This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2922062, IEEE Transactions on Image Processing



# Long-Form Video Question Answering via Dynamic Hierarchical Reinforced Networks

Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhenxin Xiao, Xiaohui Yan, Jun Yu, Deng Cai, and Fei Wu

***Abstract*—Open-ended long-form video question answering is a challenging task in visual information retrieval, which automatically generates a natural language answer from the referenced long-form video contents according to a given question. However, existing works mainly focus on short-form video question answering, due to the lack of modeling semantic representations from long-form video contents. In this paper, we introduce a dynamic hierarchical reinforced network for open-ended long-form video question answering, which employs an encoder-decoder architecture with a dynamic hierarchical encoder and a reinforced decoder. Concretely, we first propose a frame-level dynamic LSTM network with binary segmentation gate to learn frame-level semantic representations according to the given question. We then develop a segment-level highway LSTM network with question-aware highway gate for segment-level semantic modeling. Next, we devise the reinforced decoder with a hierarchical attention mechanism to generate natural language answers. We construct a large-scale long-form video question answering dataset. The extensive experiments on the long-form dataset and another public short-form dataset show the effectiveness of our method.**

***Index Terms*—Long-Form Video Question Answering, Hierarchical, Dynamic, 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. Open-ended video question answering is the essential problem of visual question answering, which automatically generates the natural language answer from the referenced video contents according to the given question. Different from multiple-choice answer prediction, which depends on appropriate candidates, open-ended answering can cover the complex scenes in the real-world application. Currently, existing video question answering approaches [1]–[4] mainly focus on the problem of short-form video question answering, where the video length is often a few seconds [4]. These methods learn video semantic representations from the recurrent neural network layer, and then generate the answer to the given question. As videos on the Internet tend to be very long, long-form video question answering has become a very practical and challenging problem, where the videos last a few minutes or even longer. Although these short-form works have achieved promising performance in short-form video question answering, they may still be ineffectively applied to the long-form video question answering due to the lack of modeling semantic representations from long-form video contents.

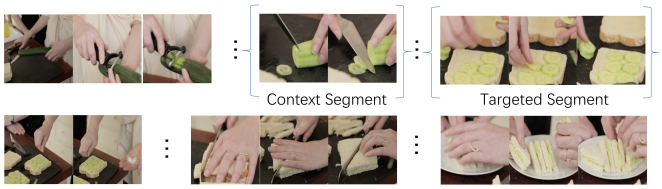



Fig. 1. Open-Ended Long-Form Video Question Answering

The long-form video contents often contain the evolving complex object interactions through frames, which have long-range semantic dependencies [5]. We illustrate a simple example in Figure 1. The answer generation to question "what does the woman do after slicing the cucumber?" requires the sufficient video semantic understanding and precise localization of the targeted segment from long-form video contents. Thus, the simple extension of the existing video question answering works based on frame-level recurrent neural networks is difficult for modeling the semantic representations from long-form video contents according to the given question [6].

Recently, the hierarchical neural encoder [7] has been proposed to learn the segment-level semantic representations with a fixed segment length. Similarly, we employ the hierarchical neural encoder structure with an attention mechanism to learn the joint semantic representations from long-form video contents according to the given question. However, the current hierarchical neural encoder method depends on the fixed-length segmentation of video, which is unreasonable for varying video contents. Although the video frames are topically consistent, they have different semantic contents [5] and can be divided into different video segments of variable length. Inspired by binary neurons [8], we develop the dynamic hierarchical encoder that unifies the frame-level dynamic video segmentation and hierarchical video modeling into a joint learning framework. On the other hand, long-form video contents generally contain quite a lot of redundant information irrelevant to the given question. Inspired by the highway network [9], we add the question-aware highway gate into the segment-level LSTM network of the hierarchical neural encoder to filter the redundant information and learn segment-

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

X. Yan is with the Poisson Lab, Huawei Technologies, Beijing 100080, China (e-mail: yanxiaohui2@huawei.com).

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

J. Yu is with College of Computer Science, Hangzhou Dianzi University, Hangzhou 310027, China (e-mail: yujun@hdu.edu.cn).

level semantic representations. Thus, to achieve high-quality long-form video question answering, we leverage the hierarchical neural encoder with frame-level dynamic segmentation and segment-level question-aware highway gate.

Currently, the reinforcement learning framework has attracted a lot of attention and shown the effectiveness on the text generation task [10], [11]. We develop the reinforcement learning process to enhance the natural language generation method based on the common neural decoder network. Furthermore, inspired by the way humans first observe the question and then look for relevant video contents based on the keywords of the question, we employ a hierarchical attention method during the answer generation process. Thus, we develop a normal LSTM decoder with hierarchical attention mechanism and trained with reinforcement learning.

