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 a video question answering (VideoQA), which requires spatiotemporal video reasoning. VideoQA aims to return an appropriate answer about textual questions referring to image frames in the video. In this paper, based on the observation that multiple entities and their movements in the video can be important clues for deriving the correct answer, we propose a two-stream spatiotemporal compositional attention network that achieves sophisticated multi-step spatiotemporal reasoning by 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 detailed appearance features of multiple entities as well as motion features guided by attending words in the textual question. Furthermore, it progressively refines internal representation and infers the answer via 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 according to the textual questions posed about visual content in the video. Using this technology, we can quickly understand the real-world events and situations in videos through natural language. Thereby, VideoQA technology plays 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 the embodied agent with visual perception [7].

In contrast to traditional visual question answering for static images [2, 14, 34], VideoQA is a more challenging task because the VideoQA system has to find relevant frames to a question and answer out of 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 video with a pre-trained convolutional networks (ConvNets)

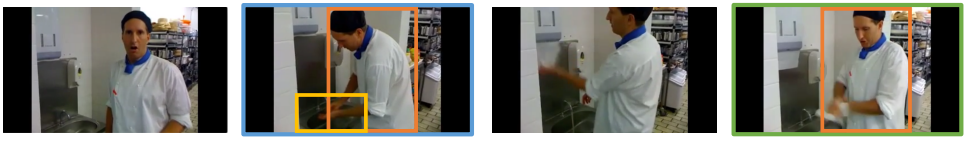

Figure 1: VideoQA example: Q1 can be correctly answered by finding a frame (or clip) from the video containing the entity in question and the motions associated with the answer. Q2 can be correctly answered by finding entities in the image frames related to the question and its answer.

model [16, 39] and a 3D ConvNets [15, 41], and then apply learnable soft weights (i.e., attention mechanism [3]) to them for capturing frame and clip-level details relevant to a given question [44, 52, 55]. Their limitation is the 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]. Moreover, some notable works use multi-step reasoning that gradually refines the motion-appearance representations of video and question representation [10, 12, 46, 48]. These multi-step video reasoning approaches achieved a competitive performance 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 events occurred in the video that involve multiple entities (e.g., humans and objects) [23, 42], these methods fail to capture the associations between region-level details of entities in the frame and their corresponding question words. As described in the examples in Figure 1, to get the right answer for VideoQA, the detailed appearance information of entities in the frame is an important clue as well as the motion information over frames.

Motivated by this observation, we develop a two-stream spatiotemporal MAC network (TS-STMAC), which performs sequential spatiotemporal reasoning on video frames according to the question content. Moreover, we use a SlowFast model that shows high performance in video understanding tasks [11] and a bottom-up attention model known to be 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 devise a two-stream spatiotemporal MAC cell, a new neural module containing a spatiotemporal attention mechanism that simultaneously finds motion features and detailed appearance features of entity's regions relevant to attending words in a question. We use it as a building block of our VideoQA framework, recurrently apply it for multi-step reasoning, and progressively infer the correct answer. Through this question-aware multi-step spatiotemporal reasoning, the model can focus on the important frames and regions ignoring useless information.

In summary, 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 to answer the question. Third, we conduct experiments on the short- and long-form VideoQA datasets to validate our method's effectiveness and show that our method outperforms state-of-the-art

approaches by a large margin on three public benchmarks.

2 Related Work

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

In contrast with modeling frame-level temporal dynamics of video, spatiotemporal reasoning approaches that focus on the frame- and region-level visual content relevant to a question are relatively less explored. Traditional approaches use a combination of recurrent neural networks (RNN) and ConvNets, which encode spatiotemporal video features and a textual question, and then jointly learn their multi-modal representations [22, 54]. However, these works lack modeling the interaction between question words and visual contents. Some words in the question often indicate the entities in the video, which can be important clues for video reasoning. To further improve the VideoQA performance, the QA model has to attend words in the question corresponding 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 not enough to capture the movement in the video, which is essential for questions about the motion 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 the word-level attention plays an important role to find frames representing motion information and image regions representing detailed appearance information relevant to a question. Our work differs in that the proposed neural module can simultaneously attend 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 answer though multi-step video reasoning to focus on important video information. We demonstrate that our sophisticated method outperforms existing temporal or spatiotemporal reasoning methods on the long-form VideoQA dataset as well as short ones.

