Multi-Turn Video Question Answering via Hierarchical Attention Context Reinforced Networks

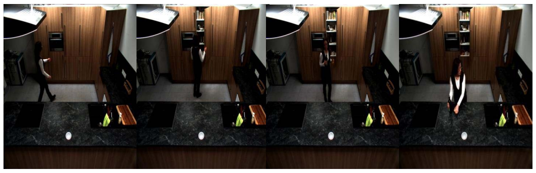
Zhou Zhao, Zhu Zhang, Xinghua Jiang, and Deng Cai, *Member, IEEE*

***Abstract*— Multi-turn 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-stream hierarchical attention context reinforced 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 a multi-step reasoning process to enhance the multi-stream hierarchical attention context network learning method. We finally predict the multiple-choice answer from the candidate answer set and further develop the reinforced decoder network to generate the open-ended natural language answer 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.**

***Index Terms*— Video question answering, multi-turn, attention, reinforcement learning.**

I. 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 an accurate answer according to the newly given question and conversation context. Unlike the single-turn video question answering, multi-turn video question answering allows users to present continuous, interrelated questions based on video contents, while historical question and answer contents are used as references for answering the current question. This is a practical and necessary extension for video question answering systems. Currently, most of the existing video question answering approaches mainly focus on the problem of single-turn video question answering [1]–[6]. Although these methods have achieved promising performance in the single-turn 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	What 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.

Fig. 1. Multi-turn video question answering.

In multi-turn 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 multi-turn 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 context information is required for the answer inference. Without an accurate understanding of the conversation context, it is hard for the video question answering system to comprehend what “them” are referring to in the question, and difficult to generate high-quality answers. Thus, the simple extension of the existing single-turn video question answering methods is difficult to

Manuscript received July 7, 2018; revised November 30, 2018 and December 29, 2018; accepted February 14, 2019. Date of publication February 27, 2019; date of current version June 13, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61602405, Grant 61836002, and Grant 61751209, in part by the Joint Research Program of ZJU, in part by the Hikvision Research Institute, and in part by the Microsoft Research Asia. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Liang Wang. *(Corresponding author: Zhou Zhao.)*

Z. Zhao and Z. Zhang are with the College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: zhaozhou@zju.edu.cn; zhangzhu950310@gmail.com).

X. Jiang and D. Cai are with the State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou 310027, China (e-mail: jiangxinghua@zju.edu.cn; dengcai@cad.zju.edu.cn).

Digital Object Identifier 10.1109/TIP.2019.2902106

provide satisfactory results. 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 is 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 extend our previous work [1] and study the problem of multi-turn video question answering from the viewpoint of multi-stream hierarchical attention context reinforced network learning. Following our previous work, 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 spatio-temporal attention networks to learn the joint representation of video contents and context-aware question embedding, where both appearance and motion semantic features in video contents are captured and fused. We next devise multi-step reasoning process to enhance the multi-stream hierarchical attention context network learning method by developing the sufficient interaction for conversation context, question and video contents. Different from our previous work, we finally predict the multiple-choice answer from the candidate answer set and further develop the reinforced decoder network to generate the open-ended answer for multi-turn video question answering. Compared with the multiple-choice answer prediction limited to pre-defined candidates, open-ended question answering is more practical and challenging to directly generate natural language answers. Our reinforced decoder networks can train the answer sequences as a whole by sampling operations to improve the open-ended performance. Moreover, more experiments and more in-depth analysis are displayed than before. We name the overall network as HACRN. And when a certain question is given, HACRN 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 multi-stream hierarchical attention context reinforced 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.
- Besides conventional multiple-choice answer, we develop the reinforced decoder network to generate the open-ended natural language answer for multi-turn video question answering, which extends the answer form and explores the utilization of the reinforcement learning frameworks in such a challenging task.
- We construct two large-scale datasets for multi-turn video question answering and validate the effectiveness of our proposed method through extensive experiments.

The rest of this paper is organized as follows. We briefly review some related works about visual question answering and dialogue modeling in Section II. In Section III, we introduce the problem of multi-turn video question answering from the viewpoint of hierarchical attention context reinforced network learning. We then present a variety of experimental results in Section IV. Finally, Section V concludes this paper.

II. Related Work

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

A. Visual Question Answering

The visual question answering task is to provide an accurate answer for the natural language question from the given visual contents [8]. The existing approaches can be categorized into image-based question answering methods [8]–[10] and video-based question answering ones [1]–[6].

For image question answering, Malinowski and Fritz [11] develop the multi-world probabilistic approach for open-ended image question answering. Kim *et al.* [12] employ the multimodal residual network, which extends the deep residual network framework and effectively models the joint representation from visual and language information. To exploit complex visual relations in image question answering task, Li and Jia [9] propose the question representation update method that iteratively selects the relevant image regions related to the query and updates the question representation. Recently, the attention mechanisms are applied in different image understanding tasks [13], [14]. Lu *et al.* [13] attend crucial regions of an image for caption generation and Peng *et al.* [14] utilize the attention based joint embedding to learn fine-grained image representations. For image question answering, Lu *et al.* [15] devise the co-attention mechanism that jointly reasons about question and image attention. Yang *et al.* [10] develop the stacked attention method that adopts multiple steps of reasoning to locate the relevant visual clues that lead to the answer of the question layer-by-layer. Shih *et al.* [16] introduce the spatial attention mechanism for image question answering, which maps textual representations and visual features into a shared feature space. Anderson *et al.* [17] propose a combined attention mechanism including bottom-up and top-down to calculate the attention about objects and salient image regions. Patro *et al.* [18] propose an exemplar-based method to obtain a differential attention region, which is different from image-based attention and close to human attention. Unlike the conventional

[1] This work is the extension of our previous paper [7], which is accepted by IJCAI(18). The added benefits of the journal paper are clearly and concisely explained in a cover letter that accompanies the submission. And the previous paper is submitted as a supporting document.

single-turn image question answering, Das *et al.* [19] study the image question answering based on previous question-answering history. Furthermore, a survey of existing image question answering methods can be found in [20].

As a natural extension of image-based question answering, the video-based question answering has been introduced as a more challenging task [1]. The fill-in-the-blank approaches [2], [3] complete the missing entry in the video description by ranking candidate answers based on both visual content and contextual video description. Tapaswi *et al.* [4] propose the three-way scoring function for movie question answering based on both the relevance between given question and textual movie subtitles, and textual movie subtitles and answers. Similar to image understanding, attention mechanisms are widely used in video content understanding [21], [22]. Peng *et al.* [21] apply spatial-temporal attention modeling for video classification and Wang *et al.* [22] present an attention based non-local operation to capture long-range dependencies in videos. In the field of video question answering, Zhao *et al.* [5] propose the hierarchical spatio-temporal attention mechanism to learn the joint representation about the dynamic video information according to the given query. Jang *et al.* [6] devise the dual-LSTM method with attention mechanism and Zeng *et al.* [1] extend the end-to-end memory network with additional LSTM layers for video question answering. Zhao *et al.* [23] divide entire videos into several segments and adopt hierarchical attention method to model segment and video presentation for answer generation. And Gao *et al.* [24] propose a motion-appearance co-memory network to simultaneously learn the motion and appearance features for subsequent answer prediction.

Unlike the previous studies, we study the problem of multi-turn video question answering based on both the visual contents and its conversational context.

B. Dialogue Modeling

Given a dialogue context in natural language, the response generation task is to provide the relevant utterance to the given conversational context [25]. Serban *et al.* [26] extend the hierarchical recurrent encoder-decoder neural network for responding learning in dialogue systems, which generates responses word-by-word and opens up the possibility of realistic and flexible interactions. Weston [27] propose the dialog-based language learning based on memory network, where supervision is from the response of the conversation partner naturally and implicitly. Serban *et al.* [28] devise the multi-resolution neural network mechanism that generates natural language from a high-level coarse token sequence and a natural language token sequence concurrently.

