Multi-Turn Video Question Answering via Multi-Stream Hierarchical Attention Context Network

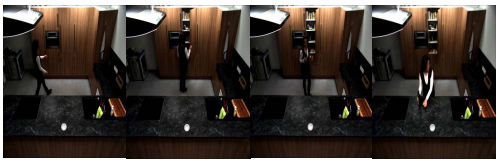
Zhou Zhao[1], **Xinghua Jiang**[1], **Deng Cai**[2], **Jun Xiao**[1*], **Xiaofei He**[2] **and Shiliang Pu**[3]
[1]College of Computer Science, Zhejiang University
[2]State Key Lab of CAD&CG, Zhejiang University
[3]Hikvision Research Institute
{csezhaozhou,dengcai}@gmail.com, jiangxinghua@zju.edu.cn
junx@cs.zju.edu.cn, xiaofeihe@cad.zju.edu.cn, pushiliang@hikvision.com

Abstract

Conversational video question answering is a challenging task in visual information retrieval, which generates the accurate answer from the referenced video contents according to the visual conversation context and given question. However, the existing visual question answering methods mainly tackle the problem of single-turn video question answering, which may be ineffectively applied for multi-turn video question answering directly, due to the insufficiency of modeling the sequential conversation context. In this paper, we study the problem of multi-turn video question answering from the viewpoint of multi-step hierarchical attention context network learning. We first propose the hierarchical attention context network for context-aware question understanding by modeling the hierarchically sequential conversation context structure. We then develop the multi-stream spatio-temporal attention network for learning the joint representation of the dynamic video contents and context-aware question embedding. We next devise the hierarchical attention context network learning method with multi-step reasoning process for multi-turn video question answering. We construct two large-scale multi-turn video question answering datasets. The extensive experiments show the effectiveness of our method.

1 Introduction

Visual question answering is the visual information delivery mechanism that enables users to issue their queries and then collect the answers from the referenced visual contents. Multi-turn video question answering is a challenging task in visual question answering, which automatically generates the accurate answer according to the newly given question and conversation context. Currently, most of the existing visual question answering approaches mainly focus on the problem of single-turn video question answering [Zeng *et al.*, 2017; Zhu *et al.*, 2015; Mazaheri *et al.*, 2016; Tapaswi *et al.*, 2016; Zhao *et al.*, 2017; Jang *et al.*, 2017]. Although the existing proposed methods have achieved promising performance in the single-turn video question answering task, they may still be ineffectively extended to the problem of multi-turn video question answering, due to the lack of modeling the visual conversation context for answer inference.



Conversation Context	Person A	Where is the woman?
	Person B	In a kitchen
	Person A	Where did the woman open?
	Person B	The cabinet
	Person A	How many potatoes did the woman take?
	Person B	Two potatoes
Video Question Answering		
	Person A	Where did the woman place them on?
	Person B	On the counter

Figure 1: Multi-Turn Video Question Answering.

In conversational video question answering task, the context information is particularly important to video question understanding, due to the casual and short video question content. We illustrate a simple example of conversational video question answering in Figure 1. We show that in order to generate the right answer for the question "where did the woman place them on?", the collective conversation con-

*Corresponding author is Jun Xiao.

text information is required for the answer inference. Thus, the simple extension of the existing single-turn video question answering methods is difficult to provide the satisfactory results. Unlike the single-turn video question answering, the multi-turn video question answering method generates the answer from the referenced video content according to the given question as well as the conversation context. The historical conversation context is often in a hierarchical structure and has two levels of sequential relationships, which are the words in conversation turn and conversation turns in the context. Furthermore, not all the conversation context information are equally important for multi-turn video question answering. Therefore, in order to achieve high-quality multi-turn video question answering, it is important to model the hierarchical sequential relationships among conversation context and to identify the important contextual information for multi-turn video question answering.

In this paper, we study the problem of multi-turn video question answering from the viewpoint of hierarchical attention context network learning. We first propose the hierarchical recurrent neural networks with attention mechanisms to model the sequential relationships among conversation context as well as the importance of contextual information for context-aware question understanding. We then devise the multi-stream hierarchical neural networks with saptio-temporal attention mechanisms to learn the joint representation of video contents and context-aware question embedding. We next develop the hierarchical attention context network learning method with multi-step reasoning process for multi-turn video question answering, named as MHACN. When a certain question is given, MHACN can generate the answer for it based on the referenced video contents and its conversation context. The main contributions of this paper are as follows:

- Unlike the previous studies, we present the problem of multi-turn video question answering from the viewpoint of hierarchical attention context network learning. We propose the multi-stream hierarchical attention context network that learns the joint representation of dynamic video content according to the context-aware question understanding.
- We incorporate the multi-step reasoning process for the proposed multi-stream hierarchical attention context network to enable the progressive joint representation learning of the multi-stream attentional video and context-aware question embedding, which further improves the performance of multi-turn video question answering.
- We construct two large-scale datasets for multi-turn video question answering and validate the effectiveness of our proposed method through extensive experiments.

