Two-Stream Spatiotemporal Compositional Attention Network for VideoQA

Taiki Miyanishi[1,3]
miyanishi@atr.jp
Takuya Maekawa[2]
maekawa@ist.osaka-u.ac.jp
Motoaki Kawanabe[1,3]
kawanabe@atr.jp

[1] Advanced Telecommunications Research Institute International (ATR) Kyoto, Japan
[2] Graduate School Information Science and Technology, Osaka University Osaka, Japan
[3] RIKEN Center for Advanced Intelligence Project (AIP) Kyoto, Japan

Abstract

This study tackles video-question answering (VideoQA), which requires spatiotemporal video reasoning. VideoQA aims to return an appropriate answer about textual questions by referring to image frames in video. Based on the observation that multiple entities and their movements in a video can be important clues for deriving correct answers, we propose a two-stream spatiotemporal compositional attention network that achieves sophisticated multi-step spatiotemporal reasoning using both motion and detailed appearance features. In contrast to the existing video reasoning approach that uses frame-level or clip-level appearance and motion features, our method simultaneously attends the region-level appearance features of multiple entities as well as the motion features guided by the attending words in the textual question. Furthermore, it progressively refines internal representation and infers answers by multiple reasoning steps. We evaluate our method on short- and long-form VideoQA benchmarks, MSVD-QA, MSRVTT-QA, and ActivityNet-QA, and achieve state-of-the-art accuracy on these datasets.

1 Introduction

The goal of video-question answering (VideoQA) is to produce an appropriate answer based on textual questions that inquire about the visual content in a video. Using this technology, we can quickly understand real-world events and situations in videos through natural language. VideoQA technology will play an important role in a wide range of practical applications, such as information access to personal visual histories [9], question answering (QA) for tutorial videos [6], video dialogue systems [4], and embodied agents with visual perception [7].

In contrast to traditional visual question answering for static images [2, 14, 34], VideoQA is more challenging because its system has to find relevant frames to a question and provide answers from possibly unnecessary image frames in the video. To address this problem, existing VideoQA approaches use the appearance and motion features extracted from a series of frames and clips in a video with a pre-trained convolutional network (ConvNets)

Figure 1: VideoQA example: Q1 can be correctly answered by finding a frame (or clip) from video containing the entity in question and the motions associated with the answer. Q2 can be correctly answered by finding entities in the image frames related to the question and its answer.

model [16, 39] and a 3D ConvNets [15, 41] scheme and apply learnable soft weights (i.e., an attention mechanism [3]) to them to capture frame- and clip-level details relevant to a given question [44, 52, 55]. The limitation of current VideoQA approaches is their use of a single encoded vector for representing the semantics of questions. To capture the more complex semantic relationships between question words and frames (and clips), several works simultaneously attend visual contents and their related part of words in a question [13, 25, 31, 32, 33]. Some notable works use multi-step reasoning that gradually refines the motion-appearance representations of video and question representations [10, 12, 46, 48]. These multi-step video reasoning approaches achieved competitive performances on short- and long-form VideoQA datasets. Previous results of these existing works suggest the effectiveness of motion-appearance features, simultaneous attention over words and visual contents, and progressive refinement through multi-step video reasoning. However, even though the events that occurred in the video involve multiple entities (e.g., humans and objects) [23, 42], these methods fail to capture the associations between the region-level details of the entities in the frame and their corresponding question words. As described in the examples in Fig. 1, to get correct answers for VideoQA, the detailed appearance information of the entities in the frame is an important clue as well as the motion information over the frames.

Motivated by this observation, we developed a two-stream spatiotemporal MAC network (TS-STMAC), which performs sequential spatiotemporal reasoning on video frames based on the question content. We used a SlowFast model that shows high performance in video-understanding tasks [11] and a bottom-up attention model useful for image VQA tasks [1] for extracting robust motion and detailed appearance features. Our TS-STMAC network is a natural extension of the Memory, Attention, and Composition (MAC) network [20], which yields promising results in spatial-reasoning tasks [21, 29] based on compositional attention. More concretely, we devised a two-stream spatiotemporal MAC cell, which is a new neural module containing a spatiotemporal attention mechanism that simultaneously finds motion features and detailed appearance features of the entity's regions relevant to the attending words in a question. We use it as a building block for our VideoQA framework, recurrently apply it for multi-step reasoning, and progressively infer the correct answer. Through such question-aware multi-step spatiotemporal reasoning, the model can focus on the critical frames and regions and ignore useless information.