With the development of attention mechanisms, Xing *et al.* [29] propose a hierarchical recurrent attention network for multi-turn response generation and Mei *et al.* [30] study the coherent conversation continuation based on dynamic attention mechanisms. To model 1-to-n relationships between a sentence and its diverse responses, Zhou *et al.* [31] assume that there exists some latent responding mechanisms and build an encoder-diverter-decoder framework with different responding mechanisms for dialogue response generation. Wu *et al.* [32] propose a sequential matching network for the retrieval-based multi-turn response selection task, which adopts multi-level matching mechanisms to distill multi-granularity matching information. Williams *et al.* [33] introduce the hybrid code networks that combine domain-specific knowledge and relevant action templates for task-oriented dialog systems. To model the semantic direction of a sentence that is important to developing coherent, interesting dialogues, Li *et al.* [34] devise the reinforcement learning framework for neural response generation by simulating dialogues between two agents. Furthermore, Li *et al.* [35] apply adversarial training to open-domain dialogue generation under the reinforcement learning framework, aiming to produce sequences that are indistinguishable from human-generated dialogue utterances. Dhingra *et al.* [36] propose the end-to-end reinforcement learning for dialogue agents over knowledge bases, acquiring real-world knowledge by interacting with an external database.

Unlike the previous studies, the multi-turn video question answering task is to provide the answer from the multi-modal visual contents and textual conversational contexts.

III. Proposed Framework

In this section, we first introduce the problem of multi-turn video question answering, and then propose the multi-stream hierarchical attention context reinforced network for the problem.

A. 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 video $\mathbf{v} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{T^{(f)}})$ contains $T^{(f)}$ frames. The frame-level representation for video $\mathbf{v}$ is denoted by $\mathbf{v}^{(f)} = (\mathbf{v}_1^{(f)}, \mathbf{v}_2^{(f)}, \ldots, \mathbf{v}_{T^{(f)}}^{(f)})$. The $\mathbf{v}_i^{(f)} = \{\mathbf{v}_{i1}^{(f)}, \mathbf{v}_{i2}^{(f)}, \ldots, \mathbf{v}_{iK}^{(f)}\}$ is the set of region features in the i-th frame by pre-trained 2D-ConvNet [37]. We then define a segment as a set of consecutive 16 frames in videos, where each segment overlaps 8 frames with adjacent segments. For example, the 1-th segment $\mathbf{s}_1$ consists of $\mathbf{f}_1$ to the $\mathbf{f}_{16}$, and the 2-th segment $\mathbf{s}_2$ consists of $\mathbf{f}_9$ to $\mathbf{f}_{24}$. Thus the video can be denoted as a segment sequence $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{T^{(s)}})$. 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 [38]. 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, which is composed of question $\mathbf{q}_k$ and answer $\mathbf{a}_k$. We then denote the newly question by $\mathbf{q} \in Q$ and the answer by $\mathbf{a} \in A$.

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) [39] 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. We denote the dimension of the hidden states in these LSTM networks by d_h, so the $\mathbf{h}_i^{(f)}$, $\mathbf{h}_i^{(s)}$, $\mathbf{h}_i^{(q)}$ and $\mathbf{h}_i^{(a)}$ have the same dimension d_h.

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 reinforced network such that when a new question is issued, HACRN can generate the answer for it based on the referenced video content and current visual conversation context.

B. Multi-Stream Hierarchical Attention Context Network

In this section, we present the multi-stream hierarchical attention context network learning framework to obtain the question-aware video representation for multi-turn video question answering.

1) Context-Aware Question Understanding: 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 [40], to learn the representation of the k-th round of the conversation context $\mathbf{u}_k$ by fusing the output states of question $\mathbf{h}_k^{(q)}$ and answer $\mathbf{h}_k^{(a)}$, given by

$$\mathbf{u}_k = g(\mathbf{W}_u^1 \mathbf{h}_k^{(q)} + \mathbf{W}_u^2 \mathbf{h}_k^{(a)}), \tag{1}$$

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 $\mathbf{W}_u^1 \in \mathbb{R}^{d_u \times d_h}$ and $\mathbf{W}_u^2 \in \mathbb{R}^{d_u \times d_h}$ are used for the fusion of question and answer representations. The d_h is the dimension of $\mathbf{h}_k^{(q)}$ and $\mathbf{h}_k^{(a)}$ from the LSTM networks and d_u is the dimension of joint representations. We consider that the $g(\cdot)$ is the element-wise scaled hyperbolic tangent function, which has shown the good performance for multi-modal representation fusion in [41]. 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 $\mathbf{h}^{(u)} = (\mathbf{h}_1^{(u)}, \mathbf{h}_2^{(u)}, \ldots, \mathbf{h}_M^{(u)})$, where $\mathbf{h}_i^{(u)} \in \mathbb{R}^{d_h}$.

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

$$s_i^{(q,u)} = \mathbf{w}_{qu}^\top tanh(\mathbf{W}_{qu}^1 \mathbf{h}^{(q)} + \mathbf{W}_{qu}^2 \mathbf{h}_i^{(u)} + \mathbf{b}_{qu}), \tag{2}$$

where $\mathbf{h}^{(q)}$ is the output state of question $\mathbf{q}$ using LSTM networks. The $\mathbf{W}_{qu}^1 \in \mathbb{R}^{d_m \times d_h}$, $\mathbf{W}_{qu}^2 \in \mathbb{R}^{d_m \times d_h}$ are parameter matrices and $\mathbf{b}_{qu} \in \mathbb{R}^{d_m}$ is the bias vector. The $\mathbf{w}_{qu}^\top \in \mathbb{R}^{d_m}$ is the row vector for computing the attention-over-context score. The d_m is the middle dimension. 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. And 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 $\hat{\mathbf{h}}^{(q)} = \mathbf{h}^{(q)} + \mathbf{h}^{(q,u)}$.

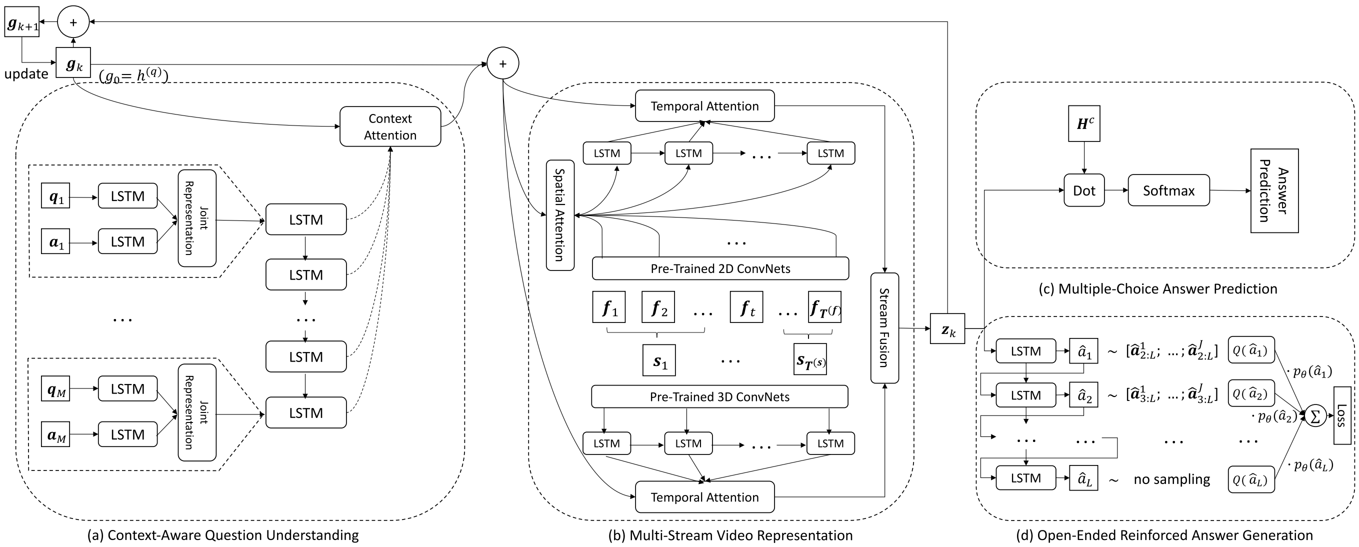
2) Multi-Stream Video Representation: Video contents contain the appearance and motion features simultaneously, thus capturing both appearance and motion semantic information is necessary for high-quality video question answering, which corresponds to the frame-level and segment-level video representation learning.

