Video Question Answering via Hierarchical Spatio-Temporal Attention Networks

Zhou Zhao[1], Qifan Yang[1], Deng Cai[2], Xiaofei He[2] and Yueting Zhuang[1]
[1]College of Computer Science, Zhejiang University
[2]State Key Lab of CAD&CG, Zhejiang University
{zhaozhou,yzhuang}@zju.edu.cn, {yqf.init,dengcai,xiaofeihe}@gmail.com

Abstract

Open-ended video question answering is a challenging problem in visual information retrieval, which automatically generates the natural language answer from the referenced video content according to the question. However, the existing visual question answering works only focus on the static image, which may be ineffectively applied to video question answering due to the lack of modeling the temporal dynamics of video contents. In this paper, we consider the problem of open-ended video question answering from the viewpoint of spatio-temporal attentional encoder-decoder learning framework. We propose the hierarchical spatio-temporal attention network for learning the joint representation of the dynamic video contents according to the given question. We then develop the spatio-temporal attentional encoder-decoder learning method with multi-step reasoning process for open-ended video question answering. We construct a large-scale video question answering dataset. The extensive experiments show the effectiveness of our method.

1 Introduction

Visual information retrieval (VIR) is the visual information delivery mechanism that enables users to post their queries and obtain the relevant information in visual media. Visual question answering is the essential problem in VIR sites, which automatically returns the relevant answer from the reference visual content according to the user's given question. Most of the existing works consider the problem of open-ended visual question answering as the multimodal understanding task, which learn the joint representation from the multimodal features of the given visual content and textual question, and then generate the natural language answer. However, the existing works mainly tackle the problem of static image question answering [Antol *et al.*, 2015; Shih *et al.*, 2016; Lu *et al.*, 2016; Li and Jia, 2016; Xiong *et al.*, 2016]. Although the existing works have achieved promising performance in image question answering, they may still be ineffectively applied to video question answering due to the lack of modeling the temporal dynamics of video content.

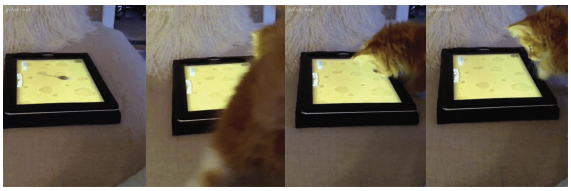


Figure 1: Open-ended Video Question Answering.

The video contents often contain the complex interactions of the targeted objects to the given question that evolves over time [Yao *et al.*, 2015]. Thus, the simple extension of existing image question answering methods based on the single temporally collapsed video representation is likely to generate unsatisfactory answers. We illustrate a simple example of open-ended video question answering in Figure 1. We show that the answer generation to the question "what is the cat doing?" requires the collective information from multiple video frames. Recently, temporal attention mechanism has been proposed to extract the critical frame information across the entire video for representation learning [Yao *et al.*, 2015]. We then employ the temporal attention mechanism to model the collective information of video content for video question answering. On the other hand, the sequential order of the frames is also important for video representation [Fernando and Gould, 2016]. Thus, leveraging the sequential complex interactions of the targeted objects according to the question from the collective frame-level video representation is critical for modeling the temporal dynamics of video content in video question answering.

In this paper, we present the problem of open-ended video question answering from the viewpoint of spatio-temporal attentional encoder-decoder learning framework. We propose the hierarchical spatio-temporal attention networks that jointly learn the representation of the sequentially critical frames with the targeted objects according to the question. We then develop the encoder-decoder learning framework that enables the joint representation learning of the multimodal spatio-temporally attentional video and textual question through multiple reasoning steps for open-ended video

question answering, named as r-STAN. When a certain question is issued, r-STAN can generate natural language answer for it based on the reference video content. The main contributions of this paper are as follows:

- Unlike the previous studies, we study the problem of open-ended video question answering from the viewpoint of spatio-temporal attentional encoder-decoder learning framework. We propose the spatio-temporal attention networks that learn the joint representation from the critical video frames of the targeted objects according to the question.
- We incorporate the multi-step reasoning process for the proposed spatio-temporal attention networks to enable the progressive joint representation learning of the multimodal spatio-temporal attentional video and textual question to further improve the performance of video question answering.
- We construct a large-scale dataset for open-ended video question answering and validate the effectiveness of our proposed method through extensive experiments.