The main contributions of this work are threefold. First, we devise a TS-STMAC cell that simultaneously captures the relationship between entity regions and motion over frames based on the attended question words. Second, we incorporate this TS-STMAC cell into a recurrent network that performs iterative spatiotemporal reasoning for VideoQA. This multi-step reasoning progressively refines the internal network representation for answering questions. Third, we conduct experiments on short- and long-form VideoQA datasets to validate our method's effectiveness and show that it outperforms the state-of-the-art approaches by a large margin on three public benchmarks.

2 Related Work

VideoQA resembles an extension of image-based visual question answering (VQA) into the video domain. This task requires both language and video understanding to infer correct answers from complex semantics. Most current approaches use temporal reasoning methods with attention mechanisms over the temporal dimension for extracting the important frame information from a video [35, 44, 45, 47, 50, 56]. Although these works use frame-level attention for videos, some VideoQA models use segment-level attention [52, 53, 55] to address the long-range dependency of the video context. Instead of explicitly using segments in the video, we use motion features extracted from short clips to represent segment information. Due to the video's nature, some complex questions in VideoQA tasks cannot be solved without viewing multiple frames in the video. To capture temporal relationships over the frames, some methods use self-attention mechanisms or temporal relational modeling and graph ConvNets [25, 31, 33]. Our method can also consider the temporal relationship over frames using representations of the internal state obtained from past inference steps and the input frames in the current step. In contrast to the static images used for the standard VQA, video contains dynamic information that captures real-world events. Methods have been proposed that take into account motion and appearance information that represents dynamics in the video guided by questions [10, 12, 46, 48]. These methods show high performance in multiple VideoQA benchmarks. In contrast, our method models the fine-grained appearance information from object-detection networks as well as the motion information from video recognition networks.

In contrast with modeling the frame-level temporal dynamics of video, spatiotemporal reasoning approaches, which focus on the frame- and region-level visual content relevant to a question, remain relatively unexplored. Traditional approaches use a combination of recurrent neural networks (RNN) and ConvNets, which encode spatiotemporal video features and a textual question, and jointly learn their multi-modal representations [22, 54]. However, these works dont model the interaction between question words and visual contents. Some words in the question often indicate entities in the video, which can be important clues for video reasoning. To further improve the VideoQA performance, the QA model has to indicate words in the question that correspond to the image regions and video frames [24, 51]. In addition to attending both textual and visual content, recent works use the fine-grained appearance of video frames with external knowledge [27] or spatial relationships among entities in the video frames [19, 26]. However, only using appearance information is inadequate to capture movement in videos, which is essential for questions about the motions of humans and objects. To overcome this limitation, we use motion features over frames as well as detailed appearance features. Several works use motion-appearance features for spatiotemporal video reasoning [26, 40]. However, these works lack an attention mechanism for question words, even though word-level attention plays an important role in finding frames that represent motion information and image regions that represent detailed appearance information relevant to a question. Our work differs because our proposed neural module can simultaneously attend to question words, frames, and image regions to represent their associations. Moreover, our question-aware spatiotemporal network uses this neural module as a building block and can progressively infer relevant answers by multi-step video reasoning to focus on important video information. Our sophisticated method outperforms existing temporal or spatiotemporal reasoning methods on both short- and long-form VideoQA datasets.

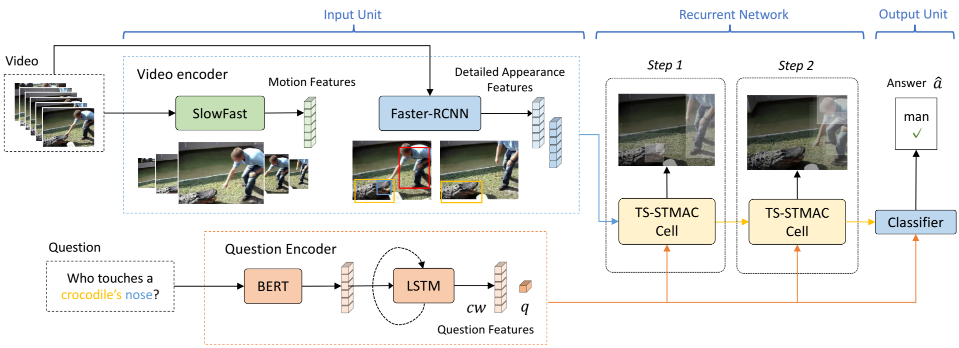

Figure 2: Illustration of our two-stream spatiotemporal MAC (TS-STMAC) network. First, video encoder extracts motion and detailed appearance features from short clips and frames using SlowFast networks and Faster-RCNN (top left). Question encoder extracts text features from question words using BERT and LSTM (bottom left). Then a neural module TS-STMAC cell takes these features as inputs and computes the interaction between question and video features by focusing on frames (or clips) and regions relevant to the question. The network repeats this process multiple times to progressively refine the internal representation. Finally, the classifier predicts the final answer using question embedding and the final memory state of the TS-STMAC cell. The regions in the selected frames with higher attention values at each step are shown more brightly.

