from VGQA.

As shown in Figure 4, The inputs to CAB consist of three parts: the visual feature v^c, the context feature c^c, corresponding question embedding q^c. At first, we repeat the context feature and concatenate it with the feature of each position of visual features v^c. Next, we fuse the corresponding question embedding (spatial question embedding or temporal question embedding) and the context-aware visual feature to generate attentions. Then, we use the attentions to excavate the visual features related to answer. The procedure

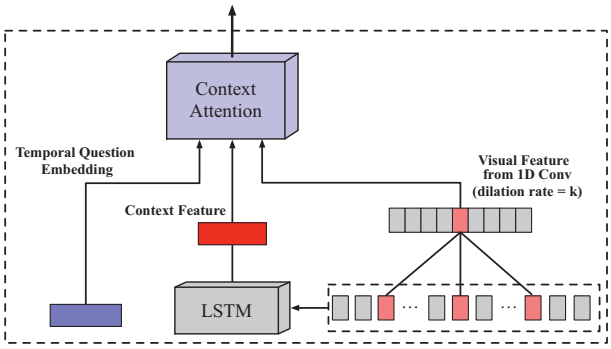

Figure 5: Temporal Contextual Attention Block (CAB) head.

can be formulated as:

$$\begin{aligned} \widetilde{v_i} &= Concatenate(v_i^c, c^c) \\ J^c &= W_v^c \widetilde{v_i} \odot W_q^c q^c \\ s_i^c &= \frac{exp(W_a^c J_i^c)}{\sum_{i=1}^{N} exp(W_a^c J_i^c)} \\ v^a &= \sum_{i=1}^{N} s_i^c v_i^c \end{aligned} \tag{5}$$

where W_v^c, W_q^c and W_a^c are learnable weights and i is the index of position in visual features. v^a is the output of the CAB.

Spatial Attention with CAB (SCAB). For the SCAB, we use the video feature V, extracted from CNN as the input visual feature. Next we average the V along spatial dimension to obtain the global feature for each frame as the input context feature and use the spatial question embedding as the input question embedding. Then the SCAB is applied in the spatial dimension and selects the question-related spatial region in each frame under guidance of spatial question embedding.

Temporal Attention with CAB (TCAB). There are some differences between the SCAB and the TCAB. As Figure 5 shows, we feed the output feature of the SCAB into a 1D temporal convolutional layer and use the output of convolutional layer as input visual feature. The context feature in the TCAB is modeled by LSTM model. Then, with temporal question embedding as input question feature, TCAB selects the question-related temporal region in the input visual feature under guidance of temporal question embedding.

We find that answering different questions require the temporal visual information of different durations. Therefore, based on CAB, we design a **Multi-Scale TCAB** (MS-TCAB) and adopt it in our QueST model. The MS-TCAB consists of K parallel TCAB heads and concatenate all the output of them as output of the MS-TCAB. Different dilation rates (Yu and Koltun 2015) are adopted for the 1D temporal convolutional layers in different TCAB heads, which can models the temporal visual information of different durations explicitly and helps to find the clues of answer more accurately.

Joint Representation

Through the attention mechanism on question, we obtain the spatial question embedding and the temporal question embedding. Now we combine them by concatenation to acquire global question embedding, denoted as q^f. The output visual feature of the sequent SCAB and MS-TCAB on initial video feature V is denoted as v^f. Then the global question embedding q^f and video embedding v^f are fused to generate a joint feature J^f for question answering.

$$J^f = W_v^f v^f \odot W_q^f q^f + b^f \tag{6}$$

where W_v^f, W_q^f and b^f are learnable parameters.

Answer Module

For most tasks in Video QA, the questions can be divided into three types: multi-choice, open-ended words and open-ended numbers.

For the open-ended words task, they are formulated as a classification task. A fully connected layer (FC) followed by a softmax function is employed on the joint embedding J^f to generate score for each answer. Then, a cross entropy loss is used to train the network.

For the multi-choice task, the question is attached with some candidate answers. We first model each candidate answer as same as the way initializing question embedding $\boldsymbol{Q}$ and generate candidate answer embedding, which is a vector of the same dimension with output joint embedding J^f of QueST. Then, joint embedding J^f is element-wise multiplied by each candidate answer embedding to generate the new joint embedding for each candidate answer. Next, a sharing weight FC is employed to project the new joint embedding to a real number. At last a softmax function is adopted to normalize the score along the candidate answers to predict the probability for each answer. Cross entropy loss is adopted here.