In this paper, we extend our previous work [1] and study the problem of open-ended long-form video question answering from the viewpoint of dynamic hierarchical reinforced network learning. Following our previous work [12], we first employ the dynamic hierarchical encoder structure. Specifically, we propose the frame-level dynamic LSTM network with binary segmentation gate to learn frame-level semantic representations according to the given question. Different from our previous paper, we then develop the segment-level highway LSTM network with question-aware highway gate to model segment-level semantic representations. For the decoding process, we devise the reinforced decoder network with hierarchical attention mechanism unused in our previous work to generate the natural language answer for open-ended video question answering. We name the dynamic hierarchical reinforced network learning framework as DHRN. When a certain question is issued, DHRN can generate a natural language answer for it based on the referenced video contents. The main contributions of this paper are as follows:

- Unlike the previous studies, we present the problem of open-ended long-form video question answering from the viewpoint of dynamic hierarchical reinforced network learning, which models the long-range semantic dependencies from long-form video contents and generates the high-quality natural language answer according to the given question.
- We develop the dynamic hierarchical encoder, composed of the frame-level dynamic LSTM network and the segment-level highway LSTM network, to dynamically segment long-form video contents and learn video semantic representations. We then devise the reinforced decoder network with a hierarchical attention mechanism to generate the natural language answer.
- We construct a large-scale dataset for open-ended long-form video question answering and validate the effectiveness of our proposed method through extensive experiments on the long-form dataset and another public short-form dataset.

The rest of this paper is organized as follows. We briefly review some related work about visual question answering in Section II. In Section III, we introduce the problem of open-ended long-form video question answering from the viewpoint of dynamic hierarchical reinforced network learning. We then present a variety of experimental results in Section IV. Finally, Section V concludes this paper.

[1]This work is the extension of our previous paper [12], 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.

## II. Related Work

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

The visual question answering task is to provide an accurate answer for the natural language question from the referenced visual contents [13]. Different from other visual information retrieval tasks [14], [15], visual question answering requires the comprehensive understanding of cross-modal information from visual and textual contents. From the viewpoint of the form of answers, the multiple-choice visual question answering [16] chooses a correct answer from the given answer candidate set according to the visual contents and question, and open-ended visual question answering [17], [18] directly generates the natural language answer, which is more complex and universal.

The existing approaches for visual question answering can be categorized into image-based question answering methods [13], [18]–[28] and video-based question answering ones [1]–[4], [29], [29]–[36].

### A. Image Question Answering

In the field of image question answering, Malinowski et al. [18] develop the multi-world probabilistic approach for open-ended image question answering. Kim et al. [19] employ the multi-modal residual network, which extends the deep residual network framework and effectively models the joint representation from visual and language information. To exploit the complex image question answering task, QRU method [20] is proposed with the reasoning process that iteratively selects the relevant image regions and updates the question representation.

With the development of attention mechanism, Shih et al. [21] propose the spatial-attention mechanism that selects the relevant image regions to the given question. Lu et al. [22] devise the co-attention network to jointly reason about question and image attention. Yang et al. [23] develop the stacked attention method that advises multiple-step reasoning to locate the relevant image contents layer-by-layer for image question answering. Anderson et al. [24] present the combined attention mechanism including top-down and bottom-up to compute the attention about salient objects and image regions. Different from image-based attention, Patro et al. [25] propose an exemplar-based attention mechanism, which is close to human attention, to access a differential attention region.

Moreover, besides the conventional single-turn image question answering, the visual dialog task [26] provides the answer based on the previous question-answering history and referenced video contents. A survey of existing image question answering methods can be found in [27].

### B. Video Question Answering

As a natural extension of image-based question answering, the video-based question answering has drawn lots of research interest and been proposed as a more challenging task.

The fill-in-the-blank approaches [29], [30] complete the missing entry in the video description by ranking candidate answers based on both visual contents and contextual video description. Tapaswi et al. [31] 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. And Wang et .al. [37] devise a layered memory network to represent movie contents with more semantic information in both frame-level and clip-level. Zeng et al. [1] extend the memory-based image question answering method with the additional LSTM layer for video question answering and Gao et al. [32] develop the dynamic memory network to model temporal semantic representations of video and answer questions.

Similar to the field of image question answering, attention mechanism is widely used in video question answering. Xue et al. [33] advise the co-attention reasoning network that applies the attention mechanism on both videos and questions. Xu et al. [3] propose the attentional reasoning network, which is based on the attention memory unit for generating the natural language answer. Considering the combination of spatial information in separate frames and motion information in video streams [38], Jang et al. [4] develop the spatio-temporal reasoning algorithm to employ the spatial and temporal attention on video. Zhao et al. [34] present the hierarchical dual-level attention network to learn the object static and dynamic information of video contents. And Ye et al. [35] propose the attribute-augmented attention network to jointly detect frame-level visual attribute.

With the development of the encoder-decoder learning framework, Gao et al. [36] encode the video contents by LSTM networks and then decode the answer to the given question by the language model. Zhao et al. [2] propose spatio-temporal attention encoder-decoder learning framework with the multi-step reasoning process for video question answering. And Zhu et al. [29] present the encoder-decoder approach to model temporal structures of video contents and employ the ranking loss to answer multiple-choice questions. Besides single-turn video question answering, Zhao et al. [39] propose a hierarchical attention method to solve the multi-turn video question answering task.

Although these works have achieved promising performance in short-form video question answering, they may still be ineffectively applied to the long-form video question answering due to the lack of modeling the semantic representations from long-form video contents. Unlike the previous studies, we study the problem of open-ended long-form video question answering from the viewpoint of dynamic hierarchical reinforced network learning.