3 Approach

3.1 Problem Definition

In this work, we consider the following VideoQA task. Given video $v \in \mathcal{V}$ and question $q \in \mathcal{Q}$ about this video, the VideoQA method outputs an answer: $\hat{a} \in \mathcal{A}$. Our goal is to predict answer $\hat{a}$ that matches true answer a^*.

Video Embeddings: Video consists of the sequence of frames that have multiple regions representing entities. For motion representation, we use a Kinetics-600 classification model of SlowFast networks that achieved high performance for action detection tasks [11]. We extract the motion feature ($f_t^a \in \mathbb{R}^{2304}$) from the t-th clip and use a series of motion features $f^a = \{f_t^a\}_{t=1}^T$ for representing the video, where T is the number of clips. For detailed appearance information, we extract the region features ($f_t^b = \{f_{i,t}^b\}_{i=1}^N$) from the t-th frame using Faster R-CNN [38], trained with the Visual Genome dataset [30], where each $f_i^b \in \mathbb{R}^{2048}$ corresponds to a region feature of an entity, and N is the number of detected entities with the highest confidence scores. Following past VQA work [1], we set $N = 36$ and use the image feature in the region multiplied by its confidence scores as the region feature. For the appearance features of the video, we use a series of sets of region features $f^b = \{f_t^b\}_{t=1}^T$. The VideoQA models input is a tuple of these motion-appearance features and the following question features.

Question Embeddings: For question representation, we use a BERT model [8]. To deal with unknown words that appear in the training data but not in the test data, we first split a question into words of length M by the WordPiece tokenizer [43] and extracted a feature vector from the last layer of a pre-trained 12-layer BERT model for each word. Note that we fine-tuned this layer during the VideoQA training. Then we encoded the question using a one-layer bi-directional LSTM (biLSTM) [18], which is used for guiding the models multi-step reasoning. We used a series of output states from LSTM $\{cw_i\}_{i=1}^M$ as contextual question word embeddings. We also used $q \in \mathbb{R}^{2d}$ as a question-sentence embedding, which is represented by the concatenation of the final hidden states from the backward and forward LSTMs. We also applied a linear transformation to q for representing step-aware question embedding $q_i \in \mathbb{R}^d$ at the i^{th} reasoning step.

3.2 Two-Stream Spatiotemporal MAC Network

For VideoQA, we developed a two-stream spatiotemporal MAC (TS-STMAC) network that consists of an input unit, a core recurrent network, and an output unit. Fig. 2 shows an overview of our proposed model. The input unit transforms the raw video and a question into distributed vector representations. The core recurrent network sequentially addresses the question by decomposing it into a series of operations (control) that retrieve information from the video (clip- and frame region-level features) and aggregates the results into internal memory. As the core recurrent network, we repeatedly used the following TS-STMAC cells at each step.

We introduced a two-stream spatiotemporal MAC cell, which is the building block for our VideoQA model. Our proposed cell mainly consists of two neural components: temporal and spatial MAC cells. Because both cells are based on MAC cells [20], we start with a brief explanation of this neural module, which has been used for spatial-reasoning tasks [28].

MAC Cells: A MAC cell is a neural module designed to apply attention-based operations to perform reasoning. It holds two hidden states at the i-th step: control $c_i \in \mathbb{R}^d$ and memory $m_i \in \mathbb{R}^d$. Control state c_i stores the information on the reasoning operation that should be performed. Memory state m_i has an intermediate result that was computed in the recurrent reasoning process. The MAC cell updates the control and memory states for each reasoning step $i = 1, \dots, S$ using three internal units: control, read, and write. It iteratively aggregates information from a knowledge source according to the control state in the following steps: (i) The control unit focuses on some words of the question using an attention mechanism [3] and updates control state c_i. (ii) The read unit attends to some parts of knowledge base $\{k\}_{i=1}^{K}$ (e.g., image features for VQA) and retrieves information r_i from them based on the current control and previous memory states c_i and m_{i-1}, where K denotes the size of the knowledge base. (iii) The write unit updates the memory based on retrieved information r_i and previous memories $\{m_0, \dots m_{i-1}\}$. The following are the equations of the reasoning step in the MAC cell:

$$c_i = \mathrm{ControlUnit}(c_{i-1}, \{cw_j\}_{j=1}^{M}, q_i) \tag{1}$$

$$r_i = \mathrm{ReadUnit}(m_{i-1}, \{k_j\}_{j=1}^{K}, c_i) \tag{2}$$

$$m_i = \mathrm{WriteUnit}(\{m_{j-1}\}_{j=1}^{i}, r_i, c_i). \tag{3}$$

Due to space limitations, see [20] for more details about these neural units. As mentioned in Section 1, using motion and detailed appearance information is important to solve VideoQA. However, normal MAC cells can only handle one of them. Therefore, we extended this MAC cell and created a TS-STMAC cell that can handle both motion and detailed appearance features for spatiotemporal reasoning.

Two-Stream Spatiotemporal MAC Cell: Figure 3 shows the proposed two-stream spatiotemporal MAC (TS-STMAC) cell architecture, which consists of two cells: a spatial cell and a temporal MAC cell. The latter is used for representing the temporal structure of the video. We used the motion features of the clips in video $\{f_j^a\}_{j=1}^{T}$ as this cells input. The temporal MAC cell updates the controller and the memory states based on the motion features.

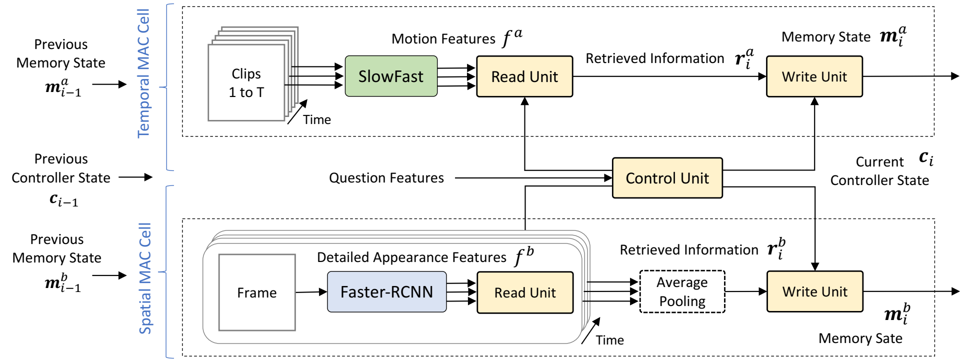

Figure 3: Overview of our two-stream spatiotemporal MAC (TS-STMAC) cell, which consists of a temporal and a spatial MAC cell. Temporal MAC cell (top) takes motion features as input and updates internal representation m^a that holds temporal information over clips based on control state c. Spatial MAC cell (bottom) takes detailed appearance features as input and updates internal representation m^b that holds spatial information over regions in frames based on c.

As with a standard MAC cell, a temporal MAC cell is given by

$$c_i = \mathrm{ControlUnit}(c_{i-1}, \{cw_j\}_{j=1}^{M}, q_i) \tag{4}$$

$$r_i^a = \mathrm{ReadUnit}_{\mathrm{temporal}}(m_{i-1}^a, \{f_t^a\}_{t=1}^{T}, c_i) \tag{5}$$

$$m_i^a = \mathrm{WriteUnit}_{\mathrm{temporal}}(\{m_{j-1}^a\}_{j=1}^{i}, r_i^a, c_i), \tag{6}$$

where $m^a \in \mathbb{R}^d$ and $r^a \in \mathbb{R}^d$ denote the memory state and the retrieved information of the temporal MAC cell, which holds the temporal information of the video content based on controller state c_i. ControlUnit, $\mathrm{ReadUnit}_{\mathrm{temporal}}$, and $\mathrm{WriteUnit}_{\mathrm{temporal}}$ are the same units of Eqs. 1, 2, and 3.