2 Multi-Turn Video Question Answering via Attention Context Networks

2.1 Problem Formulation

Before presenting our method, we first introduce some basic notions and terminologies. We denote the video by $\mathbf{v} \in V$ and conversation context by $\mathbf{u} \in U$, respectively. The frame-level representation for video $\mathbf{v}$ by $\mathbf{v}^{(f)} = (\mathbf{v}_1^{(f)}, \mathbf{v}_2^{(f)}, \ldots, \mathbf{v}_{T^{(f)}}^{(f)})$, where $T^{(f)}$ is the number of frames in video $\mathbf{v}$. The $\mathbf{v}_i = \{\mathbf{v}_{i1}^{(f)}, \mathbf{v}_{i2}^{(f)}, \ldots, \mathbf{v}_{iK}^{(f)}\}$ is the set of region features in the i-th frame by pre-trained 2D-ConvNet. The segment-level representation of video $\mathbf{v}$ is denoted by $\mathbf{v}^{(s)} = (\mathbf{v}_1^{(s)}, \mathbf{v}_2^{(s)}, \ldots, \mathbf{v}_{T^{(s)}}^{(s)})$, where $T^{(s)}$ is the number of segments in video $\mathbf{v}$ and $\mathbf{v}_j^{(s)}$ is the embedding of the j-th segment by pre-trained 3D-ConvNet. We denote the conversation context $\mathbf{u} \in U$ by $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M)$, where $\mathbf{u}_k$ is the k-th round question answering conversation. We then denote the question by $\mathbf{q} \in Q$ and the answer by $\mathbf{a} \in A$, and the k-th round conversation $\mathbf{u}_k$ is composed of question $\mathbf{q}_k$ and answer $\mathbf{a}_k$.

Since the video representations and conversation context are sequential data with variant length, it is natural to choose the variant recurrent neural network called long-short term memory network (LSTM) [Hochreiter and Schmidhuber, 1997] to learn their feature representations. We first denote the output states of frame-level video representations using LSTM by $\mathbf{h}^{(f)} = (\mathbf{h}_1^{(f)}, \mathbf{h}_2^{(f)}, \ldots, \mathbf{h}_{T^{(f)}}^{(f)})$, where $\mathbf{h}_i^{(f)}$ is the output state of the i-th frame in video $\mathbf{v}$. We then consider the output states of segment-level video representations by $\mathbf{h}^{(s)} = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_{T^{(s)}}^{(s)})$, where $\mathbf{h}_j^{(s)}$ is the output state of the j-th segment in video $\mathbf{v}$. We next denote the output state of question representation by $\mathbf{h}^{(q)}$ and the output state of answer representation by $\mathbf{h}^{(a)}$, respectively.

Using the notations above, the problem of multi-turn video question answering is formulated as follows. Given the set of videos V, conversation context U, questions Q and the associated answers A, our goal is to learn the multi-stream hierarchical attention context network such that when a new question is issued, MHACN can generate the answer for it based on the referenced video content and current visual conversation context.

2.2 Multi-Stream Hierarchical Attention Context Network Learning

We first propose the context-aware question understanding method to learn the coherent question representation with conversation context. We consider that the conversation context is in a hierarchical structure and has two levels of sequential relations among questions, answers and each round of conversation context within the structure. Furthermore, we note that not all parts of conversation context are equally important for question understanding. Therefore, we propose the hierarchical recurrent neural networks with fusion mechanisms to model the conversation context and then devise the attention-over-context mechanism to learn the context-aware question representation.