## III. Long-Form Video Question Answering via Dynamic Hierarchical Reinforced Networks

In this section, we introduce the problem of open-ended long-form video question answering from the viewpoint of dynamic hierarchical reinforced network learning. We first propose the dynamic hierarchical encoder, composed of the frame-level dynamic LSTM network and the segment-level highway LSTM network, to hierarchically learn the frame-level and segment-level semantic representations and capture long-range semantic dependencies according to the question. We then develop the reinforced decoder with a hierarchical attention mechanism to generate the natural language answer. The overall framework of our proposed method is shown in Figure 2.

### A. The Problem

Before presenting the learning framework, we first introduce some basic notions and terminologies. We denote the question by $\mathbf{q} \in Q$, the video $\mathbf{v} \in V$ and the answer by $\mathbf{a} \in A$, where $Q$, $V$ and $A$ are the sets of questions, videos and answers, respectively. Since the video is composed of sequential frames, the frame-level representation of video $\mathbf{v}$ is given by $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N)$ of length $N$, where $\mathbf{v}_i$ is the $i$-th frame. We then encode the word-level representation of the question by $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_R)$ of length $R$, where $\mathbf{q}_r$ is the $r$-th word embedding. And the ground truth natural language answer is denoted by $\mathbf{a} = (a_1, a_2, \ldots, a_M)$ of length $M$, where $a_t$ is the $t$-th word token. In the frame-level dynamic LSTM network, we will divide the frame sequences into video segments of variable length by binary segmentation gate. We denote the collection of video segments by $\{S_1, S_2, \ldots, S_K\}$ of size $K$, where $S_j$ is the $j$-th set of segmented frames and $\mathbf{v}_i \in S_j$ means that the $i$-th frame belongs to segment $S_j$. The segment-level representation of segment $S_k$ is denoted by $\mathbf{s}_k$, and the sequential representation of video segments is given by $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$.

Since all the video, question and answer are sequential data with variant length, it is natural to choose the variant recurrent neural network called long-short term memory network (LSTM) [40] to learn their feature representation by

$$
\begin{aligned}
\mathbf{i}_t &= \delta(\mathbf{W}_i \mathbf{x}_t + \mathbf{G}_i \mathbf{h}_{t-1} + \mathbf{b}_i), \\
\hat{\mathbf{c}}_t &= tanh(\mathbf{W}_c \mathbf{x}_t + \mathbf{G}_t \mathbf{h}_{t-1} + \mathbf{b}_f), \\
\mathbf{f}_t &= \delta(\mathbf{W}_f \mathbf{x}_t + \mathbf{G}_f \mathbf{h}_{t-1} + \mathbf{b}_f), \\
\mathbf{c}_t &= \mathbf{i}_t \cdot \hat{\mathbf{c}}_t + \mathbf{f}_t \cdot \mathbf{c}_t, \\
\mathbf{o}_t &= \delta(\mathbf{W}_o \mathbf{x}_t + \mathbf{G}_o \mathbf{h}_{t-1} + \mathbf{V}_o \mathbf{c}_t + \mathbf{b}_o), \\
\mathbf{h}_t &= \mathbf{o}_t \cdot tanh(\mathbf{c}_t),
\end{aligned}
\tag{1}
$$

where $\delta$ represents the sigmoid activation function. The memory cell $\mathbf{c}_t$ maintains the history of the inputs observed up to the time step. Update operations on the memory cell are modulated by three gates $\mathbf{i}_t$, $\mathbf{f}_t$ and $\mathbf{o}_t$, which are all computed as a combination of the current input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$, followed by a sigmoid activation. Specifically, we denote the frame-level semantic representations of video $\mathbf{v}$ by $\mathbf{h}^{(f)} = (\mathbf{h}_1^{(f)}, \mathbf{h}_2^{(f)}, \ldots, \mathbf{h}_N^{(f)})$, segment-level semantic representations of video by $\mathbf{h}^{(s)} = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_K^{(s)})$, question semantic representations by $\mathbf{h}^{(q)} = (\mathbf{h}_1^{(q)}, \mathbf{h}_2^{(q)}, \ldots, \mathbf{h}_R^{(q)})$ and that of generated answer $\hat{\mathbf{a}}$ by $\mathbf{h}^{(\hat{a})} = (\mathbf{h}_1^{(\hat{a})}, \mathbf{h}_2^{(\hat{a})}, \ldots, \mathbf{h}_M^{(\hat{a})})$ using LSTM networks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2922062, IEEE Transactions on Image Processing