For the open-ended numbers task, such as counting, it is formulated as a regression task. A FC is used to predict a real number for answering and the Mean Square Error loss is used to train the network.

During training phase, we combine the losses described above with the Diversity Loss (described in Equation (4)) with a coefficient λ to train our QueST model. The λ is set to 0.25 in the experiments.

Experiments

In this section, we evaluate the proposed QueST on three standard Video QA datasets: TGIF-QA, MSRVTT-QA and MSVD-QA. We first present the experimental results and comparisons with the state-of-the-art methods on TGIF-QA. Then, we provide the ablation studies to investigate the proposed attention modules in QueST. At last, we report results and comparisons on MSRVTT-QA and MSVD-QA.

Experiments on TGIF-QA

TGIF-QA (Jang et al. 2017) is a large-scale dataset for Video QA, which consists of 165,165 question-answer pairs collected from 71,741 GIFs. In TGIF-QA, there are 4 types of tasks: repetition action, state transition, frame QA and

Table 1: Details of TGIF-QA dataset.

QA pairs	Action	Trans.	Frame QA	Count
Training	20,475	52,704	39,392	26,843
Testing	2,274	6,232	13,691	3,554
Total	22,749	58,936	53,083	30,397

Table 2: Comparisons with the state-of-the-art method on TGIF-QA dataset. R denotes ResNet152 feature, C denotes C3D feature and F denotes additional modality feature of optical flow.

Method	Action	Trans.	FrameQA	Count
ST-S.-T.(R+C)	57.0	59.6	47.8	4.56
ST-TP(R+C)	60.8	67.1	49.3	4.40
Co-Mem(R+F)	68.2	74.3	51.5	4.10
PSAC(R)	70.4	76.9	55.7	4.27
HME(R+C)	73.9	77.8	53.8	**4.02**
QueST(R)	**75.9**	**81.0**	**59.7**	4.19

repetition count. Repetition action and state transition are multiple choice tasks. Question is attached with five options. Frame QA is similar to Image QA and is an open-ended word task. Repetition count requires the model to count the number of repetitions of a certain action.

Experiment Settings. In experiments, we use the standard training/testing splitting as provided in (Jang et al. 2017) and the details for each task are shown in Table 1. Given a gif in TGIF-QA, we evenly sample 10 frames to represent the video. Then the output of *res4c* layer ($\in \mathbb{R}^{14\times14\times1,024}$) in ResNet-152 followed an average pooling with a stride of 2 is selected as the visual feature for each frame, which contains more appearance information than the deeper layers in ResNet. Given a question, a pre-trained 300-D GloVe embedding is used to convert each word to word embedding. Then, we train our QueST model with Adam (Kingma and Ba 2014) optimizer for each task. Normally, we set the size of minibatch as 64 and the initial learning rate as 0.001.

Comparisons with the State-of-the-art Methods. Our proposed QueST method has been compared with recent state-of-the-art methods and the results are shown in Table 2. The results show that QueST outperforms the state-of-the-art methods (*i.e.* HME (Fan et al. 2019) and PSAC (Li et al. 2019)) by 2.0%, 3.2% and 4.0% of accuracy on action, transition and frame QA tasks. For count task, our model also obtains better performance with other methods by only using ResNet features. It is noted that, QueST only uses RGB-ResNet features as visual input and still obtains superior performance than some other multi-modality methods, *i.e.*, ST-SP-TP (ST-S.-T.) and ST-TP (Jang et al. 2017), Co-Mem (Gao et al. 2018) and HME(Fan et al. 2019), on most of the tasks.

Ablation Study on Video-Guided Question Attention. We further investigate the effectiveness of Video-Guided Question Attention (VGQA). In the VGQA, we utilize the global visual features to refine the initial question embedding via attention on word dimensions. Then, we introduce Diversity Loss and self-attention to generate spatial question embedding and temporal question embedding. Here we remove the word attention (WA) and Diversity Loss in the VGQA respectively to investigate the effectiveness of these modules. And we also design a model setting named ST-VQA where VGQA generates only one question embedding as the way used in (Jang et al. 2017). The experimental results is shown in Table 3. We can observe that removing word attention from the full model can lead to performance degeneration. The reason is that word attention introduces visual information to co-model the question and highlight the information related to video. From the results, we can find that the performance of QueST without Diversity Loss is lower than full model on most tasks and is similar to QueST with single question embedding, *i.e.* the model setting of ST-VQA, which suggests the Diversity Loss can help to divide the question information into two parts and is beneficial for modeling question in the spatial and temporal aspects respectively. Compared with ST-S.-T. in Table 2, which also applies spatio-temporal attentions on videos but obtains even worse performance than temporal-only attention model, our VGQA can construct better spatio-temporal attentions and achieve superior performance.