We employ the LSTM networks to learn the representation of the question and the answer in the k-th round of the conversation context, denoted by $\mathbf{h}_k^{(q)}$ and $\mathbf{h}_k^{(a)}$. We then employ the joint representation of question-answer pair mechanism [Zhou *et al.*, 2015], to learn the representation of the k-th round of the conversation context $\mathbf{u}_k$ by fusing the out-

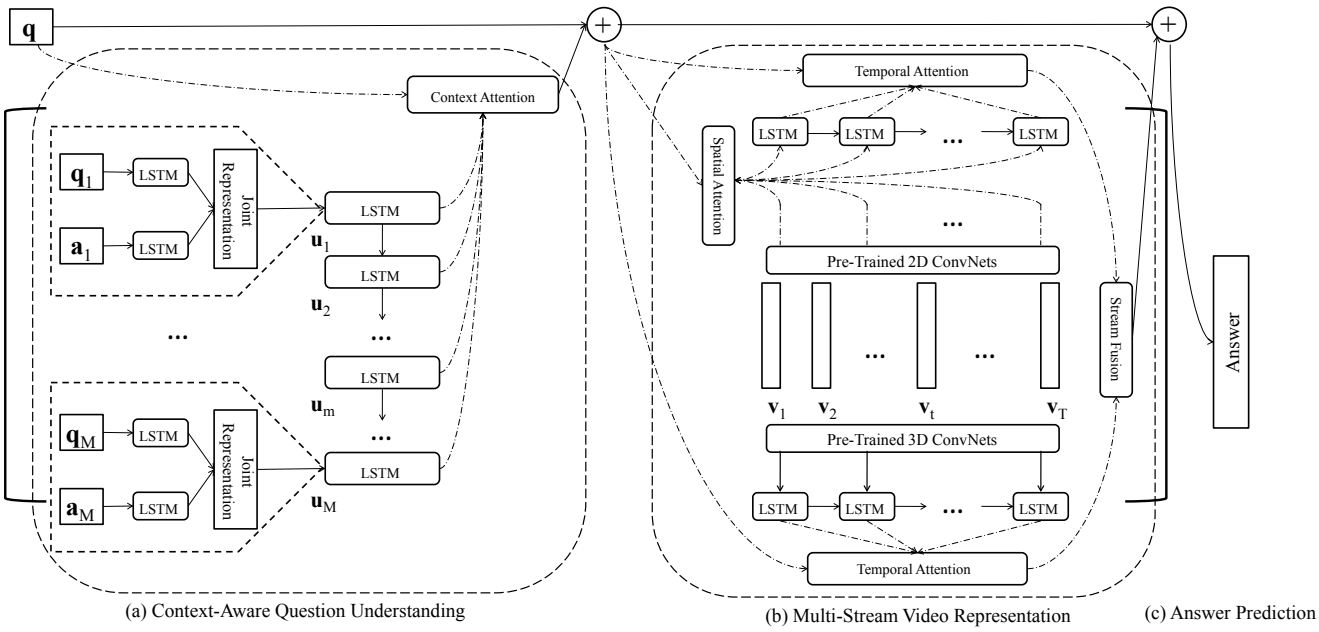


Figure 2: The Overview of Multi-Stream Hierarchical Attention Context Network for Multi-Turn Video Question Answering. (a) We perform the context-aware question understanding with attention mechanisms. (b) We learn the question-aware joint video representation based on multi-stream hierarchical attention network and stream fusion mechanism. (c) We learn the answer prediction model based on softmax loss and question-aware joint video representation for multi-turn video question answering.

put states of question $\mathbf{h}_k^{(q)}$ and answer $\mathbf{h}_k^{(a)}$, given by

$$\mathbf{u}_k = g(\mathbf{W}^{(q)}\mathbf{h}_k^{(q)} + \mathbf{W}^{(a)}\mathbf{h}_k^{(a)}), \quad (1)$$

where $+$ denotes the element-wise addition for the joint representation of the question and answer contents (i.e., $\mathbf{h}_k^{(q)}$ and $\mathbf{h}_k^{(a)}$). The projection matrix $\mathbf{W}^{(q)}$ and $\mathbf{W}^{(a)}$ are used for the fusion of question and answer representations. We consider that the $g(\cdot)$ is the element-wise scaled hyperbolic tangent function, which has shown the good performance for multimodal representation fusion in [Orr and Müller, 2003]. We then learn the representation of conversation context using LSTM networks based on the joint representations $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M)$, denoted by $\mathbf{h}^{(u)} = (\mathbf{h}_1^{(u)}, \mathbf{h}_2^{(u)}, \ldots, \mathbf{h}_M^{(u)})$.