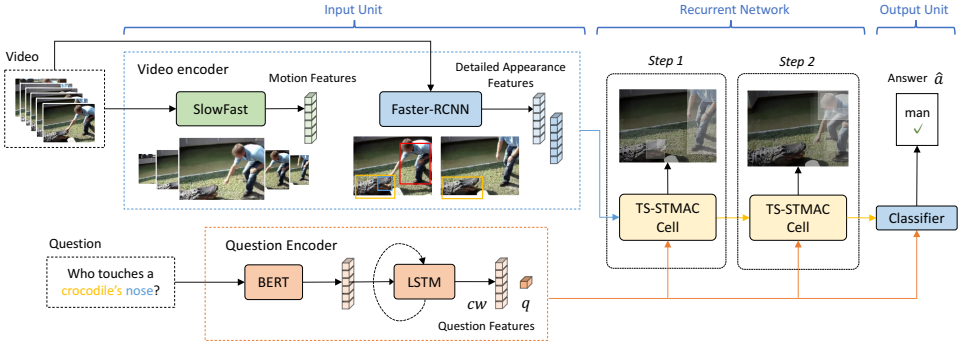

Figure 2: Illustration of our two-stream spatiotemporal MAC (TS-STMAC) network. First, video encoder extracts motion and detailed appearance features from short clips and frames using SlowFast networks and Faster-RCNN (top left). Question encoder extracts text features from question words using BERT 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 attending to frames (or clips) and regions relevant to the question. The network repeats this process multiple times to progressively refined the internal representation. Finally, the classifier predicts the final answer using the 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 in brighter.

3 Approach

3.1 Problem Definition

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

Video Embeddings. The video consists of the sequence of frames which 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 [1]. We extract the motion feature ($\boldsymbol{f}_t^a \in \mathbb{R}^{2304}$) from the t-th clip and use a series of motion features $f^a = \{\boldsymbol{f}_t^a\}_{t=1}^T$ for representing the video, where T is the number of clips. For detailed appearance information, we extract region features ($f_t^b = \{\boldsymbol{f}_{i,t}^b\}_{i=1}^N$) from the t-th frame using Faster R-CNN [38] trained with the Visual Genome dataset [30], where each $\boldsymbol{f}_i^b \in \mathbb{R}^{2048}$ corresponds to a region feature of an entity, and N is the number of detected entities with the highest confidence scores. Following the past VQA work [1], we set $N = 36$. We 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 input of VideoQA model 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 do not in the test data, we first split a question into words with a length of M by the Word Piece tokenizer [43]. We extract a feature vector from the last layer of a pre-trained 12-layer BERT model for each word. Note that we fine-tune this layer during VideoQA training. Then, we encode the question using a one-layer bi-directional LSTM (biLSTM) [18], which is used for guiding the model's

multi-step reasoning. We use a series of output states from LSTM $\{\boldsymbol{cw}_i\}_{i=1}^{M}$ as contextual question word embeddings. We also use $\boldsymbol{q} \in \mathbb{R}^{2d}$ as a question sentence embedding, which is represented by the concatenation of the final hidden states from the backward and forward LSTMs. Furthermore, we apply a linear transformation to $\boldsymbol{q}$ for representing a step-aware question embedding $\boldsymbol{q}_i \in \mathbb{R}^d$ at i^{th} reasoning step.

3.2 Two-Stream Spatiotemporal MAC Network

For VideoQA, we develop a two-stream spatiotemporal MAC (TS-STMAC) network that consists of an input unit, a core recurrent network, and an output unit. Figure 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 reasons over the question by decomposing it into a series of operations (control) that retrieve information from the video (clip- and frame region-level features) and aggregate the results into internal memory. As the core recurrent network, we repeatedly use the following TS-STMAC cells at each step.