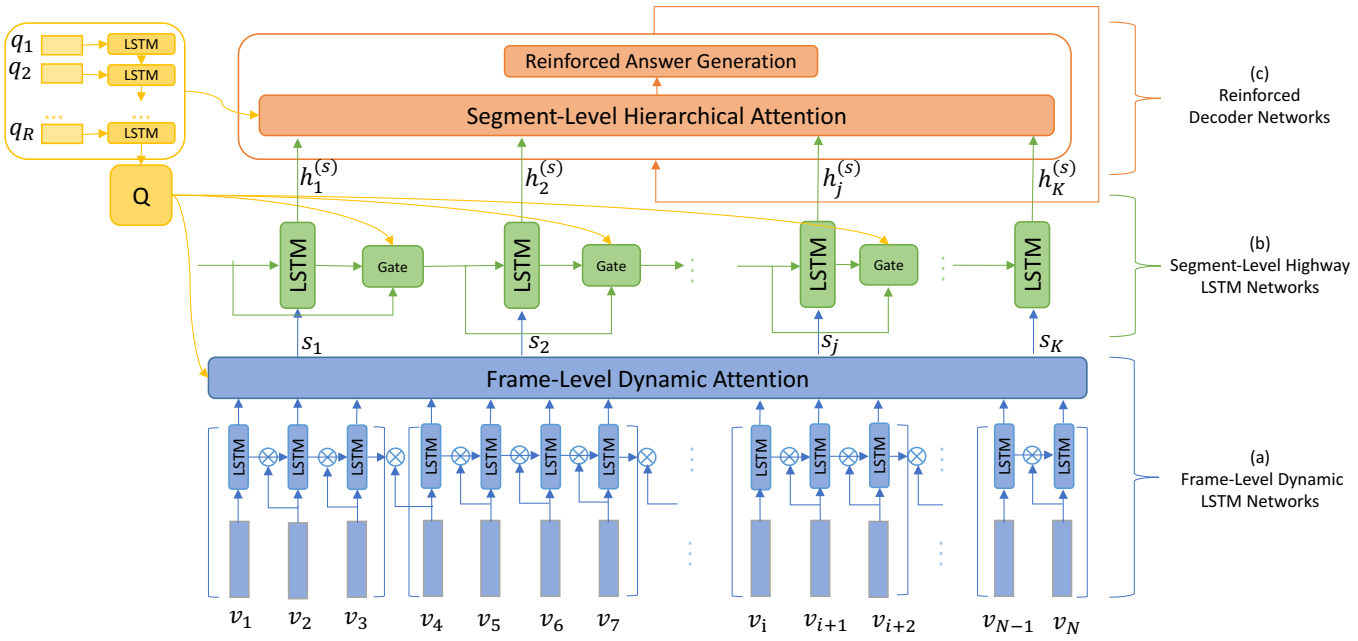



Fig. 2. The Framework of Dynamic Hierarchical Reinforced Networks for Open-Ended Long-Form Video Question Answering. (a) The frame-level dynamic LSTM networks segment the long-form video contents by binary segmentation gate and learn the frame-level semantic representations according to the given question. (b) The segment-level highway LSTM networks filter the redundant information by the question-aware highway gate and model the segment-level semantic representation. (c) The reinforced decoder networks with hierarchical attention mechanism generate the natural language answers.

Using the notations above, the problem of open-ended long-form video question answering is formulated as follows. Given the set of videos $V$, questions $Q$ and answers $A$, our goal is to learn the encoder-decoder network $g(f(\mathbf{v},\mathbf{q}),\mathbf{q})$ where the dynamic hierarchical encoder network $f(\mathbf{v},\mathbf{q})$ learns the joint semantic representations of the video and question, and the reinforced decoder network $g(f(\mathbf{v},\mathbf{q}),\mathbf{q})$ generates the natural language answer $\hat{\mathbf{a}} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_M)$ for open-ended long-form video question answering.

### B. Dynamic Hierarchical Encoder Networks

In this section, we propose the dynamic hierarchical encoder network $f(\mathbf{v},\mathbf{q})$ that unifies the frame-level dynamic LSTM network learning and segment-level highway LSTM network learning into a common framework, which performs the dynamic video segmentation and hierarchically learns frame-level and segment-level semantic representations to capture long-range semantic dependencies according to the question.

*1) Frame-Level Dynamic LSTM Networks:* The video contents often contain a number of frames with the targeted objects to the given question that evolves over time [41]. Inspired by binary neuron [8], we propose the frame-level dynamic LSTM network that segments the complex events in videos into separate segments of variable length to learn frame-level semantic representations and obtains the question-aware sequential representation of video segments according to the given question.

We first extract the frame-level video feature by $\mathbf{v}^{(f)} = (\mathbf{v}_1^{(f)}, \mathbf{v}_2^{(f)}, \ldots, \mathbf{v}_N^{(f)})$, and then learn their semantic representations $\mathbf{h}^{(f)} = (\mathbf{h}_1^{(f)}, \mathbf{h}_2^{(f)}, \ldots, \mathbf{h}_N^{(f)})$ using LSTM networks. To enable the video segmentation, we define an dynamic LSTM network with binary segmentation gate function, which decides whether to transfer the LSTM parameters (i.e., hidden state $\mathbf{h}_t^{(f)}$ and memory cell $\mathbf{c}_t^{(f)}$) of the current frame to update the LSTM parameters of the next frame (i.e., hidden state $\mathbf{h}_{t+1}^{(f)}$ and memory cell $\mathbf{c}_{t+1}^{(f)}$) or reinitialize them. When the end frame of a video segment is estimated, we reset the LSTM parameters of the next frame to segment the video. Formally, the $t$-th binary segmentation gate is defined as a step function, which is computed as a non-linear combination of the feature of the $t+1$-th frame (i.e., $\mathbf{v}_{t+1}^{(f)}$) and the hidden state of the $t$-th frame from LSTM networks (i.e., $\mathbf{h}_t^{(f)}$), given by