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 and capture appearance semantic information. Since the global representation of the frame may fail to capture all necessary information for answering the question [9], 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 [9], 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 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

$$s_{ij}^{(q,r)} = \mathbf{w}_{qr}^\top tanh(\mathbf{W}_{qr}^1 \hat{\mathbf{h}}^{(q)} + \mathbf{W}_{qr}^2 \mathbf{v}_{ij}^{(f)} + \mathbf{b}_{qr}), \tag{3}$$

where $\mathbf{W}_{qr}^1 \in \mathbb{R}^{d_m \times d_h}$, $\mathbf{W}_{qr}^2 \in \mathbb{R}^{d_m \times d_v}$ are parameter matrices and $\mathbf{b}_{qr} \in \mathbb{R}^{d_m}$ is the bias vector. The $\mathbf{w}_{qr}^\top \in \mathbb{R}^{d_m}$ is the row vector for computing the frame-level spatial attention score. The d_v is the dimension of frame region features. 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


g_{k+1}
+
update
g_k
$(g_0 = h^{(q)})$
Context Attention
q_1
a_1
LSTM
Joint Representation
q_M
a_M
+
Temporal Attention
Spatial Attention
Pre-Trained 2D ConvNets
f_1 f_2 … f_t … $f_{T^{(f)}}$
s_1 … $s_{T^{(s)}}$
Pre-Trained 3D ConvNets
Stream Fusion
z_k
H^c
Dot
Softmax
Answer Prediction
$\hat{a}_1$ ~ $[\hat{\boldsymbol{a}}^1_{2:L}; \ldots; \hat{\boldsymbol{a}}^J_{2:L}]$
$\hat{a}_2$ ~ $[\hat{\boldsymbol{a}}^1_{3:L}; \ldots; \hat{\boldsymbol{a}}^J_{3:L}]$
$\hat{a}_L$ ~ no sampling
$Q(\hat{a}_1)$
$Q(\hat{a}_2)$
$Q(\hat{a}_L)$
$\cdot p_\theta(\hat{a}_1)$
$\cdot p_\theta(\hat{a}_2)$
$\cdot p_\theta(\hat{a}_L)$
Σ
Loss

(a) Context-Aware Question Understanding

(b) Multi-Stream Video Representation

(c) Multiple-Choice Answer Prediction

(d) Open-Ended Reinforced Answer Generation

Fig. 2. The Overview of multi-stream hierarchical attention context reinforced 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 attention network and stream fusion mechanism. (c) We learn the multiple-choice answer prediction based on question-aware joint video representation. (d) we develop the open-ended reinforced answer generation for multi-turn video question answering.

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

$$s_i^{(q,f)} = \mathbf{w}_{qf}^\top tanh(\mathbf{W}_{qf}^1 \hat{\mathbf{h}}^{(q)} + \mathbf{W}_{qf}^2 \mathbf{h}_i^{(f)} + \mathbf{b}_{qf}), \qquad (4)$$

where $\mathbf{W}_{qf}^1 \in \mathbb{R}^{d_m \times d_h}$, $\mathbf{W}_{qf}^2 \in \mathbb{R}^{d_m \times d_h}$ are parameter matrices and $\mathbf{b}_{qf} \in \mathbb{R}^{d_m}$ is the bias vector. The $\mathbf{w}_{qf}^\top \in \mathbb{R}^{d_m}$ is the row 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 and capture motion semantic information. Specifically, we first extract the segment-level feature from $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{T^{(s)}})$ using 3D-ConvNet [38], which outputs a spatio-temporal representation for each input segment. With the segment-level features $\mathbf{v}^{(s)} = (\mathbf{v}_1^{(s)}, \mathbf{v}_2^{(s)}, \ldots, \mathbf{v}_{T^{(s)}}^{(s)})$, we then learn their latent state representations using LSTM by $\mathbf{h}^{(s)} = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_{T^{(s)}}^{(s)})$. 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

$$s_i^{(q,s)} = \mathbf{w}_{qs}^\top tanh(\mathbf{W}_{qs}^1 \hat{\mathbf{h}}^{(q)} + \mathbf{W}_{qs}^2 \mathbf{h}_i^{(s)} + \mathbf{b}_{qs}), \qquad (5)$$

where $\mathbf{W}_{qs}^1 \in \mathbb{R}^{d_m \times d_h}$, $\mathbf{W}_{qs}^2 \in \mathbb{R}^{d_m \times d_h}$ are parameter matrices and $\mathbf{b}_{qs} \in \mathbb{R}^{d_m}$ is the bias vector. The $\mathbf{w}_{qs}^\top \in \mathbb{R}^{d_m}$ is the row 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.

3) Multi-Step Reasoning Process: For multi-turn video question answering, the understanding of historical conversation context, modeling of question semantic information and representation learning of complex video contents are mutually influenced and improved by each other, thus sufficient interaction and reasoning of these contents are necessary and beneficial for question answering. We then incorporate the multi-step reasoning process [42] 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

$$\mathbf{g}_0 = \mathbf{h}^{(q)}, \qquad (6)$$

$$\mathbf{z}_k = y_{\mathbf{g}_k}(\mathbf{u}, \mathbf{v}), \qquad (7)$$

$$\mathbf{g}_{k+1} = \mathbf{g}_k + \mathbf{z}_k, \qquad (8)$$

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

4) Multiple-Choice Answer Prediction: Given the question-aware video representation $\mathbf{z}$, we first develop the multiple-choice method for multi-turn video question answering. Following the existing visual question answering models [8], [9], [12], we model the problem of multi-turn video question answering as a classification task with pre-defined candidate answers. We first learn the semantic representation $\mathbf{h}_i^{(c)}$ of each candidate answer by another LSTM networks and obtain the answer representation matrix $\mathbf{H}^{(c)} = [\mathbf{h}_1^{(c)}; \mathbf{h}_2^{(c)}; \cdots ; \mathbf{h}_P^{(c)}] \in \mathbb{R}^{d_h \times P}$, where P is the number of candidate answers and the dimension of answer representation is same as the final video representation. Given the question-aware video representation $\mathbf{z} \in \mathbb{R}^{d_h}$, a softmax function is employed to classify $\mathbf{z}$ into one of the possible answers as

$$p_a = softmax(\mathbf{z}^\top \cdot \mathbf{H}^{(c)}), \tag{9}$$

where $p_a \in \mathbb{R}^P$ is a probability distribution for the L candidate answers. 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.

C. Reinforced Decoder Network Learning

In this section, we propose the reinforced decoder network $g(\cdot)$ based on the question-aware video representation $\mathbf{z}$ to generate the open-ended answer for multi-turn video question answering.