2 Video Question Answering via Spatio-Temporal Attention Networks

In this section, we study the problem of open-ended video question answering from the viewpoint of spatio-temporal attentional encoder-decoder learning framework. We first develop the spatio-temporal attentional encoder networks with multi-step reasoning process to learn the joint representation of multimodal saptio-temporal attentional video and textual question progressively. We then devise the recurrent decoder network to generate the natural language answer for open-ended video question answering.

Before presenting the learning framework, we first introduce some basic notions and terminologies. We denote the question by $\mathbf{q} \in Q$, the video by $\mathbf{v} \in V$ and the answer by $\mathbf{a} \in A$, where Q, V and A are the sets of questions, videos and answers, respectively. Since the video is composed of sequential frames, the frame-level representation of video $\mathbf{v}$ is given by $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N)$ of length N, where $\mathbf{v}_2$ is the second frame. We then denote the word-level representation of natural language answer by $\mathbf{a}$ by $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M)$ of length M, where $\mathbf{a}_M$ is the Mth word token. Since both the video and answer are sequential data with variant length, it is natural to choose the variant recurrent neural network called gated recurrent unit (GRU) [Chung *et al.*, 2014] to learn the feature representation by:

$$\mathbf{r}_t = \delta(\mathbf{W}_{xr}\mathbf{x}_t + \mathbf{W}_{hr}\mathbf{h}_{t-1} + \mathbf{b}_r), \quad (1)$$

$$\mathbf{z}_t = \delta(\mathbf{W}_{xz}\mathbf{x}_t + \mathbf{W}_{hz}\mathbf{h}_{t-1} + \mathbf{b}_z), \quad (2)$$

$$\tilde{\mathbf{h}}_t = tanh(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}(\mathbf{r}_t \odot \mathbf{h}_t) + \mathbf{b}_h), \quad (3)$$

$$\mathbf{h}_t = \mathbf{z}_t \odot \tilde{\mathbf{h}}_t + (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1}, \quad (4)$$

where $\mathbf{x}_t$ and $\mathbf{h}_t$ are the input and output vectors, $\mathbf{z}_t$ and $\mathbf{r}_t$ are the update and reset gate vectors, $\mathbf{W}$s and $\mathbf{b}$s are the parameter matrices and bias vector. We note that gate vector $\mathbf{z}_t$ in Equation (2) is the trade-off parameter for updating hidden state $\mathbf{h}_t$ from the previous state $\mathbf{h}_{t-1}$ and the currently estimated one $\tilde{\mathbf{h}}_t$. Specifically, we learn the sequential feature representation of both video and answer by directional GRU, which consists of a forward GRU and a backward GRU. The backward GRU has the same network structure with the forward one while its input sequence is reversed. We denote the hidden state of the forward GRU for the video at time t by $\mathbf{h}_t^f$, and the hidden state of the backward GRU by $\mathbf{h}_t^b$. Thus, the tth hidden state of video $\mathbf{v}$ from the bidirectional GRU layer is denoted by $\mathbf{h}_t = [\mathbf{h}_t^f, \mathbf{h}_t^b]$, and the hidden states of video $\mathbf{v}$ is given by $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N)$.

Using the notations above, the problem of open-ended video question answering is formulated as follows. Given the set of videos V, questions Q and answers A, our goal is to learn the encoder-decoder network model $g(f(\mathbf{v}, \mathbf{q}))$ where the encoder network $f(\mathbf{v}, \mathbf{q})$ that learns the joint representation of the video and question, and the decoder network $\hat{\mathbf{a}} = g(f(\mathbf{v}, \mathbf{q}))$ generates the answer $\hat{\mathbf{a}}$ for open-ended video question answering. We present the details of the spatio-temporal attention network learning framework in Figure 2.

2.1 Hierarchical Spatio-Temporal Attentional Encoder Network Learning

In this section, we propose the encoder neural network $f(\cdot)$ to learn the joint representation of video and question with hierarchical spatio-temporal attention and multiple reasoning updates for the problem of open-ended video question answering.