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

MAC Cell: The MAC cell is a neural module designed to apply attention-based operations to perform reasoning. The cell holds two hidden states at i-th step: control $\boldsymbol{c}_i \in \mathbb{R}^d$ and memory $\boldsymbol{m}_i \in \mathbb{R}^d$. The control state $\boldsymbol{c}_i$ stores the information on the reasoning operation that should be performed. The memory $\boldsymbol{m}_i$ state has the intermediate result that has been computed in the recurrent reasoning process. The MAC cell updates the control and memory states for each reasoning step $i = 1, \ldots, S$ using three internal units: control, read, and write units. The MAC cell iteratively aggregates information from some knowledge source according to the control state in the following steps. (i) The control unit attends some words of the question by using attention mechanism [3] and updates the control state $\boldsymbol{c}_i$. (ii) The read unit attends to some parts of a knowledge base $\{\boldsymbol{k}\}_{i=1}^{K}$ (e.g., image features for VQA) and retrieves information $\boldsymbol{r}_i$ from them according the current control and previous memory states $\boldsymbol{c}_i$ and $\boldsymbol{m}_{i-1}$, where K denotes the size of knowledge base. (iii) The write unit updates the memory based on the retrieved information $\boldsymbol{r}_i$ and previous memories $\{\boldsymbol{m}_0, \ldots \boldsymbol{m}_{i-1}\}$. The equations of the reasoning step in the MAC cell are shown as follows:

$$\boldsymbol{c}_i = \text{ControlUnit}(\boldsymbol{c}_{i-1}, \{\boldsymbol{cw}_j\}_{j=1}^{M}, \boldsymbol{q}_i) \tag{1}$$

$$\boldsymbol{r}_i = \text{ReadUnit}(\boldsymbol{m}_{i-1}, \{\boldsymbol{k}_j\}_{j=1}^{K}, \boldsymbol{c}_i) \tag{2}$$

$$\boldsymbol{m}_i = \text{WriteUnit}(\{\boldsymbol{m}_{j-1}\}_{j=1}^{i}, \boldsymbol{r}_i, \boldsymbol{c}_i) \tag{3}$$

Due to the space limitation, see the work in [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, the normal MAC cell can only handle one of them. To address this issue, we extend this MAC cell and create 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 spatial and temporal MAC cells. The temporal MAC cell is used for representing the temporal structure of the video. We use motion features of clips in the video $\{\boldsymbol{f}_j^a\}_{j=1}^{T}$ as the input of this cell. The

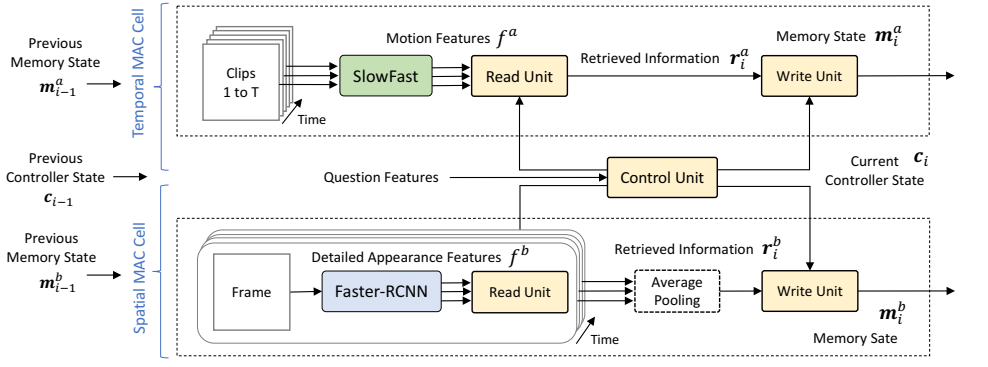

Figure 3: Overview of our two-stream spatiotemporal MAC (TS-STMAC) cell, which consists of two temporal and spatial MAC cells. The temporal MAC cell (top) takes motion features as input and updates its internal representation $\boldsymbol{m}^a$ that holds temporal information over clips based on the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