Table 3: Ablation experiments on the VGQA Module.

Setting	Action	Trans.	Frame QA	Count
w/o WA	75.0	80.7	58.5	4.36
w/o DL	74.6	79.3	57.3	**4.17**
ST-VQA	74.7	78.5	56.8	4.33
Full	**75.9**	**81.0**	**59.7**	4.19

Table 4: Ablation experiments on the CAB Module.

Setting	Action	Trans.	Frame QA	Count
ST-VQA	73.9	78.6	56.6	4.28
ST-CAB	74.1	79.1	59.0	4.23
MS-ST-CAB	**75.9**	**81.0**	**59.7**	**4.19**

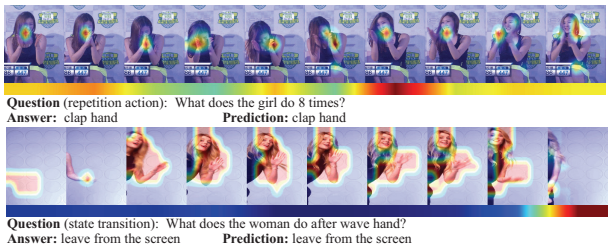

Figure 6: Visualization of results of QueST in TGIF-QA. For each QA pair, we visualize weights of the SCAB and attach weights of TCAB under each frame. Take the sample in first row as example, the attention focuses on the key regions about the action.

Table 5: Details of MSRVTT-QA dataset.

QA pairs	what	who	how	when	where
Training	108,792	43,592	4,067	1,626	504
Val.	8,337	3,439	344	106	52
Testing	49,869	20,385	1,640	677	250
Total	166,998	67,416	6,051	2,409	806

Table 6: Details of MSVD-QA dataset.

QA pairs	what	who	how	when	where
Training	19,485	10,479	736	161	72
Val.	3,995	2,168	185	51	16
Testing	8,149	4,552	370	58	28
Total	31,629	17,199	1,291	270	116

Ablation Study on Contextual Attention on Video. Here we conduct the ablation experiments on the Question-Guided Contextual Attention (CAB) including SCAB and TCAB. The experimental results are listed in Table 4.

ST-VQA denotes attention methods in (Jang et al. 2017) are used in QueST instead of our proposed CABs. ST-CAB denotes question-guided contextual spatio-temporal attentions (SCAB+TCAB) are adopted in QueST and MS-ST-CAB denotes SCAB and multi-scale TCAB (MS-TCAB) are adopted in QueST.

Compared with ST-VQA, models with CABs obtain better performance on all tasks. MS-ST-CAB improves the accuracy of the tasks of action, transition and frame QA by 2.0%, 2.4% and 3.1% respectively and also reduces the MSE loss. The results demonstrate that the contextual information is important for the Video QA task and has not been sufficiently utilized in existing spatio-temporal attention methods in Video QA. Compared with ST-CAB, MS-ST-CAB model the temporal information in different duration with CABs explicitly and obtains the gains of 1.8%, 1.9% and 0.7% of accuracy for the tasks of action, transition and frame QA respectively.

Visualization. We visualize the attention weights of CAB in Figure 6 to demonstrate the effectiveness of our QueST. We can notice that both spatial and temporal attentions can be accurately detected and thus lead to better Video QA performance. In the given clapping hand example, although it is somehow difficult to count the times of the action even for human, the model can figure out that the action which the woman do most times is clapping hands.

Experiments on MSRVTT-QA and MSVD-QA

To further evaluate the effectiveness of the proposed model, we have also tested our QueST on other Video QA datasets: MSRVTT-QA (Xu et al. 2017) and MSVD-QA (Xu et al. 2017). The tasks in MSRVTT-QA and MSVD-QA are open-ended word tasks and questions can be divided into 5 types, including what, who, how, when and where, by the first word of the question. We list the details of training/validation/testing splitting for each task MSRVTT-QA and MSVD-QA in Table 5 and Table 6 respectively. We compare our proposed QueST with recent methods, *i.e.*, ST-VQA (Xu et al. 2017), Co-Mem (Gao et al. 2018), GRA (Xu et al. 2017), HME (Fan et al. 2019).

Table 7: Experiments on testing set of MSRVTT-QA.