We next learn the context-aware question representation with attention-over-context mechanisms. Given the input question $\mathbf{q}$ and the representation of conversation context $\mathbf{h}^{(u)} = (\mathbf{h}_1^{(u)}, \mathbf{h}_2^{(u)}, \ldots, \mathbf{h}_M^{(u)})$, the attention-over-context score $s_i^{(q,u)}$ is given by

$$s_i^{(q,u)} = \mathbf{w}^{(q,u)} tanh(\mathbf{W}^{(q)}\mathbf{h}^{(q)} + \mathbf{W}^{(u)}\mathbf{h}_i^{(u)} + \mathbf{b}_s^{(q,u)}), \quad (2)$$

where $\mathbf{h}^{(q)}$ is the output state of question $\mathbf{q}$ using LSTM networks. The $\mathbf{W}^{(q)}$, $\mathbf{W}^{(u)}$ are parameter matrices and $\mathbf{b}_s^{(q,u)}$ is the bias vector. The $\mathbf{w}^{(q,u)}$ is the parameter vector for computing the attention-over-context score. For each round of conversation context $\mathbf{u}_i$, its activation for the given question $\mathbf{q}$ by the softmax function is given by $\alpha_i^{(q,u)} = \frac{\exp(s_i^{(q,u)})}{\sum_i \exp(s_i^{(q,u)})}$, which is the normalization of the attention-over-context scores. Thus, the conversation context attended question representation is given by $\mathbf{h}^{(q,u)} = \sum_i \alpha_i^{(q,u)}\mathbf{h}_i^{(u)}$. Therefore, the context-aware question representation is given by $\hat{\mathbf{h}}^{(q)} = \mathbf{h}^{(q)} + \mathbf{h}^{(q,u)}$.

Given the context-aware question representation $\hat{\mathbf{h}}^{(q)}$, we first develop the hierarchical spatio-temporal attention networks to learn the frame-level question-aware video representation. Since the global representation of the frame may fail to capture all necessary information for answering the question [Li and Jia,], it is natural to choose the spatial attention mechanism to automatically localize the targeted regions in each frame according to the question. Following the existing spatial attention mechanism [Li and Jia,], we employ the object generator to produce a set of candidate regions that are most likely to be an object. We extract the frame-level feature using 2D-ConvNet [Krizhevsky *et al.*,] by $\mathbf{v}^{(f)} = (\mathbf{v}_1^{(f)}, \mathbf{v}_2^{(f)}, \ldots, \mathbf{v}_{T^{(f)}}^{(f)})$, where $\mathbf{v}_i^{(f)} = \{\mathbf{v}_{i1}^{(f)}, \mathbf{v}_{i2}^{(f)}, \ldots, \mathbf{v}_{iK}^{(f)}\}$ is the set of region features of the i-th frame. The $\mathbf{v}_{i1}^{(f)}, \mathbf{v}_{i2}^{(f)}, \ldots, \mathbf{v}_{i(K-1)}^{(f)}$ are the candidate region features and $\mathbf{v}_{iK}^{(f)}$ is the whole frame feature. Given the region feature of the i-th frame $\mathbf{v}_{ij}^{(f)} \in \mathbf{v}_i^{(f)}$ with context-aware question representation $\hat{\mathbf{h}}^{(q)}$, its spatial attention score $s_{ij}^{(q,r)}$ is given by

$$s_{ij}^{(q,r)} = \mathbf{w}^{(q,r)} tanh(\hat{\mathbf{W}}^{(q)}\hat{\mathbf{h}}^{(q)} + \mathbf{W}^{(r)}\mathbf{v}_{ij}^{(f)} + \mathbf{b}_s^{(q,r)}), \quad (3)$$

where $\hat{\mathbf{W}}^{(q)}$, $\mathbf{W}^{(r)}$ are parameter matrices and $\mathbf{b}_s^{(q,r)}$ is the bias vector. The $\mathbf{w}^{(q,r)}$ is the parameter vector for computing the frame-level spatial attention score. For each region feature, its activation for the given context-aware question representation $\hat{\mathbf{h}}^{(q)}$ is given by $\alpha_{ij}^{(q,r)} = \frac{\exp(s_{ij}^{(q,r)})}{\sum_j \exp(s_{ij}^{(q,r)})}$, which is the normalization of the spatial attention score. The spatially attended frame representation is given by $\hat{\mathbf{v}}_i^{(f)} = \sum_j \alpha_{ij}^{(q,r)} \mathbf{v}_{ij}^{(f)}$.