temporal MAC cell updates the controller and memory states based on the motion features. As with a standard MAC cell, it is given by

$$\boldsymbol{c}_i = \text{ControlUnit}(\boldsymbol{c}_{i-1}, \{\boldsymbol{cw}_j\}_{j=1}^{M}, \boldsymbol{q}_i) \tag{4}$$

$$\boldsymbol{r}_i^a = \text{ReadUnit}_{\text{temporal}}(\boldsymbol{m}_{i-1}^a, \{\boldsymbol{f}_t^a\}_{t=1}^{T}, \boldsymbol{c}_i) \tag{5}$$

$$\boldsymbol{m}_i^a = \text{WriteUnit}_{\text{temporal}}(\{\boldsymbol{m}_{j-1}^a\}_{j=1}^{i}, \boldsymbol{r}_i^a, \boldsymbol{c}_i) \tag{6}$$

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

The spatial MAC cell is used for representing the spatial structure of the video frames. This cell 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 $\boldsymbol{r}_{i,t}^b$ from region features $\{\boldsymbol{f}_{j,t}^b\}_{j=1}^{N}$ of t^{th} frame selectively focusing on specific regions based on the control state $\boldsymbol{c}_i$:

$$\boldsymbol{r}_{i,t}^b = \text{ReadUnit}_{\text{spatial}}(\boldsymbol{m}_{i-1}^b, \{\boldsymbol{f}_{j,t}^b\}_{j=1}^{N}, \boldsymbol{c}_i), \tag{7}$$

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

$$\boldsymbol{r}_i^b = \text{pool}(\{\boldsymbol{r}_{i,1}^b, \boldsymbol{r}_{i,2}^b, \cdots, \boldsymbol{r}_{i,T}^b\}) \tag{8}$$

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

$$\boldsymbol{m}_i^b = \text{WriteUnit}_{\text{spatial}}(\boldsymbol{m}_{i-1}^b, \boldsymbol{r}_i^b, \boldsymbol{c}_i) \tag{9}$$

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

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

Output Unit: To compute the final answer, we use a simple classifier using the question and the final memory states of the spatial and temporal MAC cells after applying S times cell computation as input:

$$\boldsymbol{o}' = \boldsymbol{W}_1[\boldsymbol{q}; \boldsymbol{m}_S^a; \boldsymbol{m}_S^b] + \boldsymbol{b}_1, \quad \boldsymbol{o} = \text{softmax}(\text{ELU}(\boldsymbol{W}_2\boldsymbol{o}' + \boldsymbol{b}_2)) \tag{10}$$

where $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{b}_1$, and $\boldsymbol{b}_2$ are the learnable parameters, ELU is an exponential linear unit [5]. The final output of the classifier is given by

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

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 used MSVD-QA [44], MSRVTT-QA [44], and ActivityNet-QA [49] datasets for evaluation. MSVD-QA and MSRVTT-QA are short-form VideoQA datasets. The average lengths of videos used in these datasets are 10 and 15 sec, respectively. Both MSVD-QA and MSRVTT-QA and include five different question types (*What*, *Who*, *How*, *When*, and *Where*). In contrast, ActivityNet-QA is a more challenging VideoQA dataset that uses long videos about human activities. The average length of the videos is 116 sec. The videos are sampled from the ActivityNet dataset [17]. ActivityNet-QA includes four main question types (*Motion*, *Spatial Relationship*, *Temporal Relationship*, and *Free*). Furthermore, the *Free* questions are divided into six sub-question types (*Yes/No*, *Number*, *Color*, *Object*, *Location*, and *Other*) according to their answer types. We sampled 20 frames at equal intervals for appearance feature extraction and 20 clips for motion feature extraction. For answer candidates, we selected the top 1,000 most frequent answers in a training split.