The spatial MAC cell, which represents the spatial structure of the video frames, takes as input a series of visual feature sets $f^b = \{f_t^b\}_{t=1}^{T}$ (i.e., detailed appearance features), which are extracted from T video frames. The spatial MAC cell uses the read unit multiple times to handle a series of feature sets with arbitrary length. First, the spatial MAC cell retrieves spatial information $r_{i,t}^b$ from region features $\{f_{j,t}^b\}_{j=1}^{N}$ of t^{th} frame by selectively focusing on specific regions based on control state c_i:

$$r_{i,t}^b = \mathrm{ReadUnit}_{\mathrm{spatial}}(m_{i-1}^b, \{f_{j,t}^b\}_{j=1}^{N}, c_i), \tag{7}$$

where $m^b \in \mathbb{R}^d$ and $r^b \in \mathbb{R}^d$ denote the memory state and the retrieved information of the spatial MAC cell that holds the spatial information of the video frames. $\mathrm{ReadUnit}_{\mathrm{spatial}}$ is the same unit of Eq. 2. The spatial MAC cell repeats this process for all frames and obtains T retrieved spatial information $\{r_{i,t}^b\}_{t=1}^{T}$. After that, average pooling is applied to them for aggregating the common spatial information related to a question over the video frames as follows:

$$r_i^b = \mathrm{pool}(\{r_{i,1}^b, r_{i,2}^b, \cdots, r_{i,T}^b\}), \tag{8}$$

where pool denotes the average pooling layer. Then the spatial MAC cell updates the memory state on the spatial information:

$$m_i^b = \mathrm{WriteUnit}_{\mathrm{spatial}}(m_{i-1}^b, r_i^b, c_i), \tag{9}$$

where $\text{WriteUnit}_{\text{spatial}}$ is the same unit of Eq. 3.

Due to both the spatial and temporal MAC cells, the TS-STMAC cell can jointly model the video's spatial and temporal structures based on a textual question by attending motion-appearance features guided by question word features.

Output Unit: We compute the final answer with a simple classifier using the question and final memory states of the spatial and temporal MAC cells after applying S cell computations as input:

$$o' = W_1[q; m_S^a; m_S^b] + b_1, \quad o = \text{softmax}(\text{ELU}(W_2 o' + b_2)), \tag{10}$$

where W_1, W_2, b_1, and b_2 are learnable parameters, and ELU is an exponential linear unit [5]. The classifiers final output is given by

$$a = \text{argmax}_{a \in \mathcal{A}} o. \tag{11}$$

4 Evaluation

4.1 Experimental Setup

Datasets: On three VideoQA datasets, we compared our method with its different components and several state-of-the-art approaches. We evaluated with the MSVD-QA [44], MSRVTT-QA [44], and ActivityNet-QA [49] datasets. MSVD-QA and MSRVTT-QA are short-form VideoQA datasets. The average lengths of the videos in these datasets are 10 and 15 sec., respectively. Both MSVD-QA and MSRVTT-QA include five different question types: *what*, *who*, *how*, *when*, and *where*. In contrast, ActivityNet-QA is a more challenging VideoQA dataset that uses longer videos about human activities. The average length of its videos is 116 sec. The videos were sampled from the ActivityNet dataset [17]. ActivityNet-QA includes four main question types: (*motion*, *spatial relationship*, *temporal relationship*, and *other*). Furthermore, according to their answer types, the *Free* questions are divided into six sub-question types: (*yes/no*, *number*, *color*, *object*, *location*, and *other*. We sampled 20 frames at equal intervals for appearance feature extraction and 20 clips for motion feature extraction. For the answer candidates, we selected the top 1,000 most frequent answers in a training split.

Implementation Details: We trained our method with up to 100 epochs using AMSGrad [37] (a variant of Adam [49]) for optimization with a learning rate of $\alpha = 10^{-4}$ and a batch size of 32. We employed early stopping if the validation accuracy did not increase for ten epochs. We converted the words in the questions and answers to lower cases and set dimension d of the TS-STMAC cell to 256. For the multi-step reasoning of the TS-STMAC network, we used two reasoning steps ($S = 2$) following the average performance on the validation data across three VideoQA datasets. We also used self-attention connections between the cells.

Evaluation Metric: Following past works [10, 49], we measured the performance using accuracy. The evaluation metric is given by $\text{Accuracy} = \frac{1}{|\mathcal{Q}|}\sum_{i=1}^{|\mathcal{Q}|} 1[a_i^* = a_i]$, where indicator function $1[\cdot]$ equals 1 only if a_i^* and a_i are the same and 0 otherwise.