$$\gamma_t = 1[\delta(\mathbf{w}_\gamma^T(\mathbf{W}_{\gamma v}\mathbf{v}_{t+1}^{(f)} + \mathbf{W}_{\gamma h}\mathbf{h}_t^{(f)} + \mathbf{b}_\gamma)) < \tau], \quad (2)$$

where the $1[\cdot]$ is a step function and $\delta(\cdot)$ is a sigmoid function. The $\mathbf{w}_\gamma^T$ is a learnable row vector, $\mathbf{W}_{\gamma v}$, $\mathbf{W}_{\gamma h}$ and $\mathbf{b}_\gamma$ are the learnable weights and bias. The $\tau$ is the threshold parameter of the step function. For example, the 3rd binary segmentation gate $\gamma_3$ is computed according to the $\mathbf{h}_3^{(f)}$ (i.e., the hidden state of the 3rd frame from LSTM networks) and $\mathbf{v}_4^{(f)}$ (i.e., the video feature of the 4-th frame). The binary gate function $\gamma_t$ then decides whether to transfer the parameters of the $t$-th frame from LSTM networks. That is, $\gamma_t$ represents the irrelevance between the semantic representations of the $t$-th frame and the video feature of the $t$+1-th frame. When $\gamma_t = 1$, we reinitialize the LSTM parameters of the $t$+1-th frame and the semantic representation of the $t$+1-th frame is computed by $\mathbf{h}_{t+1}^{(f)} = \text{LSTM}(\mathbf{v}_{t+1}^{(f)}, \mathbf{h}_0, \mathbf{c}_0)$, where $\mathbf{h}_0$ and $\mathbf{c}_0$ are the zero hidden state and memory cell. This means that we divide the adjacent two frames into two separate segments where the $t$-th frame is the end frame of the current segment and the $t$+1-th frame is the start frame of the next segment.

Given the frame representations $(\mathbf{h}_1^{(f)}, \mathbf{h}_2^{(f)}, \ldots, \mathbf{h}_N^{(f)})$ with binary gate values $(\gamma_1, \gamma_2, \ldots, \gamma_{N-1})$, we then learn the joint question-aware segment representation. If the value of binary gate $\gamma_t = 1$, the question-aware representation of current segment $S_k$ is computed and then passed to the segment-level highway LSTM networks. Given the last question representation $\mathbf{h}_R^{(q)}$ from the question LSTM networks, the frame-level dynamic attention score for the $t$-th frame $\mathbf{v}_t \in S_k$ is given by

$$\alpha_t^{(f)} = \mathbf{P}^{(f)} tanh(\mathbf{W}_h^{(f)} \mathbf{h}_t^{(f)} + \mathbf{W}_q^{(f)} \mathbf{h}_R^{(q)} + \mathbf{b}^{(f)}), \quad (3)$$

where $\mathbf{W}_h^{(f)}$, $\mathbf{W}_q^{(f)}$ are parameter matrices, $\mathbf{b}^{(f)}$ is the bias vector and the $\mathbf{P}^{(f)}$ is the parameter vector for computing attention scores. Here we apply the additive attention method for better performance even if the dot-product attention is faster [42]. For each frame $\mathbf{v}_t \in S_k$, its activation by the softmax function is given by $\beta_t^{(f)} = \frac{\exp(\alpha_t^{(f)})}{\sum_{\mathbf{v}_t \in S_k} \alpha_t^{(f)}}$. Thus, the attentional question-aware representation for segment $S_k$ is then given by $\mathbf{s}_k = \sum_{\mathbf{v}_t \in S_k} \beta_t^{(f)} \mathbf{h}_t^{(f)}$. By the frame-level dynamic LSTM network, we learn the question-aware sequential representation of video segments, denoted by $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$.

*2) Segment-Level Highway LSTM Networks:* The long-form video contents generally contain quite a lot of redundant information irrelevant to the given question, which interferes with the generation of correct answers. Inspired by highway networks [9], we develop the segment-level highway LSTM network with question-aware highway gate that filters the redundant information to capture the long-range video semantic dependencies and learn the segment-level semantic representations according to the given question.

Specifically, given segment representations $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$, we devise the segment-level highway LSTM networks to learn their semantic representations $(\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_K^{(s)})$. To enable the redundant information filtering, we define a question-aware highway gate to decide how to transfer the LSTM parameters (i.e., hidden state $\mathbf{h}_t^{(s)}$ and memory cell $\mathbf{c}_t^{(s)}$) of the $t$-th segment according to the hidden state $\mathbf{h}_t^{(s)}$ of the $t$-th segment and the last question representation $\mathbf{h}_R^{(q)}$. When the $t$-th segment is irrelevant to the given question, we ignore the LSTM parameters of the $t$-th segment and keep the LSTM parameters of the $t$-1-th segment to filter the redundant information. Formally, the $t$-th highway gate is defined as a non-linear combination of the semantic representation of the $t$-th segment from LSTM networks (i.e., $\mathbf{h}_t^{(s)}$) and the last question representation $\mathbf{h}_R^{(q)}$, given by

$$\mu_t = \delta(\mathbf{w}_\mu^T(\mathbf{W}_{\mu s}\mathbf{h}_t^{(s)} + \mathbf{W}_{\mu q}\mathbf{h}_R^{(q)} + \mathbf{b}_\mu)), \quad (4)$$