Given the question-aware video representation $\mathbf{z}$ from multi-stream hierarchical attention context networks, the LSTM networks generate the t-th word of open-ended answer by sampling $\hat{a}_t \sim p_\theta(\hat{a}_t | \hat{\mathbf{a}}_{1:t-1}, \mathbf{z}) = g_t(\mathbf{h}_t^{(\hat{a})}, \mathbf{z})$, where $g_t(\cdot)$ is the LSTM recurrent answer generator. The $\mathbf{h}_t^{(\hat{a})}$ is the hidden state of decoder network at step t and $\mathbf{z}$ is the question-aware video representation.

To train the decoder network, one general strategy is under the framework of maximum likelihood estimation, given by

$$\mathcal{L}_{ML}(g(\mathbf{z})) = \sum_{t=1}^{L} \log p_\theta(\hat{a}_t | \hat{\mathbf{a}}_{1:t-1}, \mathbf{z}). \tag{10}$$

However, maximum likelihood estimation tends to make the learnt decoder network suboptimal [43], where the words in the answer are trained individually without regard to the integrity of the entire answer. Recently, reinforcement learning has been wide-used in different cross-modal tasks, such as cross-modal translation [44] and video caption [45]. Instead of maximum likelihood estimation, we employ the reinforcement learning framework to train the decoder network, where the answer sequences are trained as a whole based on sampling operations.

Algorithm 1 Reinforced Decoder Network Learning

```
Require: Training set $< U, V, Q, A >$;
Ensure: Optimize model parameters $\boldsymbol{\theta}$;
Randomly initialize the model parameters $\boldsymbol{\theta}$;
Load the pre-trained 2D-ConvNet model, 3D-ConvNet model and word embedding;
for iteration = 1 to maximum iteration time $T$ do
    Randomly sample a minibatch;
    Input the video representation $\mathbf{v}$, conversation context representation $\mathbf{u}$ and question feature $\mathbf{q}$ into multi-stream hierarchical attention context network and obtain the question-aware video representation $\mathbf{z}$ using Equation (1) $-$ (12);
    for $t = 1$ to maximum answer length $L$ do
        Sample the next word in answer by $\hat{a}_t \sim p_\theta(\hat{a}_t | \hat{\mathbf{a}}_{1:t-1}, \mathbf{z})$;
        if $t < L$ then
            Adopt Monte-Carlo simulation, randomly sample answer words starting from $t+1$ step $J$ times and get $J$ answer sequences;
            Calculate $R_{\mathbf{a}}(\hat{\mathbf{a}}) = \|\mathbf{h}^{(a)} - \mathbf{h}^{(\hat{a})}\|^2$ for each $\hat{\mathbf{a}}$;
            Calculate the average reward of the $J$ sequences as the expected cumulative reward at step $t$;
        else $\{t = L\}$
            Calculate $R_{\mathbf{a}}(\hat{\mathbf{a}}) = \|\mathbf{h}^{(a)} - \mathbf{h}^{(\hat{a})}\|^2$ for the final sequences $\hat{\mathbf{a}}$ as the cumulative expected reward at step $L$;
        end if
    end for
    Update the model parameters $\boldsymbol{\theta}$ using Equation (12);
end for
```

In the framework of reinforcement learning, we define the generation of next word as action, and the probability distribution of next word $p_\theta(\hat{a}_t | \hat{\mathbf{a}}_{1:t-1}, \mathbf{z})$ as policy. We then define the reward function by measuring the embedding similarity according to the related visual-semantic embedding works [46]. Given the generated answer $\hat{\mathbf{a}}$ and the ground-truth answer $\mathbf{a}$, the reward function is denoted by $R_{\mathbf{a}}(\hat{\mathbf{a}}) = \|\mathbf{h}^{(a)} - \mathbf{h}^{(\hat{a})}\|^2$. Specifically, the expected cumulative reward of the t-th step is defined by value function $Q(\hat{a}_t | \hat{\mathbf{a}}_{1:t-1}, \mathbf{z}) = E_{p_\theta(\hat{\mathbf{a}}_{t+1:L} | \hat{\mathbf{a}}_{1:t}, \mathbf{z})}(R_{\mathbf{a}}(\hat{\mathbf{a}}))$. We then estimate the value function $Q(\hat{a}_t | \hat{\mathbf{a}}_{1:t-1}, \mathbf{z})$ at the t-th step by utilizing the Monte-Carlo sampling, given by

$$Q(\hat{a}_t | \hat{\mathbf{a}}_{1:t-1}, \mathbf{z}) \approx \begin{cases} \frac{1}{J} \sum_{n=1}^{J} R_{\mathbf{a}}([\hat{\mathbf{a}}_{1:t}, \hat{\mathbf{a}}_{t+1:L}^{(n)}]), & t < L \\ R_{\mathbf{a}}([\hat{\mathbf{a}}_{1:t-1}, \hat{a}_t]). & t = L \end{cases} \tag{11}$$

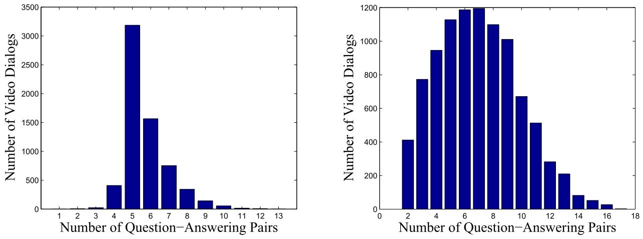
Where we randomly sample an answer sequence $\hat{\mathbf{a}}_{t+1:L}^{(n)}$ starting from the $t+1$-th step according to current state and repeat J times to calculate average as the expected cumulative reward. Following the policy gradient theorem of reinforcement learning, the gradients of the proposed decoder network is given by

$$\nabla_\theta \mathcal{L}_{RL}(g(\mathbf{z})) = \sum_{t=1}^{L} \nabla_\theta \log p_\theta(\hat{a}_t | \hat{\mathbf{a}}_{1:t-1}, \mathbf{z})) Q(\hat{a}_t | \hat{\mathbf{a}}_{1:t-1}, \mathbf{z}). \tag{12}$$

The overall reinforced decoder network learning is illustrated in Algorithm 1.

IV. Experiments

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



(a) (b)

Fig. 3. Statistic of question-answering pair numbers in the TACoS-MultiLevel and YouTubeClip datasets. (a) TACoS-MultiLevel. (b) YouTubeClip.

for multi-turn video question answering. In Section IV-A, we introduce the datasets and the implementation details. In Section IV-B, we introduce the evaluation criteria adopted in multiple-choice and open-ended experiments. In Section IV-C, we compare the proposed approach with state-of-the-art methods. In Section IV-D, we discuss the proposed approach in detail.

A. Experimental Settings

1) Dataset: We construct the multi-turn video question answering datasets from YouTubeClips [47] and TACoS-MultiLevel [48]. The YouTubeClips dataset consists of 1,987 videos and the TACoS-MultiLevel dataset is composed of 1303 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 video dialogs. We eventually have 9935 video dialogs for the TACoS-MultiLevel dataset and 6515 ones for the YouTubeClips dataset. In detail, the question-answering pair numbers of video dialogs are summarized in Figure 3. We have 37 228 video question answering pairs for the TACoS-MultiLevel dataset and 66 806 ones for the YouTubeClips dataset. Most of video dialogs in the TACoS-MultiLevel dataset have five turns of conversation and question-answering pair number of video dialogs in the YouTubeClips dataset is mostly between three and twelve. We take 90% of constructed video dialogs as the training data, 5% as the validation data and 5% as the testing ones.

For the multiple-choice method, we compute the semantic similarity between its ground-truth answer of each video dialog and all other answers based on the Euclidean distance with the pre-trained glove embedding [49], and then rank the top 50 answers as the candidate answer set. For the open-ended method, we generate the natural language answer word-by-word and evaluate the loss with its ground-truth answer by the reinforcement learning framework. The constructed multi-turn video question answering datasets will be provided later.