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 frames with the targeted information according to the question. We thus introduce the temporal attention mechanism to estimate the relevance of video frames according to the question. Given the spatially attended frames $\hat{\mathbf{v}}^{(f)} = (\hat{\mathbf{v}}_1^{(f)}, \hat{\mathbf{v}}_2^{(f)}, \ldots, \hat{\mathbf{v}}_{T^{(f)}}^{(f)})$, we first learn their latent state representations from LSTM networks by $\mathbf{h}^{(f)} = (\mathbf{h}_1^{(f)}, \mathbf{h}_2^{(f)}, \ldots, \mathbf{h}_{T^{(f)}}^{(f)})$. Then, for each frame $\mathbf{h}_i^{(f)}$, its temporal attention score $s_i^{(q,f)}$ is given by

$$s_i^{(q,f)} = \mathbf{w}^{(q,f)} tanh(\hat{\mathbf{W}}^{(q)} \hat{\mathbf{h}}^{(q)} + \mathbf{W}^{(f)} \mathbf{h}_i^{(f)} + \mathbf{b}_s^{(q,f)}), \quad (4)$$

where $\hat{\mathbf{W}}^{(q)}$, $\mathbf{W}^{(f)}$ are parameter matrices and $\mathbf{b}_s^{(q,f)}$ is the bias vector. The $\mathbf{w}^{(q,f)}$ is the parameter vector for computing the frame-level temporal attention score. For each frame, its activation for the given context-aware question representation $\hat{\mathbf{h}}^{(q)}$ is given by $\alpha_i^{(q,f)} = \frac{\exp(s_i^{(q,f)})}{\sum_i \exp(s_i^{(q,f)})}$, which is the normalization of temporal attention score. Thus, the temporally attended frame representation is given by $\hat{\mathbf{h}}^{(f)} = \sum_i \alpha_i^{(q,f)} \mathbf{h}_i^{(f)}$.

We then develop the temporal attention networks to learn the segment-level question-aware video representation. We extract the segment-level feature using 3D-ConvNet [Tran *et al.*, 2015] by $\mathbf{v}^{(f)} = (\mathbf{v}_1^{(s)}, \mathbf{v}_2^{(s)}, \ldots, \mathbf{v}_{T^{(s)}}^{(s)})$, and learn their latent state representations using LSTM by $\mathbf{h}^{(f)} = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_{T^{(s)}}^{(s)})$. For each video segment $\mathbf{h}_i^{(s)}$, its temporal attention score based on the context-aware question representation $\hat{\mathbf{h}}^{(q)}$ is given by

$$s_i^{(q,s)} = \mathbf{w}^{(q,s)} tanh(\hat{\mathbf{W}}^{(q)} \hat{\mathbf{h}}^{(q)} + \mathbf{W}^{(s)} \mathbf{h}_i^{(s)} + \mathbf{b}_s^{(q,s)}), \quad (5)$$

where $\hat{\mathbf{W}}^{(q)}$, $\mathbf{W}^{(s)}$ are parameter matrices and $\mathbf{b}_s^{(q,s)}$ is the bias vector. The $\mathbf{w}^{(q,s)}$ is the parameter vector. For each video segment, its activation for the given context-aware question representation $\hat{\mathbf{h}}^{(q)}$ is given by $\alpha_i^{(q,s)} = \frac{\exp(s_i^{(q,s)})}{\sum_i \exp(s_i^{(q,s)})}$. Thus, the temporally attended segment representation is given by $\hat{\mathbf{h}}^{(s)} = \sum_i \alpha_i^{(q,s)} \mathbf{h}_i^{(s)}$.

Therefore, we learn the question-aware video representation using multi-stream hierarchical attention context network by $y_{\mathbf{h}^{(q)}}(\mathbf{u}, \mathbf{v}) = \hat{\mathbf{h}}^{(f)} \otimes \hat{\mathbf{h}}^{(s)}$, where $\otimes$ is the element-wise product operator. We then incorporate the multi-step reasoning process [Sukhbaatar *et al.*,] for the proposed multi-stream hierarchical attention context network to further improve the performance of multi-turn video question answering. Given the multi-stream hierarchical attention context network $y(\cdot)$, video $\mathbf{v}$ and conversation context $\mathbf{u}$, the multi-stream hierarchical attention context network learning with multi-step reasoning process is given by

$$\begin{aligned} \mathbf{z}_k &= \mathbf{z}_{k-1} + y_{\mathbf{z}_{k-1}}(\mathbf{u}, \mathbf{v}), \\ \mathbf{z}_0 &= y_{\mathbf{h}^{(q)}}(\mathbf{u}, \mathbf{v}), \end{aligned}$$

which is recursively updated. The question-aware video representation is returned after the k-th update, denoted by $\mathbf{z}_k$. The learning process of reasoning multi-stream hierarchical attention context networks is illustrated in Figure 2.