Implementation Details. We trained our method up to 100 epochs using AMSGrad [37] variant of Adam [19] for optimization, with a learning rate of $\alpha = 10^{-4}$ and a batch size of 32. We employed the early stopping if the validation accuracy does not increase for ten epochs. We converted the words in the question and answer to lower cases. We set the dimension d of the TS-STMAC cell as 256. For the multi-step reasoning of the TS-STMAC network, two reasoning steps ($S = 2$) were used following the average performance on validation data across three VideoQA datasets. We also used self-attention connections between the cells.

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

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. The best result for each dataset is marked by boldface.

addition to the proposed TS-STMAC, we prepared its variants temporal MAC (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 can be seen as 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 consider 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 the Glove ones ($\in \mathbb{R}^{300}$) and were initialized with the 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 feature extracted from SlowFast networks.

Table 1 shows accuracy using different architectures with different features. Note that TS-STMAC (BERT + SlowFast + RCNN) is our proposed method. The results show that the methods using BERT for encoding a question outperformed ones with Glove in many cases when using the same models and features. It indicates that the difference comes from BERT is 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 datasets indicating the superiority of RCNN features in the VideoQA task that can represent the 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 have complementary effects.

4.3 Comparison with the State-of-the-Art

In this section, we compare the proposed method TS-STMAC to existing state-of-the-art methods on short- and long-form VideoQA datasets. Because the number of instances in some question types are relatively small in some datasets [10], we report the number of instances of each question type overall VideoQA datasets. To compare our method to the existing ones, we used reported accuracies of their original paper 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 [28]	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. The number below each question type denotes the number of QA pairs on the *test* split. The best result for each question type is marked by 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 [29]	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 [28]	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 the ActivityNet-QA dataset. The best result for each question type is marked by boldface.

MSVD-QA Dataset: We show the VideoQA performance on MSVD-QA in Table 2 (left). We compared our method TS-STMAC with the temporal reasoning models (HME [10], CAN [28], and HCRN [31]) and the spatiotemporal reasoning model (MIN [26]). HME, CAN, and HCRN mainly use temporal information of video frames. MIN uses both spatial and temporal information of the video. We found our method significantly outperformed existing ones, and achieved overall accuracy 0.432, which is 28.2% better than the prior best of temporal reasoning method, HME (0.337). Moreover, the performance of TS-STMAC is 19.0% better than the latest temporal reasoning model HCRN (0.363). Our TS-STMAC is weaker than existing methods on *How*, *When*, *Where* questions. However, this is due to the class imbalance, where the number of instances on 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 is 11.0% better than the temporal reasoning model HCRN (0.355). From the results on both MSVD-QA and MSRVTT-QA results, we found that the proposed method shows high performance in the short-form QA dataset.

ActivityNet-QA Dataset: We report the performance on ActivityNet-QA, which is a long-form VideoQA dataset, unlike MSVD-QA and MSRVTT-QA datasets. We compared our method with the original baseline model of this dataset, ESA, and three latest temporal reasoning models (HME, CAN, and HCRN). Because the results of HME and HCRN have not come out yet, we apply HME and HCRN 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 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). Moreover, our method outperformed others on all question types. In par-

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

ticular, our method improved 65.1% performance comparing to HCRN on *Motion* questions that ask about the human activities in the video. Also, our method improved 17.2% performance comparing to HCRN on *Object* questions that ask about objects in the video. The results indicate the effectiveness of using a powerful spatiotemporal reasoning model with the combination of the detailed appearance and motion features.

4.4 Qualitative Results

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

5 Conclusion

In this paper, we proposed a new spatiotemporal video reasoning method for VideoQA. We devise a two-stream spatiotemporal MAC (TS-STMAC) cell to model the relationships between spatial and temporal structures of video as well as textual information of question. Then we proposed the TS-STMAC network that sequentially applies the TS-STMAC cell for multi-step reasoning. We evaluate our approach on three VideoQA datasets: MSVD-QA, MSRVTT-QA, and ActivityNet-QA. The qualitative and quantitative results showed the usefulness of both 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):32–73, 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.