2) Implementation Details: We process the multi-turn 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 [50], and take the 4,096-dimensional feature vector for each frame. Meanwhile, we define a segment as a set of consecutive 16 frames in videos, where each segment overlaps 8 frames with adjacent segments. We then resize each frame in segments to 112 × 112 and employed the pre-trained 3D-ConvNet [38] to extract the 4096-dimensional segment video representation vector for each segment. We then employ the pre-trained word2vec model [51] to extract the semantic representation of questions and answers. Specifically, the size of the vocabulary set is 6500 and the dimension of the word vector is set to 256.

In the training process, we use the Adam optimizer [52] to minimize the loss for all models, where the initial learning rate is set to 0.001 and the exponential decay rate is set to 0.8. To prevent the gradient is too large in the back propagation, a gradient clipping method is utilized to limit gradient norms within 1.0. We adopt the mini-batch strategy in training and the batch size is set to 64. And we apply an early stopping technology to stop the training process when the performance no longer improves in the validation dataset.

B. Evaluation Criteria

We predict the final answer in multiple-choice form and open-ended form. Thus, we adopt different evaluation criteria to validate the performance of the proposed HACRN method for each form.

For multiple-choice form, we evaluate the performance of the proposed HACRN method based on three ranking evaluation criteria MRR, P@K and MeanRank, which have been widely used in visual question answering [1], [5], [6]. Given the testing question $\mathbf{q} \in Q$ with its ground-truth answer $\mathbf{a}$, we denote the rank of the ground-truth answer for question $\mathbf{q}$ by $r_{\mathbf{a}}^{\mathbf{q}}$. We now introduce the evaluation criteria below.

- *MRR* The MRR measures the ranking quality for the ground-truth answer by an algorithm. The MRR measure is given by

$$MRR = \frac{1}{|Q|} \sum_{\mathbf{q} \in Q} \frac{1}{r_{\mathbf{a}}^{\mathbf{q}}},$$

  where $|Q|$ is the number of testing questions used.
- *P@K* The P@K measures the ranking precision of the top-ranked answers by an algorithm. The P@K measure is given by

$$P@K = \frac{\sum_{\mathbf{q} \in Q} 1[r_{\mathbf{a}}^{\mathbf{q}} \leq K]}{|Q|},$$

  where $1[\cdot]$ is the indicator function.
- *MeanRank* The MeanRank measures the average rank position of the ground-truth answer by an algorithm. The MeanRank measure is given by

$$MeanRank = \frac{\sum_{\mathbf{q} \in Q} r_{\mathbf{a}}^{\mathbf{q}}}{|Q|}.$$

For open-ended form, we evaluate the performance of the proposed HACRN method based on two evaluation criteria Accuracy [8] and WUPS [11], which have been widely adopted in open-ended visual question answering. Given the testing question $\mathbf{q} \in Q$ and its ground-truth answer $\mathbf{a} = \{a_1, a_2, \ldots, a_L\}$, we denote the generated answers $\mathbf{o}$ from our HACRN method by $\mathbf{o} = \{o_1, o_2, \ldots, o_L\}$. We now show the evaluation criteria below.

- *Accuracy* The Accuracy is the normalized criterion to evaluate the quality of the generated answer. Given the testing question set Q, the Accuracy score is calculated by

$$Accuracy = \frac{1}{|Q|}\sum_{\mathbf{q}\in Q}\Big(\prod_{i=1}^{L}\mathbf{1}[a_i = o_i]\Big),$$

where $Accuracy = 1$ means that the generated answer is exactly the same as the ground-truth answer word-by-word and $Accuracy = 0$ for any other case.

- *WUPS* The WUPS is a soft evaluation criterion based on the WUP [53] score, which measures word similarity based on WordNet [54]. Thus, Given the testing Q, the WUPS score with the threshold γ is calculated by

$$\mathrm{WUPS} = \frac{1}{|Q|}\sum_{\mathbf{q}\in Q}\min\{\prod_{a_i\in\mathbf{a}}\max_{o_j\in\mathbf{o}} WUP_\gamma(a_i, o_j), \prod_{o_i\in\mathbf{o}}\max_{a_j\in\mathbf{a}} WUP_\gamma(o_i, a_j)\},$$

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

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

where the threshold γ is set to 0 and 0.9 according to the experimental setting in [11] and we denote the two WUPS evaluation criteria by WUPS@0.0 and WUPS@0.9, respectively.

C. Performance Comparisons

We extend the existing single-turn video question answering algorithms as the baseline algorithms for the problem of multi-turn video question answering. Each algorithm has a multiple-choice form and an open-ended form, where a normal LSTM recurrent answer generator is added to the end of models for open-ended form.

- *ESA* method is the single-turn video question answering algorithm [1], which learns the joint video representation based on the given question with attention mechanisms.
- *ESA+* method is the extension of ESA algorithm [1], where we add the hierarchical 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 [6], where we add the hierarchical LSTM network for conversation context modeling and then devise the dual-LSTM method to fuse the conversation context and video contents for multi-turn video question answering.
- *STAN+* method is the extension of STAN algorithm [5], where we add the hierarchical LSTM network for context-aware question understanding and then perform spatio-temporal attention with context-aware question representation for multi-turn video question answering.
- *CDMN+* method is an extension of CDMN algorithm [24], where we add the hierarchical LSTM network to model the conversation context and then propose a motion-appearance co-memory network to simultaneously learn the motion and appearance features for multi-turn video question answering.

TABLE I

EXPERIMENTAL RESULTS OF MULTIPLE-CHOICE ON TACOS-MULTILEVEL DATASET

Method	MRR	P@1	P@5	MeanRank
ESA	0.411	0.298	0.515	11.964
ESA+	0.411	0.300	0.507	10.435
STVQA+	0.427	0.305	0.540	9.762
STAN+	0.452	0.319	0.594	8.401
CDMN+	0.454	0.317	0.597	8.376
HACRN$_{(w/o.t)}$	0.444	0.319	0.579	8.726
HACRN$_{(w/o.s)}$	0.452	0.324	0.583	8.622
HACRN$_{(w/o.m)}$	0.512	**0.391**	0.643	6.625
HACRN$_{(rp)}$	**0.526**	0.386	**0.682**	**5.804**

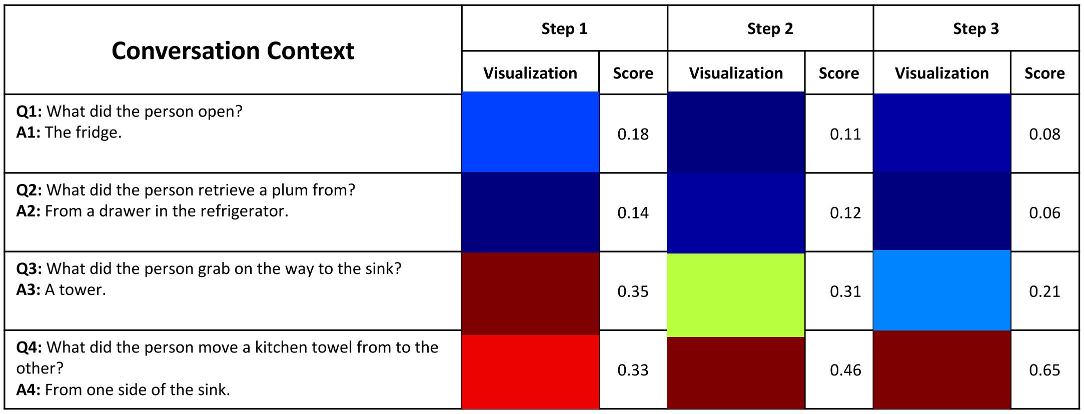
TABLE II

EXPERIMENTAL RESULTS OF MULTIPLE-CHOICE ON YOUTUBECLIP 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
CDMN+	0.422	0.278	0.584	7.074
HACRN$_{(w/o.t)}$	0.443	0.283	0.635	6.149
HACRN$_{(w/o.s)}$	0.454	0.295	0.636	6.042
HACRN$_{(w/o.m)}$	0.469	**0.315**	0.661	5.792
HACRN$_{(rp)}$	**0.470**	0.306	**0.670**	**5.496**