4.2 Ablation Experiments

To verify the contribution of the proposed modules in the TS-STMAC network, we first compared four architectures with different neural modules on three VideoQA datasets. In addition to the proposed TS-STMAC, we prepared the following variants: temporal MAC

Method	Feature			Dataset		
	Text	Motion	Appearance	MSVD-QA	MSRVTT-QA	ActivityNet-QA
TMAC	`Glove`		`ResNet`	0.371	0.368	0.365
TMAC	`Glove`	`SlowFast`		0.393	0.377	0.385
SMAC	`Glove`		`RCNN`	0.375	0.369	0.366
TS-TMAC	`Glove`	`SlowFast`	`ResNet`	0.400	0.378	0.381
TS-STMAC	`Glove`	`SlowFast`	`RCNN`	0.401	0.378	0.385
TMAC	`BERT`		`ResNet`	0.397	0.382	0.365
TMAC	`BERT`	`SlowFast`		0.413	0.388	0.381
SMAC	`BERT`		`RCNN`	0.401	0.385	0.370
TS-TMAC	`BERT`	`SlowFast`	`ResNet`	0.415	0.391	0.390
TS-STMAC	`BERT`	`SlowFast`	`RCNN`	**0.432**	**0.394**	**0.402**

Table 1: Comparison with different VideoQA architectures with different features. Best result for each dataset is denoted by boldface.

(TMAC), spatial MAC (SMAC), and two-stream temporal MAC (TS-TMAC) networks. TMAC used a single temporal MAC cell as a core recurrent network that can use either motion or appearance features as inputs. It resembles a simple baseline that applied the MAC network [20] with temporal attention over frames to the VideoQA task. SMAC used a single spatial MAC cell that can use detailed appearance features for video reasoning. TS-TMAC used two temporal MAC cells to consider both clip-level motion and frame-level appearance features. As described in Section 3.2, TS-STMAC used spatial and temporal MAC cells to address both motion and detailed appearance features. We also compared the performance with different textual, motion, and appearance features to evaluate their complementary effects. For comparison to `BERT` word embeddings, we prepared `Glove` examples ($\in \mathbb{R}^{300}$) that were initialized with Glove [36]. To validate the effectiveness of the region-level appearance feature `RCNN` extracted from Faster-RCNN (i.e., detailed appearance features), we prepared a frame-level appearance feature `ResNet` ($\in \mathbb{R}^{2048}$) extracted from ResNet101 [16]. `SlowFast` denotes the clip-level motion features extracted from SlowFast networks.

Table 1 shows the accuracy using different architectures with different features. Note that TS-STMAC (`BERT` + `SlowFast` + `RCNN`) is our proposed method. The methods using `BERT` for encoding a question outperformed those with `Glove` in many cases when using identical models and features. This result indicates that the difference from BERT provides better embeddings than Glove and can address the unknown words in a question. Moreover, TMAC (`BERT` + `ResNet`) outperformed SMAC (`BERT` + `RCNN`), and TS-STMAC (`BERT` + `SlowFast`+ `RCNN`) outperformed TS-TMAC (`BERT` + `SlowFast`+ `ResNet`) across all the datasets, indicating the superiority of the `RCNN` features in the VideoQA task that can represent detailed appearance information in video frames. Compared with TMAC (`BERT` + `SlowFast`), which used only motion features and SMAC (`BERT` + `RCNN`), which used detailed appearance features, TS-STMAC (`BERT` + `SlowFast` + `RCNN`) improved the performance in all cases. These results suggest that modeling both motion and detailed appearance features provides complementary effects.

4.3 Comparison with the State-of-the-Art

In this section, we compare our proposed method, TS-STMAC, to the existing state-of-the-art methods on short- and long-form VideoQA datasets. Because the number of instances in some question types is relatively small in some datasets [10], we report the number of instances of each question type in the overall VideoQA datasets. To compare our method to the existing ones, we used the reported accuracies of their original papers unless otherwise stated.

Method	MSVD-QA						MSRVTT-QA					
	What	Who	How	When	Where	All	What	Who	How	When	Where	All
	8,149	4,552	370	58	28	13,157	49,869	20,385	1,640	677	250	72,821
HME [10]	0.224	0.501	0.730	0.707	0.429	0.337	0.265	0.436	0.824	0.760	0.286	0.330
CAN [48]	0.211	0.479	**0.841**	**0.741**	**0.571**	0.324	0.267	0.434	**0.837**	0.753	0.352	0.332
MIN [26]	0.242	0.495	0.838	**0.741**	0.536	0.350	0.295	0.450	0.832	0.747	**0.424**	0.354
HCRN [31]	0.255	0.518	0.773	**0.741**	0.500	0.363	0.295	0.451	0.821	0.783	0.344	0.355
Ours: TS-STMAC	**0.337**	**0.569**	0.786	0.724	0.464	**0.432**	**0.336**	**0.488**	0.831	**0.786**	0.336	**0.394**

Table 2: Experimental results on MSVD-QA and MSRVTT-QA datasets: Number below each question type denotes number of QA pairs on the *test* split. Best result for each question type is marked in boldface.