where $\delta(\cdot)$ is a sigmoid function and the $\mathbf{w}_\mu^T$ is a learnable row vector. The $\mathbf{W}_{\mu s}$, $\mathbf{W}_{\mu q}$ and $\mathbf{b}_\mu$ are the learnable weights and bias. The highway gate $\mu_t$ then upgrades the LSTM parameters transferred to next segment(i.e., the $t$+1-th segment) by

$$\begin{aligned} \mathbf{h}_t^{(s)} &\leftarrow \mathbf{h}_{t-1}^{(s)} \cdot (1 - \mu_t) + \mathbf{h}_t^{(s)} \cdot \mu_t, \\ \mathbf{c}_t^{(s)} &\leftarrow \mathbf{c}_{t-1}^{(s)} \cdot (1 - \mu_t) + \mathbf{c}_t^{(s)} \cdot \mu_t, \end{aligned} \quad (5)$$

that is, $\mu_t$ represents the relevance between the $t$-th segment and the given question. This means when $\mu_t = 0$, we filter the semantic representation of the $t$-th segment by $\mathbf{h}_t^{(s)} \leftarrow \mathbf{h}_{t-1}^{(s)}$ and $\mathbf{c}_t^{(s)} \leftarrow \mathbf{c}_{t-1}^{(s)}$. Therefore, by the segment-level highway LSTM network, we learn the segment-level question-aware semantic representations of videos, denoted by $\mathbf{h}^{(s)} = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_K^{(s)})$.

Unifying the frame-level dynamic LSTM network learning and segment-level highway LSTM network learning into a common framework, the dynamic hierarchical encoder network is given by $f(\mathbf{v}, \mathbf{q}) = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_K^{(s)})$.

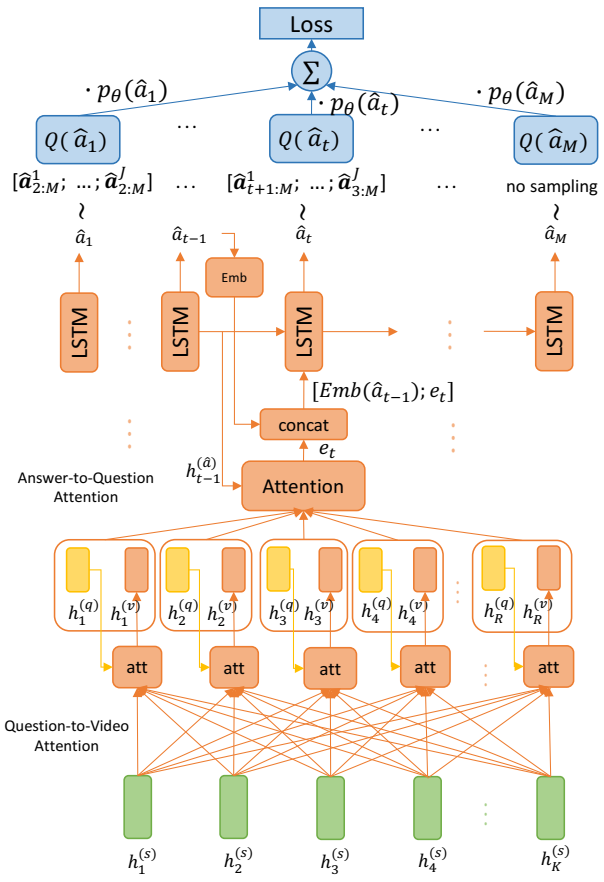



Fig. 3. The Framework of Reinforced Decoder Networks with Hierarchical Attention Mechanism for Open-Ended Answer Generation. The hierarchical attention mechanism, including the answer-to-question attention and question-to-video attention, is developed at each step of natural language generation to learn the context vector based on the segment-level semantic representations of video contents and the question semantic representation.

## C. Reinforced Decoder Networks

In this section, we propose the reinforced decoder network $g(\cdot)$ to generate high-quality natural language answer for open-ended long-form video question answering. Specifically, it's a normal LSTM decoder with hierarchical attention mechanism and trained with reinforcement learning.

*1) Hierarchical Attention Mechanism:* When answering a question, people tend to read the question first and then look for the video contents that match the keywords of the given question. Inspired by this, we advise the hierarchical attention mechanism at each step of natural language answer generation, which develops the hierarchical attention structure between

answer and question, and question and video, as shown in Figure 3.

Given the segment-level semantic representations of video contents $f(\mathbf{v},\mathbf{q}) = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_K^{(s)})$ from the dynamic hierarchical encoder network, the LSTM answer generator predicts the answer sequences word by word. At the $t$-th step, the $t$-th word is generated by sampling $\hat{a}_t \sim p_\theta(\hat{a}_t|\hat{\mathbf{a}}_{1:t-1}, \mathbf{v}, \mathbf{q}) = g_t(\mathbf{h}_{t-1}^{(\hat{a})}, [\mathrm{Emb}(\hat{a}_{t-1}); \mathbf{e}_t])$, where $g_t(\cdot)$ is the LSTM answer generator, $\mathbf{h}_{t-1}^{(\hat{a})}$ is the decoder state at step $t$-1 and $\mathrm{Emb}(\hat{a}_{t-1})$ is the word embedding of the $t$-1-th generated word in answer. The $\mathbf{e}_t$ is the context vector at step $t$ and we next illustrate how the $\mathbf{e}_t$ is modeled.