Inspired by attention mechanism [Xu *et al.*, 2015], we propose the hierarchical spatio-temporal attention networks to learn the joint representation from the relevant frames of the targeted regions according to the question. Since the global representation of the frame may fail to capture all necessary information for answering the question [Li and Jia, 2016], it is natural to choose the spatial attention model to automatically localize the targeted regions in each frame according to the question. Following the existing spatial attention model [Li and Jia, 2016], we employ the object generator to produce a set of candidate regions that are most likely to be an object. The frame representation is thus given by the set of the candidate region features and the whole frame region feature, denoted by $F = \{F_1, F_2, \ldots, F_N\}$. The $F_j = \{\mathbf{f}_{j1}, \mathbf{f}_{j2}, \ldots, \mathbf{f}_{jK}\}$ is the feature set of the j-th frame, where $\mathbf{f}_{j1}, \mathbf{f}_{j2}, \ldots, \mathbf{f}_{j(K-1)}$ are the candidate region features and $\mathbf{f}_K$ is the whole frame region feature. Given question $\mathbf{q}$ and the region feature of j-th frame representation $\mathbf{f}_{ji} \in F_j$, the spatial attention score $s_{ji}^{(s)}$ is given by

$$s_{ji}^{(s)} = \mathbf{w}^{(s)} tanh(\mathbf{W}_{qs}\mathbf{q} + \mathbf{W}_{fs}\mathbf{f}_{ji} + \mathbf{b}_s), \quad (5)$$

where $\mathbf{W}_{qs}$ and $\mathbf{W}_{fs}$ are parameter matrices and $\mathbf{b}_s$ is bias vector. For each region $\mathbf{f}_{ji}$, the activations in spatial dimension by the softmax function is given by $\alpha_{ji} = \frac{\exp(s_{ji}^{(s)})}{\sum_i \exp(s_{ji}^{(s)})}$, which is the normalization of the spatial attention score. The spatially attended frame representation is then given by $\mathbf{v}_j^{(s)} = \sum_i \alpha_{ji}\mathbf{f}_{ji}$.

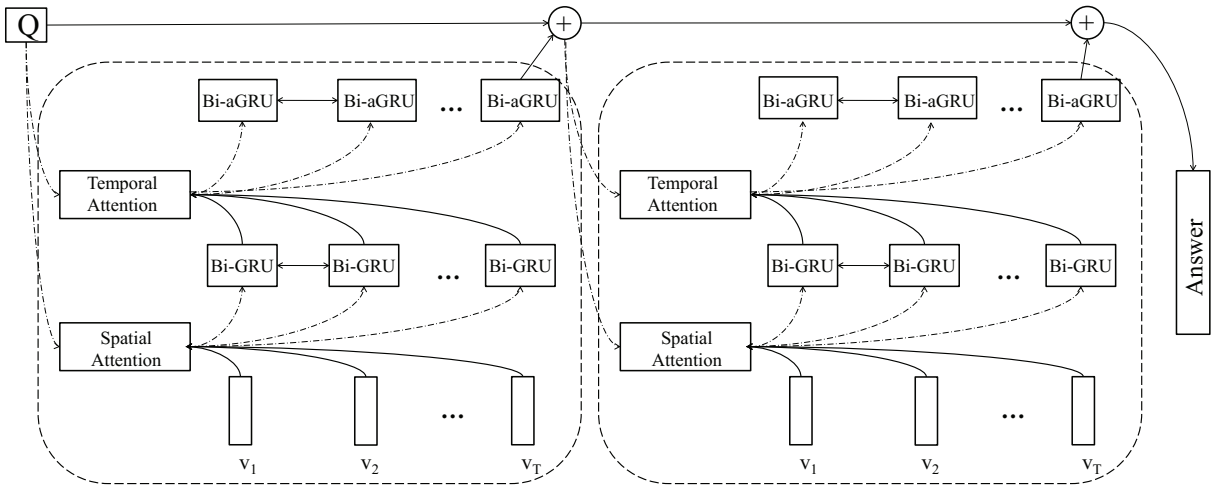


Figure 2: The Overview of Open-Ended Video Question Answering via Hierarchical Spatial-Temporal Attentional Encoder-Decoder Learning Framework (r-STAN in case of r = 2). The hierarchical spatio-temporal attentional encoder networks learn the joint representation of multimodal spatio-temporal attentional video and textual question with multiple reasoning steps, and the recurrent decoder network generates the natural language answer for open-ended video question answering.