Method	ActivityNet-QA									
	Motion	Spatial	Temporal	Yes/No	Color	Object	Location	Number	Other	All
	800	800	800	2,094	697	318	386	606	1,499	8,000
ESA [49]	0.125	0.144	0.025	0.594	0.298	0.142	0.259	0.446	0.284	0.318
HME [10]	0.174	0.159	0.023	0.607	0.304	0.132	0.277	0.475	0.297	0.331
CAN [48]	0.211	0.173	0.036	0.626	0.311	0.201	0.306	0.480	0.333	0.354
HCRN [31]	0.215	0.171	0.031	0.657	0.316	0.220	0.298	0.454	0.336	0.362
Ours: TS-STMAC	**0.355**	**0.183**	**0.039**	**0.683**	**0.364**	**0.258**	**0.316**	**0.500**	**0.376**	**0.402**

Table 3: Experimental results on ActivityNet-QA dataset. Best result for each question type is marked in boldface.

MSVD-QA Dataset: We show the VideoQA performance on MSVD-QA in Table 2 (left). We compared our TS-STMAC with the temporal reasoning models (HME [10], CAN [48], and HCRN [31]) and the spatiotemporal reasoning model (MIN [26]). HME, CAN, and HCRN mainly use the temporal information of video frames. MIN uses both the spatial and temporal information of the video. Our method significantly outperformed the existing ones and achieved an overall accuracy 0.432, which is 28.2% better than the prior best of HME, the temporal reasoning method (0.337). TS-STMACs performance is 19.0% better than HCRN, the latest temporal reasoning model (0.363). Our TS-STMAC is weaker than the existing methods on *how*, *when*, and *where* questions. This problem is caused by a class imbalance, where the number of instances of these questions is relatively small.

MSRVTT-QA Dataset: In Table 2 (right), we compared our method with HME, CAN, MIN, and HCRN on the MSRVTT-QA dataset. As in the MSVD-QA dataset, our method significantly outperformed the others on two major question types: *what* and *who*. Our method achieved the best overall accuracy of 0.394, which is 11.3% better than the spatiotemporal reasoning model, MIN (0.354), and 11.0% better than the temporal reasoning model, HCRN (0.355). From both the MSVD-QA and MSRVTT-QA results, our proposed method showed a high performance in the short-form QA dataset.

ActivityNet-QA Dataset: Next we report the performance on ActivityNet-QA, which is a long-form VideoQA dataset, unlike the MSVD-QA and MSRVTT-QA datasets. We compared our method with the original baseline model of this dataset, ESA, and the three latest temporal reasoning models: HME, CAN, and HCRN. Because the HME and HCRN results have not been published yet, we applied both to ActivityNet-QA with default parameters based on their public code. Table 3 summarizes the experimental results of nine question types on ActivityNet-QA. Our proposed method outperformed the other methods and achieved the best accuracy of 0.402, which is 11.0% better than the best of the temporal reasoning model, HCRN (0.362). Our method also outperformed the others on all the question types. In particular, it improved its performance by 65.1% compared to HCRN on

Figure 4: Visualization of typical examples by TS-STMAC network: We visualize the spatial attentions of objects with colored regions and attending words in a question at each reasoning step. Regions with higher spatial attention values are shown brighter. The more attending words are shown with darker color.

motion questions on the human activities in the video. Also, our method improved its performance by 17.2% compared to HCRN on *object* questions in the video. These results indicate the effectiveness of using a powerful spatiotemporal reasoning model that combines detailed appearance and motion features.

4.4 Qualitative Results

Finally, we demonstrate how multi-step spatiotemporal reasoning works by visualizing examples. Fig. 4 shows some typical examples from the reasoning process of the TS-STMAC network. We selected the frames based on a score, which is the product of the temporal attention to a frame and the top five spatial attentions on the regions at each reasoning step. We also show words that receive more attention from the controller unit. The results show the cell tends to identify relevant frames and regions through multi-step reasoning, suggesting that our method effectively incorporated the spatial and temporal features as well as textual information into VideoQA.