Unlike the previous video question answering works, our HACRN method performs the context-aware question understanding and learns multi-stream attention video representation with multi-step reasoning process for the problem. To exploit the effect of multi-stream attention process, we denote the method without the frame-level hierarchical spatio-temporal attention context network by **HACRN$_{\mathbf{(w/o.s)}}$**, and the method without the segment-level temporal attention context network by **HACRN$_{\mathbf{(w/o.t)}}$**. Next, to validate the effect of the multi-step reasoning process, we denote the multi-stream method without multi-step reasoning process by **HACRN$_{\mathbf{(w/o.m)}}$**, and the multi-stream method with multi-step reasoning process by **HACRN$_{\mathbf{(rp)}}$**. Moreover, to demonstrate the effect of reinforced decoder networks in open-ended form, we denote our method with reinforced decoder networks by **HACRN$_{\mathbf{(rl)}}$** and the one without reinforced decoder networks by **HACRN$_{\mathbf{(rp)}}$** as before.

Table I shows the multiple-choice experimental results of the methods on MRR, P@1, P@5 and MeanRank using TACoS-MultiLevel dataset. Table II demonstrates the multiple-choice evaluation results of the methods using YoutubeClip dataset. In the same way, table III and table IV show open-ended experimental results of the methods using TACoS-MultiLevel dataset and YoutubeClip dataset, respectively. The hyper-parameters 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



Conversation Context	Step 1		Step 2		Step 3	
	Visualization	Score	Visualization	Score	Visualization	Score
Q1: What did the person open? **A1:** The fridge.	🟦	0.18	🟦	0.11	🟦	0.08
Q2: What did the person retrieve a plum from? **A2:** From a drawer in the refrigerator.	🟦	0.14	🟦	0.12	🟦	0.06
Q3: What did the person grab on the way to the sink? **A3:** A tower.	🟥	0.35	🟩	0.31	🟦	0.21
Q4: What did the person move a kitchen towel from to the other? **A4:** From one side of the sink.	🟥	0.33	🟥	0.46	🟥	0.65

Question: What did the person turn on? Answer: The sink.

Fig. 4. The Conversation context attention results of multi-step reasoning Process.

TABLE III
EXPERIMENTAL RESULTS OF OPEN-ENDED ON TACOS-MULTILEVEL DATASET

Method	Accuracy	WUPS@0.0	WUPS@0.9
ESA	0.121	0.275	0.154
ESA+	0.132	0.287	0.172
STVQA+	0.154	0.300	0.177
STAN+	0.158	0.312	0.183
CDMN+	0.169	0.323	0.189
HACRN$_{(w/o.t)}$	0.174	0.328	0.200
HACRN$_{(w/o.s)}$	0.182	0.329	0.213
HACRN$_{(w/o.m)}$	0.175	0.333	0.205
HACRN$_{(rp)}$	0.191	0.344	0.214
HACRN$_{(rl)}$	**0.199**	**0.350**	**0.221**

TABLE IV
EXPERIMENTAL RESULTS OF OPEN-ENDED ON YOUTUBECLIP DATASET

Method	Accuracy	WUPS@0.0	WUPS@0.9
ESA	0.058	0.190	0.076
ESA+	0.067	0.223	0.098
STVQA+	0.079	0.245	0.107
STAN+	0.089	0.263	0.115
CDMN+	0.094	0.271	0.119
HACRN$_{(w/o.t)}$	0.105	0.293	0.124
HACRN$_{(w/o.s)}$	0.100	0.298	0.127
HACRN$_{(w/o.m)}$	0.102	0.290	0.132
HACRN$_{(rp)}$	0.107	0.300	0.134
HACRN$_{(rl)}$	**0.113**	**0.320**	**0.139**

methods on each evaluation criterion. The experimental results reveal a number of interesting points:

- The methods based on context-aware question understanding, ESA+, STVQA+, STAN+, CDMN+, HACRN$_{(w/o.t)}$, HACRN$_{(w/o.s)}$, HACRN$_{(w/o.m)}$, HACRN$_{(rp)}$, HACRN$_{(rl)}$ outperform the single-turn video question answering method ESA, which suggests the context-aware question representation is critical for the problem.
- The methods based on hierarchical attention context network learning, HACRN$_{(w/o.t)}$, HACRN$_{(w/o.s)}$, HACRN$_{(w/o.m)}$, HACRN$_{(rp)}$, HACRN$_{(rl)}$ outperform all baselines, which validates the video information retrieval and conversation context modeling ability of our HACRN method.
- The multi-stream hierarchical attention context network method HACRN$_{(rp)}$ achieves better performance than the methods HACRN$_{(w/o.t)}$ and HACRN$_{(w/o.s)}$. This suggests that both the frame-level and segment-level attention mechanisms are important for the problem of multi-turn video question answering.
- The HACRN$_{(rp)}$ method with multi-step reasoning process outperforms the HACRN$_{(w/o.m)}$ method without it, which demonstrates the multi-step reasoning process can improve the performance of multi-turn video question answering.
- In the open-ended case, our HACRN$_{(rl)}$ method achieves the better performance than the method HACRN$_{(rp)}$. This fact shows that the reinforced decoder network is effective for the problem of open-ended multi-turn video question answering.

These analyses show that each component of multi-stream hierarchical attention context reinforced networks is necessary for multi-turn video question answering, including the context-aware question understanding, multi-stream video attention mechanisms, multi-step reasoning process and reinforce decoder network.

D. In-Depth Analysis of the Proposed Framework

1) Qualitative Analysis: To demonstrate how the attention mechanisms works, we display conversation context attention results in Figure 4 and video attention results in Figure 5. As we can see in Figure 4, conversation context attention results of multi-step reasoning process are visualized using a thermodynamic diagram [55] and the accurate attention scores are shown by numerical scores. Following the reasoning process, we observe that the model pays more attention to the question-answer pair more relevant to the given query. In Figure 5, we show the video attention distribution by two curves, where the frame-level video attention is denoted by the red dashed curve and the segment-level video attention is