Given question representations $\mathbf{h}^{(q)} = (\mathbf{h}_1^{(q)}, \ldots, \mathbf{h}_R^{(q)})$, we propose the hierarchical attention to obtain the context vector $\mathbf{e}_t$, which consists of the answer-to-question attention and question-to-video attention. Concretely, we first compute the answer-to-question attention weight between the $r$-th word token and the $t$-1-th decoder state given by

$$\alpha_r^{(\hat{a})} = \mathbf{P}^{(\hat{a})} tanh(\mathbf{W}_h^{(\hat{a})} \mathbf{h}_{t-1}^{(\hat{a})} + \mathbf{W}_q^{(\hat{a})} \mathbf{h}_r^{(q)} + \mathbf{b}^{(\hat{a})}). \quad (6)$$

We then activate it by the softmax function $\beta_r^{(\hat{a})} = \frac{\exp(\alpha_r^{(\hat{a})})}{\sum_r \exp(\alpha_r^{(\hat{a})})}$. Thus, the context vector at step $t$ is given by $\mathbf{e}_t = \sum_r \beta_r^{(\hat{a})} \mathbf{h}_r^{(v)}$, where $\mathbf{h}_r^{(v)}$ is computed by question-to-video attention as below. To model each $\mathbf{h}_r^{(v)}$, we take the segment-level video representations of videos $\mathbf{h}^{(s)} = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_K^{(s)})$ and develop question-to-video attention to learn word-aware video representations, given by

$$\begin{aligned} \alpha_{rk}^{(q)} &= \mathbf{P}^{(q)} tanh(\mathbf{W}_h^{(q)} \mathbf{h}_r^{(q)} + \mathbf{W}_h^{(q)} \mathbf{h}_k^{(s)} + \mathbf{b}^{(q)}), \\ \beta_{rk}^{(q)} &= \frac{\exp(\alpha_{rk}^{(q)})}{\sum_k \exp(\alpha_{rk}^{(q)})}, \ \mathbf{h}_r^{(v)} = \sum_k \beta_{rk}^{(q)} \mathbf{h}_k^{(s)}. \end{aligned} \quad (7)$$

By the LSTM decoder network with hierarchical attention mechanism, we finally generate the natural language answer $\hat{\mathbf{a}} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_M)$.

*2) Reinforcement Learning For Decoder Networks:* One common approach to train the proposed decoder network is the framework of maximum likelihood estimation, given by

$$\mathcal{L}_{ML} = \sum_{t=1}^{M} \log p_\theta(a_t|\mathbf{a}_{1:t-1}, \mathbf{v}, \mathbf{q}). \quad (8)$$

However, the training based on maximum likelihood estimation makes the learnt decoder network suboptimal [43]. In this work, we train the proposed decoder network under the framework of reinforcement learning.

In the setting of reinforcement learning, we define the generation of next answer word as action, and the decoding probability $p_\theta(a_t|\mathbf{a}_{1:t-1}, \mathbf{v}, \mathbf{q})$ as the policy. Following the existing visual-semantic embedding work [44], we choose the reward function based on the embedding similarity between the ground-truth answer $\mathbf{a}$ and the generated answer $\hat{\mathbf{a}}$, given by $R_{\mathbf{a}}(\hat{\mathbf{a}}) = \| \frac{1}{M_a} \sum_t \mathrm{Emb}(a_t) - \frac{1}{M_{\hat{a}}} \sum_t \mathrm{Emb}(\hat{a}_t) \|^2$. Specifically, we define the expected cumulative reward at each decoding step using value function by $Q(\hat{a}_t|\hat{\mathbf{a}}_{1:t-1}, \mathbf{v}, \mathbf{q}) = E_{p_\theta(\hat{\mathbf{a}}_{t+1:M}|\hat{\mathbf{a}}_{1:t}, \mathbf{v}, \mathbf{q})} R_{\mathbf{a}}(\hat{\mathbf{a}})$. The value function $Q(\hat{a}_t|\hat{\mathbf{a}}_{1:t-1}, \mathbf{v}, \mathbf{q})$ is then estimated by aggregating the Monte-Carlo simulation at each decoding step, given by

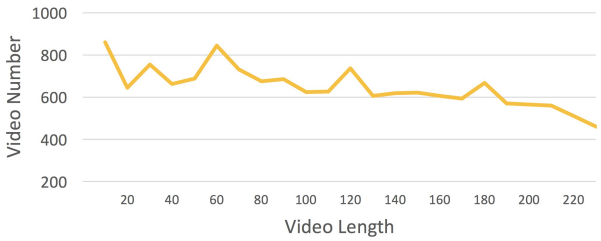



Fig. 4. Length Statistics of Videos for ActivityNet Dataset

$$\begin{aligned} & Q(\hat{a}_t|\hat{\mathbf{a}}_{1:t-1}, \mathbf{v}, \mathbf{q}) \\ \approx & \begin{cases} \frac{1}{J} \sum_{n=1}^{J} R_{\mathbf{a}}([\hat{\mathbf{a}}_{1:t}, \hat{\mathbf{a}}_{t+1:M}^{(n)}]), & t < M \\ R_{\mathbf{a}}([\hat{\mathbf{a}}_{1:t-1}, \hat{a}_t]). & t = M \end{cases} \end{aligned} \quad (9)$$