5 Conclusion

We proposed a new spatiotemporal video reasoning method for VideoQA. We devised a two-stream spatiotemporal MAC (TS-STMAC) cell to model the relationships between the spatial and temporal structures of a video as well as the textual information of questions. Then we proposed a TS-STMAC network that sequentially applies the TS-STMAC cell for multi-step reasoning. We evaluated our approach on three VideoQA datasets: MSVD-QA, MSRVTT-QA, and ActivityNet-QA. Our qualitative and quantitative results showed the usefulness of both the spatial and temporal reasoning modules and the multi-step iterations in the reasoning.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR15E2, Japan.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[4] Guan-Lin Chao, Abhinav Rastogi, Semih Yavuz, Dilek Hakkani-Tur, Jindong Chen, and Ian Lane. Learning question-guided video representation for multi-turn video question answering. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 215–225, 2019.

[5] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[6] Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim. TutorialVQA: Question answering dataset for tutorial videos. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 5450–5455, 2020.

[7] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2018.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

[9] Chenyou Fan. EgoVQA - An egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[10] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1999–2007, 2019.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019.

[12] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6576–6585, 2018.

[13] Lianli Gao, Pengpeng Zeng, Jingkuan Song, Yuan-Fang Li, Wu Liu, Tao Mei, and Heng Tao Shen. Structured two-stream attention network for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 6391–6398, 2019.

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, 2017.

[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[19] Deng Huang Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11021–11028, 2020.

[20] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[21] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019.

[22] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2758–2766, 2017.

[23] Jingwei Ji, Ranjay Krishna, Fei-Fei Li, and Juan Carlos Niebles. Action genome: Actions as composition of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10236–10247, 2020.

[24] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11101–11108, 2020.

[25] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11109–11116, 2020.

[26] Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. Multi-interaction network with object relation for video question answering. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 1193–1201, 2019.

[27] Weike Jin, Zhou Zhao, Yimeng Li, Jie Li, Jun Xiao, and Yueting Zhuang. Video question answering via knowledge-based progressive spatial-temporal attention network. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15 (2s):1–22, July 2019.

[28] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017.

[29] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2989–2998, 2017.

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1): 3273, 2017.

[31] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[32] Xiangpeng Li, Lianli Gao, Xuanhan Wang, Wu Liu, Xing Xu, Heng Tao Shen, and Jingkuan Song. Learnable aggregating net with diversity learning for video question answering. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 1166–1174, 2019.

[33] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8658–8665, 2019.

[34] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1682–1690, 2014.

[35] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 677–685, 2017.

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

[37] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 91–99. 2015.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[40] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 239–247, 2018.

[41] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[42] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018.

[43] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144, 2016.

[44] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 1645–1653, 2017.

[45] H. Xue, W. Chu, Z. Zhao, and D. Cai. A better way to attend: Attention with trees for video question answering. *IEEE Transactions on Image Processing*, 27(11): 5563–5574, Nov 2018.

[46] Tianhao Yang, Zheng-Jun Zha, Hongtao Xie, Meng Wang, and Hanwang Zhang. Question-aware tube-switch network for video question answering. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 1184–1192, 2019.

[47] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video question answering via attribute-augmented attention network learning. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 829–832, 2017.

[48] Ting Yu, Jun Yu, Zhou Yu, and Dacheng Tao. Compositional attention networks with two-stream fusion for video question answering. *IEEE Transactions on Image Processing*, 29:1204–1218, 2020.

[49] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 9127–9134, 2019.

[50] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 4334–4340, 2017.

[51] Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, and Yueting Wu, Fei and; Zhuang. Frame augmented alternating attention network for video question answering. *IEEE Transactions on Multimedia*, 22(4), 2020.

[52] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jingkuan Song, and Xiaofei He. Open-ended long-form video question answering via hierarchical convolutional self-attention networks. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 4383–4389, 2019.

[53] Z. Zhao, Z. Zhang, S. Xiao, Z. Xiao, X. Yan, J. Yu, D. Cai, and F. Wu. Long-form video question answering via dynamic hierarchical reinforced networks. *IEEE Transactions on Image Processing*, 28(12):5939–5952, Dec 2019.

[54] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3518–3524, 2017.

[55] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3683–3689, 2018.

[56] Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao. Multichannel attention refinement for video question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(1s):1–23, March 2020.