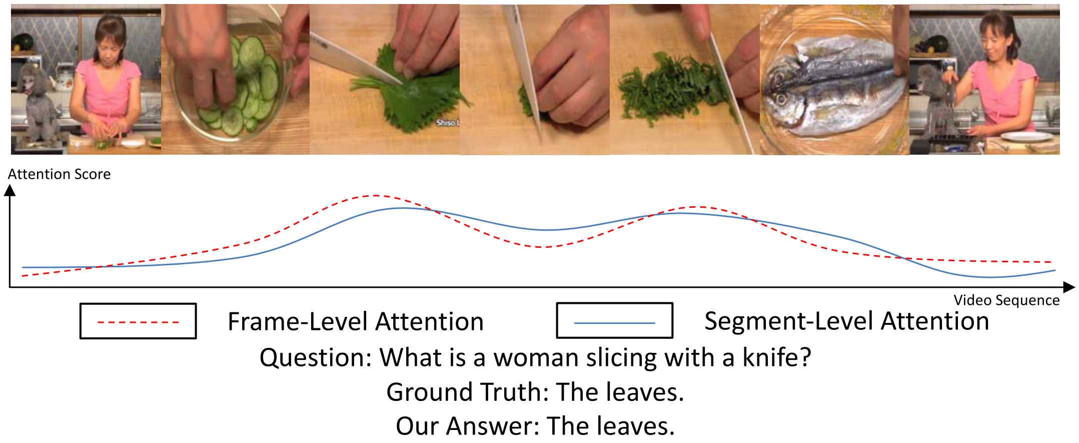
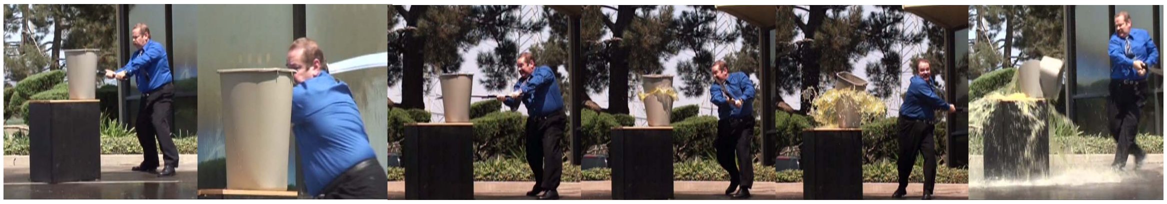


Fig. 5. The Video attention results of frame-level and segment-level.

TABLE V

A EXAMPLE OF MULTIPLE-CHOICE EXPERIMENTAL RESULTS



Conversation Context	Question	Answer
Q1: What is the man holding in his hand? **A1:** A sword. **Q2:** What does the man wear? **A2:** A blue shirt. **Q3:** Where is the man standing? **A3:** Besides a garbage can. **Q4:** What does the man slice with a sword? **A4:** A garbage can.	What flows out from it?	**ESA:** To a pot of water **ESA+:** A liquid in the bucket. **STVQA+:** Water. **STAN+:** A drop of liquid. **CDMN+:** Water. **HACRN:** Yellow water. **Ground Truth:** Yellow water.

denoted by the blue solid curve. The curves are drawn based on the real computed attention weights and then smoothed, where the horizontal axis represents the temporal sequence over the video and the vertical axis represents the absolute value of attention score for each frame or segment, varying from 0 to 1. As shown, the video attentions are localized on the important frames and segments according to the question. In addition, the segment-level video attention curve has a similar trend as the frame-level video attention curve but is more stable relatively.

Furthermore, we display a typical example of multiple-choice in Table V and an open-ended example in Table VI. According to conversation context and video information, we show the predicted or generated answers of all methods. Compared with the ground-truth answer, we intuitively observe that our HACRN method achieves better performance than all other algorithms.

2) Hyper-Parameter Analysis: In our approach, there are two essential hyper-parameters, which are the dimension of the hidden states in LSTM networks, and the dimension of the joint representation of the question-answer pairs in the conversation context. In fact, the joint representation dimension corresponds to d_u and the LSTM dimension corresponds to d_h in Section III. We vary the LSTM dimension from 32, 64, . . ., to 512, and the joint representation dimension from 32, 64, . . ., to 512. For multiple-choice form, We first investigate the effect of the LSTM dimension on MRR and P@1 on TACoS-MultiLevel and YoutubeClip datasets in Figures 6. We then show the effect of joint representation dimension on TACoS-MultiLevel dataset and YoutubeClip dataset in Figures 7. Similarly, for open-ended form, Figures 8 and Figures 9 show the effect of the LSTM dimension and joint representation dimension on Accuracy and WUPS@0.0 based on two datasets, respectively.

From these figures, we can observe that the hyper-parameters have similar effects on different evaluation criteria for the same dataset, such as MRR and P@1 on the TACoS-MultiLevel and YoutubeClip datasets. Moreover, the influences of the LSTM dimension are relatively stable for the performance of multi-turn video question answering, but the

TABLE VI

A EXAMPLES OF OPEN-ENDED EXPERIMENTAL RESULTS

Conversation Context	Question	Answer
Q1: How many people are there in room? **A1:** Three people. **Q2:** What is the woman in white doing? **A2:** The woman is posing. **Q3:** What is the woman in brown doing? **A3:** The woman is looking at the others. **Q4:** What is the man holding in his hand? **A4:** A camera.	What is the man doing with it?	**ESA:** The man is seeing. **ESA+:** The photographer. **STVQA+:** Pictures os the women. **STAN+:** A man holds a camera. **CDMN+:** A man is taking a camera. **HACRN:** The man is taking pictures. **Ground Truth:** The man is taking photographs.

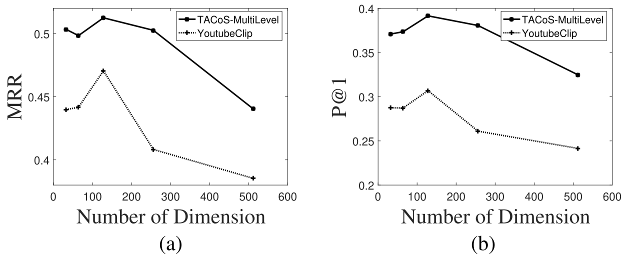


Fig. 6. Effect of the LSTM dimension on MRR and P@1 using TACoS-MultiLevel and YoutubeClip datasets. (a) MRR. (b) P@1.

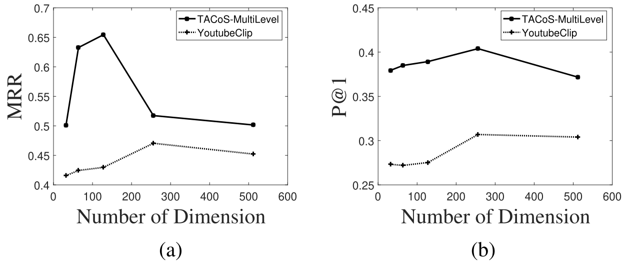


Fig. 7. Effect of joint representation dimension on MRR and P@1 using YoutubeClip and TACoS-MultiLevel datasets. (a) MRR. (b) P@1.

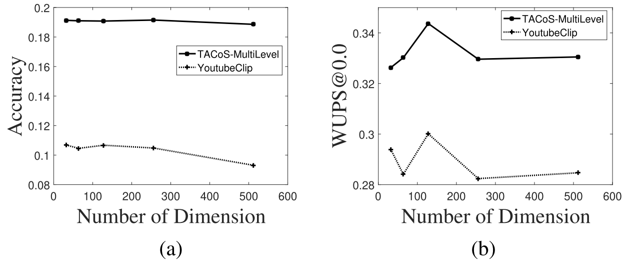


Fig. 8. Effect of the LSTM dimension on accuracy and WUPS@0.0 using TACoS-MultiLevel and YoutubeClip datasets. (a) Accuracy. (b) WUPS@0.0.

joint representation dimension has different effects for the multiple-choice and open-ended forms. For multiple-choice form, to obtain the best performance, we set the LSTM dimension to 128 and the joint representation dimension to 256. And for open-ended form, the better hyper-parameters are 128 and 32 for the LSTM dimension and joint representation dimension, respectively.

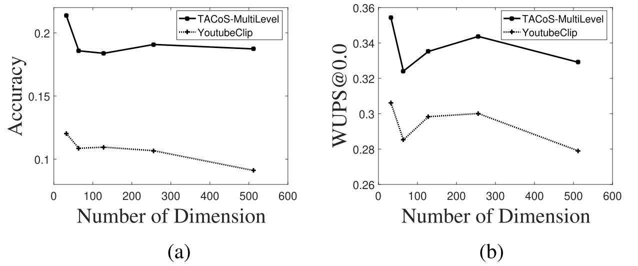


Fig. 9. Effect of joint representation dimension on accuracy and WUPS@0.0 using TACoS-MultiLevel and YoutubeClip datasets. (a) Accuracy. (b) WUPS@0.0.