On the other hand, a number of frames in the video are redundant and irrelevant to the question. Thus, it is important to localize the relevant video frames with the targeted information according to the question. We thus introduce the temporal attention model to estimate the relevance of video frames according to the question. Given the spatially attended video frames $\mathbf{v}^{(s)} = (\mathbf{v}_1^{(s)}, \mathbf{v}_2^{(s)}, \ldots, \mathbf{v}_N^{(s)})$, we learn their latent state representation from bidirectional GRU layer by $\mathbf{h}^{(s)} = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_N^{(s)})$. Then, their relevance scores based on temporal attention mechanism [Xu *et al.*, 2015] is given by

$$s_j^{(t)} = \mathbf{w}^{(t)} tanh(\mathbf{W}_{qt}\mathbf{q} + \mathbf{W}_{ht}\mathbf{h}_j^{(s)} + \mathbf{b}_t), \quad (6)$$

where $\mathbf{W}_{qt}$ and $\mathbf{W}_{ht}$ are parameter matrices and $\mathbf{b}_t$ is bias vector. For the latent state of each frame $\mathbf{h}_j^{(s)}$, its activation in temporal dimension by the softmax function is denoted by $\beta_j = \frac{\exp(s_j^{(t)})}{\sum_j \exp(s_j^{(t)})}$, which is the normalization of the temporal attention scores (i.e., $\beta_j \in (0, 1)$). Inspired by the attentional gate [Kumar and Irsoy, 2015], we employ the attentional GRU networks to learn the order-sensitive representation of the spatio-temporally attended, denoted as aGRU network. The inputs to the aGRU network are the latent state of spatially attended frames by bidirectional GRU layer $\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_N^{(s)}$, and the estimated temporal attention scores $\beta_1, \beta_2, \ldots, \beta_N$. The current estimated state $\tilde{\mathbf{h}}_j^{(t)}$ is obtained based on the input $\mathbf{h}_j^{(s)}$. The aGRU network then updates its hidden state $\mathbf{h}_j^{(t)}$ based on the mixture of current estimated state $\tilde{\mathbf{h}}_j^{(t)}$ and its previous state $\mathbf{h}_{j-1}^{(t)}$. Unlike the update rule of GRU in Equation (4), the aGRU updates the current state $\mathbf{h}_j^{(t)}$, given by

$$\mathbf{h}_j^{(t)} = \beta_j \odot \tilde{\mathbf{h}}_j^{(t)} + (1 - \beta_j) \odot \mathbf{h}_{j-1}^{(t)}, \quad (7)$$

where the update gate vector is set to the normalized temporal attention score (i.e., $\mathbf{z}_t = \beta_t$). Therefore, spatio-temporal attentional representation of video $\mathbf{v}$ according to question $\mathbf{q}$ is given by $h_{\mathbf{q}}^{sp}(\mathbf{v}) = \mathbf{h}_N^{(t)}$, where $\mathbf{h}_N^{(t)}$ is the last hidden state of the aGRU networks.

We then incorporate the multi-step reasoning process [Sukhbaatar *et al.*, 2015] for the proposed spatio-temporal attention networks to further improve the performance of open-ended video question answering. Given spatio-temporal network $h^{sp}(\cdot)$, video $\mathbf{v}$ and question $\mathbf{q}$, the spatio-temporal attention network learning with multi-step reasoning process is given by:

$$\begin{aligned} \mathbf{y}_r &= \mathbf{y}_{r-1} + h_{\mathbf{y}_{r-1}}^{sp}(\mathbf{v}), \\ \mathbf{y}_0 &= \mathbf{q}, \end{aligned}$$

where is recursively updated. The joint representation of spatio-temporal attentional video is returned after the R-th update, given by $f(\mathbf{q}, \mathbf{v}) = \mathbf{y}_R$. The learning process of reasoning spatio-temporal attention networks is illustrated in Figure 2.

We now present the decoder neural network $g(\cdot)$ based on the joint representation of spatio-temporal attentional video for answer prediction. Unlike the encoder neural network which simply encodes the spatio-temporal attentional video representation according to the question, we decoder neural network is learned to generate the answer. At each time j, the decoder computes the probability of generating k-th word by

$$p(a_{j,k} = 1 | \mathbf{a}_1, \ldots, \mathbf{a}_{j-1}, \mathbf{y}_R) = \frac{\exp(\mathbf{w}_{ky}^{(o)}\mathbf{y}_R + \mathbf{w}_{kh}^{(0)}\mathbf{h}_j^{(0)})}{\sum_k \exp(\mathbf{w}_{ky}^{(o)}\mathbf{y}_R + \mathbf{w}_{kh}^{(0)}\mathbf{h}_j^{(0)})},$$

where $\mathbf{w}_{ky}^{(o)}$ and $\mathbf{w}_{kh}^{(0)}$ are parameter vectors. The $\mathbf{h}_j^{(0)}$ is the j-th latent state of the decoder neural network. We note that instead of the decoder layer to generate free-form answers for open-ended question answering, it is also possible to be extended for multiple-choice question answering.