Following the existing visual question answering models [Antol *et al.*, 2015; Kim *et al.*, ; Li and Jia,], we model the problem of multi-turn video question answering as a classification task with pre-defined classes. Given the question-aware video representation $\mathbf{z}$, a softmax function is employed to classify $\mathbf{z}$ into one of the possible answers as

$$p_a = softmax(\mathbf{W}^{(z)} \mathbf{z} + \mathbf{b}_a^{(z)}), \quad (6)$$

where $\mathbf{W}^{(z)}$ is the parameter matrix and $\mathbf{b}_a^{(z)}$ is the bias vector. We note that instead of using softmax function for answer prediction, it is also possible to utilize LSTM, taking the question-aware video representation $\mathbf{z}$ as input, to generate the free-form answers for the open-ended multi-turn video question answering.

3 Experiments

In this section, we first introduce two conversational video question answering datasets, and then conduct several experiments on them, to show the effectiveness of our approach MHACN for multi-turn video question answering.

3.1 Data Preparation

We construct the conversational video question answering datasets from YouTubeClips [Chen and Dolan, 2011] and TACoS-MultiLevel [Rohrbach *et al.*,]. The YouTubeClips data consists of 1,987 videos and TACoS-MultiLevel data is composed of 1,303 videos. Each YouTubeClips video is composed of 60 frames and each TACoS-MultiLevel video consists of 80 frames. For each video, five pairs of crowdsourcing workers from a professional company were invited to construct five different conversational video dialogs. In total, we have 37,228 video question answering pairs for TACoS-MultiLevel data and 66,806 ones for YouTubeClips data. And most of video dialogs in TACoS-MultiLevel data have five turns of conversation question answering pairs. We take 90% of constructed conversational video dialogs as the training data, 5% as the validation data and 5% as the testing ones. For each question, we compute the semantic similarity between its ground-truth answer and all other answers based on the Euclidean distance with the pre-trained glove embedding [Pennington *et al.*,], and then rank the top 50 answers as the candidate answer set. The constructed conversational video question answering datasets will be provided later.

We then process the conversational video question answering datasets as follows. We resize each frame to 224×224 and extract the visual representation of each frame by the pre-trained VGGNet [Simonyan and Zisserman, 2014], and take

the 4,096-dimensional feature vector for each frame. We employ the pre-trained word2vec model [Mikolov *et al.*, 2013] 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. We evaluate the performance of our proposed MHACN method based on three widely-used ranking evaluation criteria MRR, P@K and MeanRank for multi-turn video question answering, which has been widely used in visual question answering [Zeng *et al.*, 2017; Jang *et al.*, 2017; Zhao *et al.*, 2017].

3.2 Performance Comparisons

We employ the existing single-turn video question answering algorithm and extend the existing video question answering methods as the baseline algorithms for the problem of multi-turn video question answering.

- **ESA** method is the single-turn video question answering algorithm [Zeng *et al.*, 2017], which learns the joint video representation based on the given question with attention mechanisms.
- **ESA+** method is the extension of ESA algorithm [Zeng *et al.*, 2017], where we add the LSTM network to model the conversation context and then fuse context representation and question embedding into the joint representation for multi-turn video question answering.
- **STVQA+** method is the extension of STVQA algorithm [Jang *et al.*, 2017], where we add the LSTM network to sequentially fuse the conversation context and spatial-temporal attended video for multi-turn video question answering.
- **STAN+** method is the extension of STAN algorithm [Zhao *et al.*, 2017], where we add the LSTM network for context-aware question understanding and then perform spatio-temporal attention with context-aware question representation for multi-turn video question answering.

Unlike the previous video question answering works, our MHACN method performs the context-aware question understanding and learn multi-stream spatio-temporal attention video representation with multiple reasoning process for the problem. To exploit the effect of multi-stream hierarchical attention process, we denote the spatio-temporal hierarchical attention context network with 2D-ConvNets by SHACN, and the temporal hierarchical attention context network with 3D-ConvNets by DHACN. Next, to validate the effect of reasoning process, we denote that our method without reasoning process by MHACN, and our method with reasoning steps by $\text{MHACN}_{(r)}$. The input words of out method are initialized by pre-trained word embeddings [Mikolov *et al.*, 2013] with size of 256, and the weights of LSTM networks are randomly by a Gaussian distribution with zero mean.

Table 1 shows the experimental results of the methods on MRR, P@1, P@5 and MeanRank using TACoS-MultiLevel dataset. Table 2 demonstrates the evaluation results of the methods using YoutubeClip dataset. The hyperparameters and parameters which achieve the best performance on the validation set are chosen to conduct the testing evaluation.