V. CONCLUSION

In this paper, we study the problem of multi-turn video question answering from the viewpoint of multi-stream hierarchical attention context reinforced network learning. We first propose the hierarchical attention context learning method with recurrent neural networks for context-aware question understanding 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 to further improve the performance of multi-turn video question answering. Besides the conventional multiple-choice form, we further develop the reinforced decoder network to generate the open-ended natural language answer and explore the utilization of the reinforcement learning framework in such a challenging task. We construct two large-scale multi-turn video question answering datasets and evaluate the effectiveness of our proposed method through extensive experiments.

REFERENCES

[1] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun, "Leveraging video descriptions to learn video question answering," in *Proc. AAAI*, 2017, pp. 4334–4340.

[2] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering temporal context for video question and answering," *Proc. Int. J. Comput. Vis.*, 2017.

[3] A. Mazaheri, D. Zhang, and M. Shah. (2016). "Video fill in the blank with merging LSTMs." [Online]. Available: https://arxiv.org/abs/1610.04062

[4] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4631–4640.

[5] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang, "Video question answering via hierarchical Spatio-temporal attention networks," in *Proc. 26th Int. Joint Conf. Artificial Intell.*, Aug. 2017, pp. 3518–3524.

[6] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2758–2766.

[7] Z. Zhao, X. Jiang, D. Cai, J. Xiao, X. He, and S. Pu, "Multi-turn video question answering via multi-stream hierarchical attention context network," in *Proc. 27th Int. Joint Conf. Artificial Intell.*, Jul. 2018, pp. 3690–3696.

[8] S. Antol *et al.*, "Vqa: Visual question answering," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[9] R. Li and J. Jia, "Visual question answering with question representation update (QRU)," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 4655–4663.

[10] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[11] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 1682–1690.

[12] J.-H. Kim *et al.*, "Multimodal residual learning for visual QA," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 361–369.

[13] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3242–3250.

[14] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5585–5599, Nov. 2018.

[15] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 289–297.

[16] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4613–4621.

[17] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[18] B. Patro and V. P. Namboodiri, "Differential attention for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7680–7688.

[19] A. Das *et al.*, "Visual dialog," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1080–1089.

[20] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. V. D. Hengel. (2016). "Visual question answering: A survey of methods and datasets." [Online]. Available: https://arxiv.org/abs/1607.05910

[21] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 773–786, Mar. 2019.

[22] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[23] Z. Zhao *et al.*, "Open-ended long-form video question answering via adaptive hierarchical reinforced networks," in *Proc. 27th Int. Joint Conf. Artificial Intell.*, Jul. 2018, pp. 3683–3689.

[24] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6576–6585.

[25] I. V. Serban *et al.*, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proc. AAAI*, 2017, pp. 3295–3301.

[26] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. AAAI*, 2016, pp. 3776–3784.

[27] J. E. Weston, "Dialog-based language learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 829–837.

[28] I. V. Serban *et al.*, "Multiresolution recurrent neural networks: An application to dialogue response generation," in *Proc. AAAI*, 2017, pp. 3288–3294.

[29] C. Xing, Y. Wu, W. Wu, Y. Huang and M. Zhou, "Hierarchical recurrent attention network for response generation," in *Proc. AAAI*, 2018, pp. 5610–5617.

[30] H. Mei, M. Bansal, and M. R. Walter, "Coherent dialogue with attention-based language models," in *Proc. AAAI*, 2017, pp. 3252–3258.

[31] G. Zhou, P. Luo, R. Cao, F. Lin, B. Chen, and Q. He, "Mechanism-aware neural machine for dialogue response generation," in *Proc. AAAI*, 2017, pp. 3400–3407.

[32] Y. Wu, W. Wu, M. Zhou, and Z. Li, "Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots," in *Proc. 55th Ann. Meeting Assoc. Comput.*, 2017.

[33] J. D. Williams, K. Asadi, and G. Zweig, "Hybrid code networks: Practical and efficient end-to-end dialog control with supervised and reinforcement learning," in *Proc. 55th Ann. Meeting Assoc. Comput. Linguistics*, 2017, pp. 496–505.

[34] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, "Deep reinforcement learning for dialogue generation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, Nov. 2016, pp. 1192–1202.

[35] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. (2017). "Adversarial learning for neural dialogue generation." [Online]. Available: https://arxiv.org/abs/1701.06547

[36] B. Dhingra *et al.*, "Towards end-to-end reinforcement learning of dialogue agents for information access," in *Proc. ACL*, Jul. 2017, pp. 484–495.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 2012, pp. 1097–1105.

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] X. Zhou, B. Hu, Q. Chen, B. Tang, and X. Wang, "Answer sequence learning with neural networks for answer selection in community question answering," in *Proc. 53rd Ann. Meeting Assoc. Comput. Linguistics*, July. 2015, pp. 713–718.

[41] G. B. Orr, and K. R. Müller, *Neural Networks: Tricks of the Trade* (Lecture Notes in Computer Science), vol. 7700. Berlin, Germany: Springer, 2003.

[42] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2440–2448.

[43] D. Bahdanau *et al.* (2016). "An actor-critic algorithm for sequence prediction." [Online]. Available: https://arxiv.org/abs/1607.07086

[44] J. Qi and Y. Peng, "Cross-modal bidirectional translation via reinforcement learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 2630–2636.

[45] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4213–4222.

[46] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[47] D. L. Chen and W. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Ann. Meeting Assoc. Comput. Linguistics*, Jun. 2011, pp. 190–200.

[48] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele, "Coherent multi-sentence video description with variable level of detail," in *Proc. German Conf. Pattern Recognit.* Springer, 2014, pp. 184–195.

[49] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[50] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition," [Online]. Available: https://arxiv.org/abs/1409.1556

[51] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013). "Efficient estimation of word representations in vector space." [Online]. Available: https://arxiv.org/abs/1301.3781

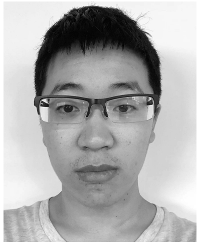
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

[53] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” in *Proc. 32nd Ann. Meeting Assoc. Comput. Linguistics*, 1994, pp. 133–138.

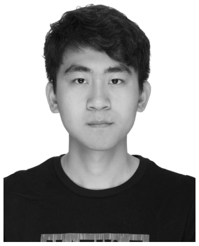
[54] C. Fellbaum, *WordNet*. Hoboken, NJ, USA: Wiley, 1998.

[55] C.-A. Palma *et al.*, “Visualization and thermodynamic encoding of single-molecule partition function projections,” *Nat. Commun.*, vol. 6, p. 6210, Feb. 2015.

Zhou Zhao received the B.S. and Ph.D. degrees in computer science from The Hong Kong University of Science and Technology, in 2010 and 2015, respectively. He is currently an Associate Professor with the College of Computer Science, Zhejiang University. His research interests include machine learning and data mining.



Zhu Zhang received the B.E. degree in computer science and technology from Zhejiang University, China, in 2018, where he is currently pursuing the master’s degree in computer science. His research interests include machine learning and computer vision.



Xinghua Jiang received the B.E. degree in computer science and technology from Zhejiang University, China, in 2016, where he is currently pursuing the master’s degree in computer science. His research interests include machine learning and computer vision.

Deng Cai received the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign in 2009. He is currently a Professor with the State Key Laboratory of CAD&CG, College of Computer Science, Zhejiang University, China. His research interests include machine learning, data mining, and information retrieval.