Table 1: Summary of Dataset

Data Splitting	Question Types			
	Object	Number	Color	Location
All	25,767	9,933	14,916	3,530
Train	19,205	7,355	10,813	2,519
Valid	2,574	976	1,390	336
Test	3,988	1,602	2,713	675

3 Experiments

3.1 Data Preparation

We construct the video question-answering datset from the annotated video clip data [Li *et al.*, 2016] with natural language descriptions, which consists of 201,068 GIFs and 287,933 descriptions. Following the state-of-the-art question generation method [Heilman and Smith, 2010], we generate the question-answer pairs from the movie descriptions. Following the existing question answering approaches [Antol *et al.*, 2015; Shih *et al.*, 2016; Yang *et al.*, 2016; Zhao *et al.*, 2015; 2016], we generate four types of questions, which are related to the object, number, color and location queries for the video. We split the generated dataset into three parts: the training, the validation and the testing sets. The four types of video question-answering pairs used for the experiments are summarized in Table 1.

We then preprocess the video question-answering dataset as follows. We first sample 25 frames from each video and then resize each frame to 224×224. We extract the visual representation of each frame by the pretrained VGGNet [Simonyan and Zisserman, 2014], and take the 4,096-dimensional feature vector for each frame. We choose 3 candidate regions for each frame. We employ the pretrained word2vec model to extract the semantic representation of questions and answers. Specifically, the size of vocabulary set is 6,500 and the dimension of word vector is set to 256. Note that we add a token <eos> to mark the end of the answer phrase, and take the token <Unk> for the out-of-vocabulary word.

3.2 Evaluation Criteria

We evaluate the performance of our proposed r-STAN method based on two widely-used evaluation criteria for visual question answering, i.e., Accuracy [Antol *et al.*, 2015] and WUPS [Malinowski and Fritz, 2014]. Given the testing question $\mathbf{q} \in Q_t$ with its ground-truth answer $\mathbf{a}$, we denote the predicted answers from our r-STAN method by $\mathbf{o}$. We now introduce the evaluation criteria below.

- **Accuracy.** The Accuracy is the normalized criteria of accessing the quality of the generated answer based on the testing question set Q_t, given by

$$Accuracy = \frac{1}{|Q_t|} \sum_{\mathbf{q} \in Q_t} (1 - \prod_{i=1}^{M} \mathbf{1}[\mathbf{a}_i \neq \mathbf{o}_i]),$$

  where $Accuracy = 1$ (best) means that the generated answer and the ground-truth ones are exactly the same, while $Accuracy = 0$ means the opposite.

- **WUPS.** The WUPS is the soft measure based on the WUP [Wu and Palmer, 1994] score to evaluate the quality of the generated answer. The WUP measures word similarity based on WordNet [Fellbaum, 1998]. Thus, given the set of generated answer words $O_q = \{o_1, o_2, \ldots, o_M\}$ and the ground-truth ones $A_q = \{a_1, a_2, \ldots, a_M\}$ for testing question $\mathbf{q}$, the WUPS score with the threshold γ is given by

$$\text{WUPS} = \frac{1}{|Q_t|} \sum_{\mathbf{q} \in Q_t} \min\{ \prod_{a_i \in A_q} \max_{o_j \in O_q} WUP_\gamma(a_i, o_j), \prod_{o_i \in O_q} \max_{a_j \in A_q} WUP_\gamma(o_i, a_j)\},$$

  where the $WUP_\gamma(\cdot)$ score is given by

$$WUP_\gamma(a_i, o_j) = \begin{cases} WUP(a_i, o_j) & WUP(a_i, o_j) \geq \gamma \\ 0.1 \cdot WUP(a_i, o_j) & WUP(a_i, o_j) < \gamma \end{cases}$$

  Following the experimental setting in [Malinowski and Fritz, 2014], we choose two WUPS evaluation criteria with the parameter γ to be 0 and 0.9, denoted by WUPS@0.0 and WUPS@0.9, respectively. Because of space limitation, we present the experimental results with $M = 1$, and illustrate the results with high value of M in the extended version of this paper.

3.3 Performance Comparisons

We extend the existing visual question answering methods as the baseline algorithms for the problem of video question answering.