Method	MRR	P@1	P@5	MeanRank
ESA	0.411	0.298	0.515	11.964
ESA+	0.411	0.3	0.507	10.435
STVQA+	0.427	0.305	0.54	9.762
STAN+	0.452	0.319	0.594	8.401
SHACN	0.444	0.319	0.579	8.726
DHACN	0.452	0.324	0.583	8.622
MHACN	0.512	**0.391**	0.643	6.625
$\text{MHACN}_{(1)}$	**0.526**	0.386	**0.682**	**5.804**

Table 1: Experimental results on TACoS-MultiLevel dataset.

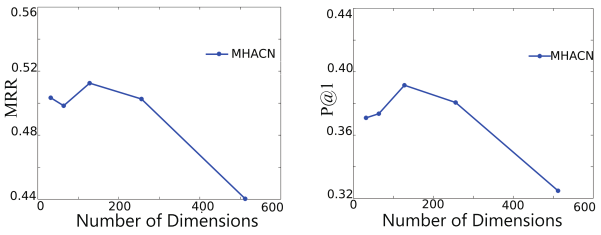
Method	MRR	P@1	P@5	MeanRank
ESA	0.333	0.224	0.418	11.571
ESA+	0.396	0.252	0.541	8.412
STVQA+	0.411	0.266	0.578	7.284
STAN+	0.418	0.274	0.577	7.258
SHACN	0.443	0.283	0.635	6.149
DHACN	0.454	0.295	0.636	6.042
MHACN	0.469	**0.315**	0.661	5.792
$\text{MHACN}_{(1)}$	**0.47**	0.306	**0.67**	**5.496**

Table 2: Experimental results on YoutubeClip dataset.

We report the average value of all the methods on three evaluation criteria. The experimental results reveal a number of interesting points:

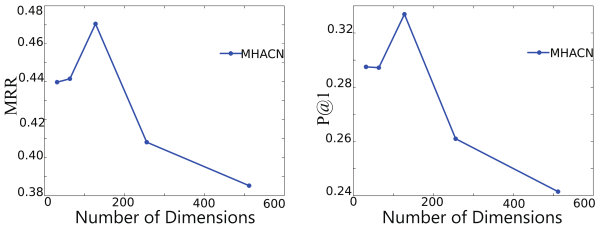
- The methods based on context-aware question understanding, ESA+, STVQA+, STAN+, SHACN, DHACN, MHACN and $\text{MHACN}_{(1)}$ outperform the single-turn video question answering method ESA, which suggests the context-aware question representation is critical for the problem.
- The multi-stream attention context network method MHACN achieves better performance than the methods SHACN and DHACN. This suggests that both the frame-level and segment-level hierarchical attention mechanisms are important for the problem of multi-turn video question answering.
- In all the cases, our $\text{MHACN}_{(1)}$ method achieves the best performance. This fact shows that the reasoning multi-stream attention context network learning framework that exploits both context-aware question understanding and multi-stream attention mechanisms, and multiple reasoning updates can further improve the performance of multi-turn video question answering.

In our approach, there are two essential parameters, which are the dimension of question representation, and the dimension of the joint representation of the question-answer pairs. We vary the dimension of question representation from 32, 60, . . ., to 512, and the dimension of joint representation from 32, 64, . . ., to 512. We first investigate the effect of question representation dimension on MRR, P@1 and P@5 on TACoS-MultiLevel dataset in Figures 3, and on YoutubeClip dataset in Figures 4. We then study the effect of joint representation dimension on TACoS-MultiLevel dataset in Figures 5.



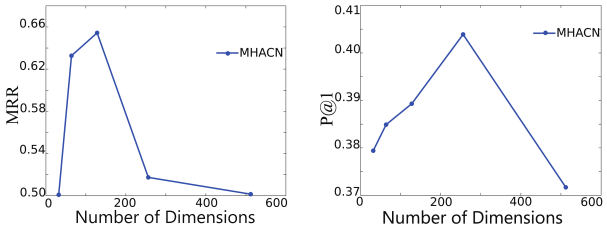
(a) MRR (b) P@1

Figure 3: Effect of Question Representation Dimension on MRR and P@1 using TACoS-MultiLevel dataset.



(a) MRR (b) P@1

Figure 4: Effect of Question Representation Dimension on MRR and P@1 using YoutubeClip dataset.



(a) MRR (b) P@1

Figure 5: Effect of Joint Representation Dimension on MRR and P@1 using TACoS-MultiLevel dataset.

4 Related Work

In this section, we briefly review some related work on visual question answering and dialogue modeling.