Method	what	who	how	when	where	all
ST-VQA	24.5	41.2	78.0	76.5	34.9	30.9
GRA	26.2	43.0	80.2	72.5	30.0	32.5
Co-Mem	23.9	42.5	74.1	69.0	42.9	32.0
HME	26.5	43.6	82.4	76.0	28.6	33.0
QueST	**27.9**	**45.6**	**83.0**	75.7	31.6	**34.6**

Table 8: Experimental results on testing set of MSVD-QA.

Method	what	who	how	when	where	all
ST-VQA	18.1	50.0	83.8	72.4	28.6	31.3
Co-Mem	19.6	48.7	81.6	74.1	31.7	31.7
GRA	20.6	47.5	**83.5**	**72.4**	**53.6**	32.0
HME	22.4	50.1	73.0	70.7	42.9	33.7
QueST	**24.5**	**52.9**	79.1	**72.4**	50.0	**36.1**

For MSRVTT-QA, the experimental results and comparisons with the state-of-art methods are listed in Table 7. In these results, our QueST outperforms the state-of-the-art methods, *i.e.* HME, by 1.6% overall accuracy on the testing set. Our QueST can obtain the gains of 1.4%, 2.0%, 0.6% on the question types of what, who and how respectively. In other question types, our QueST also obtains comparable performance. The performances on the whole dataset and the question types with a relative large scale (like what, who and how) can demonstrate the effectiveness of our QueST method. For MSVD-QA, from the results shown in 8, QueST outperforms the state-of-the-art method, *i.e.* HME, by 2.4% overall accuracy and obtains the best accuracy on the three question types (what, who and when).

Conclusion

In this paper, we propose Question-Guided Spatio-Temporal Contextual Attention Network (QueST) for Video QA, including two modules: Video-Guided Question Attention Block (VGQA) and Question-Guided Contextual Attention Blocks (CABs). Through applying VGQA and CABs, QueST divides the information of question into spatial part and temporal part, which helps to better interpret visual features under the guidance of question information of corresponding dimension. Experimental results on three benchmark Video QA datasets show that QueST can achieve significant performance improvement on Video QA compared with the state-of-the-art methods.

Acknowledgments

This work was supported by National Natural Science Funds of China (U1701262, U1801263, 61671267), and Beijing Natural Science Foundation (Grant No. 4182022).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Ben-Younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal Tucker Fusion for Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2612–2620.

Fan, C.; Zhang, X.; Zhang, S.; Wang, W.; Zhang, C.; and Huang, H. 2019. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999–2007.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *arXiv preprint arXiv:1606.01847*.

Gao, Y.; Beijbom, O.; Zhang, N.; and Darrell, T. 2016. Compact Bilinear Pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 317–326.

Gao, J.; Ge, R.; Chen, K.; and Nevatia, R. 2018. Motion-Appearance Co-Memory Networks for Video Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6576–6585.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hochreiter, S., and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural computation* 9(8):1735–1780.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2758–2766.

Kazemi, V., and Elqursh, A. 2017. Show, Ask, Attend, and Answer: A Strong Baseline for Visual Question Answering. *arXiv preprint arXiv:1704.03162*.

Kingma, D. P., and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; and Socher, R. 2016. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. In *International Conference on Machine Learning*, 1378–1387.

Li, X.; Song, J.; Gao, L.; Liu, X.; Huang, W.; He, X.; and Gan, C. 2019. Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering. In *AAAI Conference on Artificial Intelligence*, 8658–8665.

Liu, D.; Jiang, T.; and Wang, Y. 2019. Completeness Modeling and Context Separation for Weakly Supervised Temporal Action Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1298–1307.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, 91–99.

Singh, J.; Ying, V.; and Nutkiewicz, A. 2018. Attention on Attention: Architectures for Visual Question Answering (VQA). *arXiv preprint arXiv:1803.07724*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 4489–4497.

Wu, Q.; Wang, P.; Shen, C.; Dick, A.; and van den Hengel, A. 2016. Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4622–4630.

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic Memory Networks for Visual and Textual Question Answering. In *International Conference on Machine Learning*, 2397–2406.

Xu, H., and Saenko, K. 2016. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In *European Conference on Computer Vision*, 451–466.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*, 2048–2057.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *Proceedings of the ACM International Conference on Multimedia*, 1645–1653.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked Attention Networks for Image Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21–29.

Yu, F., and Koltun, V. 2015. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv preprint arXiv:1511.07122*.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In *AAAI Conference on Artificial Intelligence*, 9127–9134.

Yu, Y.; Kim, J.; and Kim, G. 2018. A Joint Sequence Fusion Model for Video Question Answering and Retrieval. In *Proceedings of the European Conference on Computer Vision*, 471–487.