- **VQA+** method is the extension of VQA algorithm [Antol *et al.*, 2015], where we add the mean-pooling layer that obtains the joint video representation from VGGNet-based frame features, and then computes the joint representation of question embedding and video representation by their element-wise multiplication for generating open-ended answers.
- **SAN+** method is the incremental algorithm based on stacked attention networks [Yang *et al.*, 2016], where we add the GRU network to fuse the sequential representation of spatially attended frames for video question answering.
- **UTC+** method is based on the temporal-context encoder-decoder framework [Zhu *et al.*, 2015], where we add the GRU-based decoder network for video question answering.
- **QRU+** method is modified from the QRU algorithm [Li and Jia, 2016], where we add the bidirectional GRU network for obtaining video representation and perform multiple question representation updates for video question answering.

Among them, methods VQA+, SAN+ and QRU+ are extended from the image-based question answering methods, while UTC+ is modified from fill-in-the-blank video question answering method. Unlike the previous visual question answering works, our r-STAN method learns the hierarchical spatio-temporal attended video representation with multiple

Table 2: Experimental results on Accuracy, WUPS@0.0 and WUPS@0.9 with all types of visual questions.

Method	Accuracy	WUPS@0.0	WUPS@0.9
VQA+	0.37	0.6851	0.4993
SAN+	0.4101	0.7159	0.5039
UTC+	0.42	0.7107	0.5066
QRU+	0.4747	0.7574	0.5692
r-STAN$_{(0)}$	0.478	0.7601	0.5753
r-STAN$_{(1)}$	0.48	0.763	**0.5807**
r-STAN$_{(2)}$	**0.4893**	**0.7728**	0.5788

reasoning process for the problem. To exploit the effect of reasoning process, we denote the our r-STAN method with r reasoning steps by r-STAN$_{(r)}$ and the one without reasoning process by r-STAN$_{(0)}$. The input words of our method are initialized by pre-trained word embeddings [Mikolov *et al.*, 2013] with size of 256, and weights of GRUs are randomly by a Gaussian distribution with zero mean.

Table 2 shows the overall experimental results of the methods on all types of questions based on three evaluation criteria. Tables 3, 4 and 5 illustrate the evaluation results on Accuracy, WUPS@0.0 and WUPS@0.9 with different types of questions, respectively. The hyperparamters and parameters which achieve the best performance on the validation set are chosen to conduct the testing evaluation. We report the average value of all the methods on three evaluation criteria. The experiments reveal a number of interesting points:

- The methods based on GRU network learning, SAN+, UTC+ and QRU+ outperform the mean-pooling based method VQA+, which suggests that the sequential frame-level representation is critical for the problem.
- The reasoning based method QRU+ achieves better performance than other baselines. This suggests that the reasoning framework that enables the multiple updates over the joint representation of video and question can also improve the performance of video question answering.
- In all the cases, our r-STAN method achieves the best performance. This fact shows that the reasoning spatio-temporal attention network learning framework that exploits both the joint spatio-temporally attended video representation, and multiple reasoning updates can further improve the performance of video question answering.

In our approach, there are three essential parameters, which are the dimension of hidden state in bi-GRU networks, the dimension of hidden state in bi-aGRU network and the size of fully connected units for decoder networks. We investigate the effect of these parameters on our method by varying both the dimension of hidden state in bi-GRU and bi-aGRU networks from 128 to 1,024, and the number of hidden units in fully connected layer from 300 to 1,200. We first illustrate the performance of our method by varying the dimension of hidden state in bi-aGRU network, the dimension of hidden state in bi-GRU network and the number of hidden untis on Accuracy in Figures 3(a), 3(b) and 3(c). We then vary these parameters to show their effect on our method using WPUS@0.9 in Figures 4(a), 4(b) and 4(c). Our method achieves the best performance when the dimension of hidden state of bi-GRU networks is set to 512, the dimension of hidden state in bi-aGRU networks is set to 512 and the number of hidden units in fully connected layer is set to 500.

Table 3: Experimental results on Accuracy with different types of visual questions.

Method	Accuracy			
	Object	Number	Color	Location
VQA+	0.3333	0.7901	0.2515	0.1764
SAN+	0.3108	0.7243	0.3841	0.366
UTC+	0.3313	0.7284	0.3741	0.3928
QRU+	0.3795	0.7704	0.4791	0.3273
r-STAN$_{(0)}$	0.3749	0.7715	0.4928	0.3541
r-STAN$_{(1)}$	0.3807	0.7674	0.4956	0.3422
r-STAN$_{(2)}$	**0.3815**	**0.7899**	**0.5064**	**0.372**

Table 4: Experimental results on WUPS@0.0 with different types of visual questions.