The visual question answering task is to provide the accurate answer for the natural language question from the given visual contents [Antol *et al.*, 2015]. The existing approaches can be categorized into image-based question answering methods [Antol *et al.*, 2015; Li and Jia, ; Yang *et al.*, 2016] and video-based question answering ones [Mazaheri *et al.*, 2016; Zeng *et al.*, 2017; Zhu *et al.*, 2015; Zhao *et al.*, 2017; Tapaswi *et al.*, 2016; Jang *et al.*, 2017]. Kim et. al. [Kim *et al.*,] employ the multimodal residual network for image question answering. Li et. al. [Li and Jia,] propose the question representation update method that iteratively selects the relevant image regions and update the question representation. Shih et. al. [Shih *et al.*, 2016] introduce the spatial attention mechanism for image question answering. Das et. al. [Das *et al.*, 2017] study the image question answering based on previous question-answering history. As a natural extension of image-based question answering, the video-based question answering has been introduced as a more challenging task [Zeng *et al.*, 2017]. The fill-in-the-blank approaches [Zhu *et al.*, 2015; Mazaheri *et al.*, 2016] complete the missing entry in the video description by ranking candidate answers based on both visual content and contextual video description. Tapaswi et. al. [Tapaswi *et al.*, 2016] propose the three-way scoring function for movie question answering Zhao et. al. [Zhao *et al.*, 2017] propose the hierarchical spatio-temporal attention networks for video question answering. Jang et. al. [Jang *et al.*, 2017] devise the dual-LSTM method with attention mechanism. Unlike the previous studies, we study the problem of conversational video question answering based on both the visual contents and its conversational context.

Given a dialogue context in natural language, the response generation task is to provide the relevant utterance to the given conversational context [Serban *et al.*, 2017b]. Serban et. al. [Serban *et al.*, 2016] extend the hierarchical recurrent encoder-decoder neural network for responding learning in dialogue systems. Weston et. al. [Weston, 2016] study the dialog-based language learning based on memory network. Serban et. al. [Serban *et al.*, 2017a] devise the multiresolution neural network model for dialogue response generation. Mei et. al. [Mei *et al.*, 2017] study the coherent conversation continuation via RNN-based dialogue models with dynamic attention mechanisms. Unlike the previous studies, the conversational video question answering task is to provide the answer from the multimodal visual contents and textual conversational contexts.

5 Conclusion

In this paper, we study the problem of multi-turn video question answering from the viewpoint of multi-step hierarchical attention context network learning. We first propose the hierarchical attention context learning method with recurrent neural networks for context-aware question understanding, which is based on the sequential conversation context structure with attention-over-context mechanisms. We then develop the multi-stream attention network that learns the joint embedding for video question answering from both the spatio-temporal attended frame-level video representation and the temporal attended segment-level video representation. We next incorporate the multi-step reasoning process for the proposed multi-stream hierarchical attention context network to further improve the performance of multi-turn video question answering. We construct two large-scale multi-turn video question answering datasets and evaluate the effectiveness of our proposed method through extensive experiments.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant No.61602405, No.61572431, Zhejiang Natural Science Foundation(LZ17F020001), Key R&D Program of Zhejiang Province(2018C01006) and Joint Research Program of ZJU and Hikvision Research Institute.

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.

[Chen and Dolan, 2011] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL: Human Language Technologies-Volume 1*, pages 190–200. ACL, 2011.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. *CVPR*, 2017.

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

[Jang *et al.*, 2017] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *CVPR*, 2017.

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

[Krizhevsky *et al.*,] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*.

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

[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.

[Mei *et al.*, 2017] Hongyuan Mei, Mohit Bansal, and Matthew R Walter. Coherent dialogue with attention-based language models. In *AAAI*, pages 3252–3258, 2017.

[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.

[Orr and Müller, 2003] Genevieve B Orr and Klaus-Robert Müller. *Neural networks: tricks of the trade*. Springer, 2003.

[Pennington *et al.*,] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*.

[Rohrbach *et al.*,] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*.

[Serban *et al.*, 2016] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784, 2016.

[Serban *et al.*, 2017a] Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron C Courville. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI*, pages 3288–3294, 2017.

[Serban *et al.*, 2017b] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301, 2017.

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

[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.

[Tran *et al.*, 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[Weston, 2016] Jason E Weston. Dialog-based language learning. In *NIPS*, pages 829–837, 2016.

[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.

[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.

[Zeng *et al.*, 2017] 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 *AAAI*, pages 4334–4340, 2017.

[Zhao *et al.*, 2017] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, volume 2, 2017.

[Zhou *et al.*, 2015] Xiaoqiang Zhou, Baotian Hu, Qingcai Chen, Buzhou Tang, and Xiaolong Wang. Answer sequence learning with neural networks for answer selection in community question answering. *ACL*, 2015.

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