Method	WUPS@0.0			
	Object	Number	Color	Location
VQA+	0.4962	0.9746	0.8867	0.262
SAN+	0.5997	0.9456	0.8029	0.5798
UTC+	0.6146	0.9459	0.7674	0.53
QRU+	0.6051	0.9606	0.9169	0.6741
r-STAN$_{(0)}$	0.6085	0.9592	0.9175	**0.6918**
r-STAN$_{(1)}$	0.616	0.9632	**0.9195**	0.6604
r-STAN$_{(2)}$	**0.6387**	**0.9645**	0.9188	0.6394

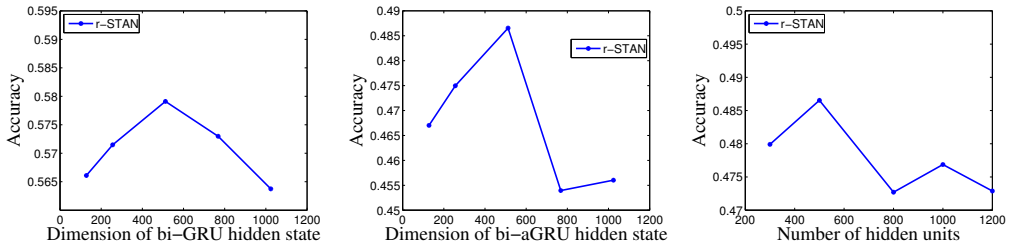
Table 5: Experimental results on WUPS@0.9 with different types of visual questions.

Method	WUPS@0.9			
	Object	Number	Color	Location
VQA+	0.3764	0.8385	0.5585	0.2099
SAN+	0.3745	0.8436	0.5216	0.4353
UTC+	0.3969	0.8423	0.4898	0.4418
QRU+	0.4301	0.8866	0.6371	0.4325
r-STAN$_{(0)}$	0.4266	0.8874	0.6611	**0.4534**
r-STAN$_{(1)}$	0.4394	0.8866	**0.6616**	0.4371
r-STAN$_{(2)}$	**0.4397**	**0.8987**	0.6438	0.4491

4 Related Work

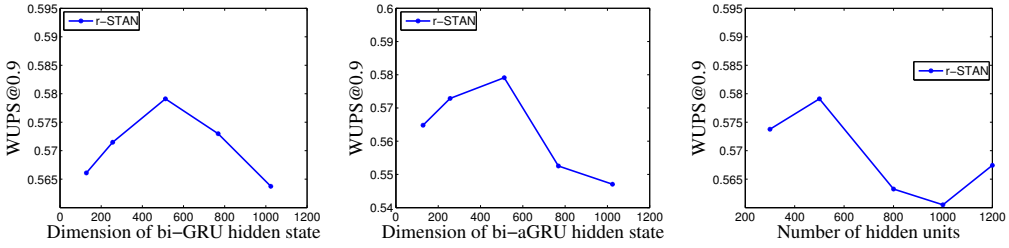
In this section, we briefly review some related work on image-based question answering and video-related question answering.

The problem of image-based question answering has attracted considerable attention recently [Antol *et al.*, 2015; Shih *et al.*, 2016; Lu *et al.*, 2016; Li and Jia, 2016; Xiong *et al.*, 2016]. Given an image and a natural language question about the image, the task of image question answering [Antol *et al.*, 2015] is to provide an accurate natural language answer. With the advancement of visual attention [Xu *et al.*, 2015], Shih et. al. [Shih *et al.*, 2016] introduce the spatial



(a) Dimension of bi-GRU state (b) Dimension of bi-aGRU state (c) Number of hidden units

Figure 3: Effect of bi-GRU hidden state dimension, bi-aGRU hidden state dimension and number of hidden units on Accuracy.



(a) Dimension of bi-GRU state (b) Dimension of bi-aGRU state (c) Number of hidden units

Figure 4: Effect of bi-GRU hidden state dimension, bi-aGRU hidden state dimension and number of hidden units on WUPS@0.9.

attention that selects the relevant image regions to the text-based questions. Yang et. al. [Yang *et al.*, 2016] develop the stacked attention networks and Lu et. al. [Lu *et al.*, 2016] propose the co-attention mechanism with joint image and question attention for image question answering. To exploit complex visual question answering tasks, QRU method [Li and Jia, 2016] employs the reasoning process with attention mechanism that iteratively selects the relevant image regions for question representation update. Xiong et. al. [Xiong *et al.*, 2016] introduce the dynamic memory networks for both image and textual question answering. Kim et. al. [Kim *et al.*, 2016] present multimodal residual networks for image question-answering, which uses element-wise multiplication for the joint residual mappings exploiting the residual learning of the attentional models. Malinowski et. al. [Malinowski and Fritz, 2014] propose a multi-world approach for open-ended image question answering. A survey of existing image question answering works can be found in [Wu *et al.*, 2016].

As a natural extension of image-based question answering, video-related question answering is introduced as a more challenging task, which has drawn a significant attention [Tapaswi *et al.*, 2016; Zhu *et al.*, 2015; Mazaheri *et al.*, 2016; Tu *et al.*, 2014; Yu *et al.*, 2015]. Tapaswi et. al. [Tapaswi *et al.*, 2016] introduce the multi-modal movie-related question answering with high-level abstract concepts, which are extracted from both visual and external information. On the other hand, the proposed models [Zhu *et al.*, 2015; Mazaheri *et al.*, 2016; Tu *et al.*, 2014] for fill-in-the-blank [Yu *et al.*, 2015] video-related question answering task are mainly based on time-invariant visual information, which are extended from the existing image-based question answering methods. Unlike the previous studies, we formulate the problem of open-ended video question answering from the viewpoint of hierarchical spatio-temporal attentional encoder-decoder network learning.

5 Conclusion

In this paper, we present the problem of open-ended video question answering from the viewpoint of spatio-temporal attentional encoder-decoder learning framework. We first propose the hierarchical spatio-temporal attention networks to learn the video representation from the critical frames of the targeted objects according to the question. We then incorporate the multi-step reasoning process to our proposed attention networks that enables the progressive joint representation learning of multimodal spatio-temporal attentional video and textual question for video question answering. We construct a large-scale video question answering dataset and evaluate the effectiveness of our proposed method through extensive experiments.

Acknowledgments

This work was supported by National Basic Research Program of China (973 Program) under Grant 2013CB336500, and National Natural Science Foundation of China under Grant 61602405, Fundamental Research Funds for the Central Universities 2016QNA5015 and the China Knowledge Centre for Engineering Sciences and Technology. The Project is also Supported by the Key Laboratory of Advanced Information Science and Network Technology of Beijing (XDXX1603).

References

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[Fellbaum, 1998] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

[Fernando and Gould, 2016] Basura Fernando and Stephen Gould. Learning end-to-end video classification with rank-pooling. In *ICML)*, 2016.

[Heilman and Smith, 2010] Michael Heilman and Noah A Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617. ACL, 2010.

[Kim *et al.*, 2016] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. *NIPS*, 2016.

[Kumar and Irsoy, 2015] Ankit Kumar and Ozan Irsoy. Ask me anything: Dynamic memory networks for natural language processing. 2015.

[Li and Jia, 2016] Ruiyu Li and Jiaya Jia. Visual question answering with question representation update (qru). In *NIPS*, pages 4655–4663, 2016.

[Li *et al.*, 2016] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*, pages 4641–4650, 2016.

[Lu *et al.*, 2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297, 2016.

[Malinowski and Fritz, 2014] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, pages 1682–1690, 2014.

[Mazaheri *et al.*, 2016] Amir Mazaheri, Dong Zhang, and Mubarak Shah. Video fill in the blank with merging lstms. *arXiv preprint arXiv:1610.04062*, 2016.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Shih *et al.*, 2016] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, pages 4613–4621, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sukhbaatar *et al.*, 2015] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NIPS*, pages 2440–2448, 2015.

[Tapaswi *et al.*, 2016] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640, 2016.

[Tu *et al.*, 2014] Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 21(2):42–70, 2014.

[Wu and Palmer, 1994] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. ACL, 1994.

[Wu *et al.*, 2016] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *arXiv preprint arXiv:1607.05910*, 2016.

[Xiong *et al.*, 2016] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *ICML*, 1603, 2016.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016.

[Yao *et al.*, 2015] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015.

[Yu *et al.*, 2015] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*, pages 2461–2469, 2015.

[Zhao *et al.*, 2015] Zhou Zhao, Lijun Zhang, Xiaofei He, and Wilfred Ng. Expert finding for question answering via graph regularized matrix completion. *TKDE*, 27(4):993–1004, 2015.

[Zhao *et al.*, 2016] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Expert finding for community-based question answering via ranking metric network learning. In *IJCAI*, pages 3000–3006. AAAI Press, 2016.

[Zhu *et al.*, 2015] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering temporal context for video question and answering. *arXiv preprint arXiv:1511.04670*, 2015.