The $\{\hat{\mathbf{a}}_{t+1:M}^{(1)}, \hat{\mathbf{a}}_{t+1:M}^{(2)}, \ldots, \hat{\mathbf{a}}_{t+1:M}^{(J)}\}$ is the set of generated answers, which are randomly sampled starting from the $t$+1-th decoding step using current state and action. The gradients $\nabla_\theta \mathcal{L}_{RL}$ of the reinforced decoder network according to the policy gradient theorem is given by

$$\sum_{t=1}^{M} \nabla_\theta \log p_\theta(a_t|\mathbf{a}_{1:t-1}, \mathbf{v}, \mathbf{q}) Q(\hat{a}_t|\hat{\mathbf{a}}_{1:t-1}, \mathbf{v}, \mathbf{q}). \quad (10)$$

Thus, we can apply the reinforcement learning framework to train the dynamic hierarchical reinforced networks. To improve the training efficiency, we first adopt the maximum likelihood loss to preliminarily train networks, and then develop the reinforcement training.

## IV. Experiments

In this section, we first introduce two video question answering datasets, one is long-form and another is short-form as a comparison, and then conduct several experiments on them to show the effectiveness of our method DHRN for open-ended long-form video question answering. In Section IV-A, we introduce the two datasets and the implementation details of our proposed method. In Section IV-B, we introduce the evaluation criteria for open-ended answers. In Section IV-C, we compare the proposed approach with state-of-the-art methods. In Section IV-D, we discuss the proposed framework in detail.

### A. Experimental Settings

*1) Dataset:* We construct the long-form video question answering dataset from the ActivityNet data [45] with natural language descriptions, which contains 20,000 videos amounting to 849 hours and 100,000 descriptions. The average time duration of videos is over 100 seconds as Figure 4 and the longest video runs for over 10 minutes. In our previous work, we generate the question-answer pairs from the video descriptions by a rule-based question generation method [17]. Following the existing visual question answering approaches [2], [13], [21], we define five types of questions related to the object, number, color, location and action. For the first four types, we generate questions in the same way

TABLE I
SUMMARIES OF THE ACTIVITYNET DATASET

| Data | Question Types | | | | | |
|---|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action | All |
| Train | 10,338 | 2,605 | 3,680 | 7,438 | 28,543 | 52,604 |
| Valid | 1,327 | 289 | 447 | 1,082 | 3,692 | 6,837 |
| Test | 1,296 | 355 | 501 | 971 | 3,826 | 6,949 |
| All | 12,961 | 3,249 | 4,628 | 9,491 | 36,061 | 66,390 |

TABLE II
QUESTION AND ANSWER LENGTHS OF THE ACTIVITYNET DATASET

| Data | Question Types | | | | |
|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action |
| Question | 11.64 | 14.38 | 8.00 | 11.22 | 7.99 |
| Answer | 2.40 | 1.00 | 1.00 | 3.34 | 5.09 |

TABLE III
SUMMARIES OF THE TGIF DATASET

| Data | Question Types | | | | | |
|---|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action | All |
| Train | 13,715 | 7,024 | 9,723 | 2,249 | 40,526 | 73,237 |
| Valid | 1,771 | 866 | 1,262 | 298 | 2,003 | 6,200 |
| Test | 1,743 | 834 | 1,259 | 261 | 1,920 | 6,017 |
| All | 17,229 | 8,724 | 12,244 | 2,808 | 44,449 | 85,454 |

TABLE IV
QUESTION AND ANSWER LENGTHS OF THE TGIF DATASET

| Data | Question Types | | | | |
|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action |
| Question | 9.10 | 10.40 | 7.00 | 9.17 | 8.68 |
| Answer | 1.00 | 1.00 | 1.00 | 1.00 | 2.41 |

as [17]. As for the action type, we likewise search the verb phrase in video descriptions as the ground-truth answers and transforming the sentence structure to the action questions. Different from the previous work, we further reconstruct the long-form video question answering dataset and generate more question-answer pairs, especially for the action-related questions. We then take 80% of constructed data as the training data, 10% as the validation data and 10% as the testing ones. The five types of long-form video question-answering pairs used for the experiments are summarized in Table I. Moreover, we count the average question and answer length of all type questions in Table II, where questions about number and color only correspond to one-word answers, i.e. the numeral and color-related word, and action questions require the longest answers.

To validate that our method DHRN achieves the better performance improvement on long videos than short ones, we conduct several experiments on the public short-form TGIF dataset, which contains 102066 gifs and corresponding question-answer pairs. The time duration of gifs is a few seconds and the average frame number is about 40. The origin TGIF dataset has four types of questions about state transition, action repetition, count repetition and frameQA. We take the first two type about the video dynamic information as the action type. The third type about how many times the action occurs is not suitable for open-ended prediction. The frameQA questions in the origin TGIF dataset have four types related to object, number, color and location, and we do not make any change on this part. Moreover, the answer form of action type in the origin TGIF dataset is multi-choice, which provides several candidate answers to be chosen. We select the correct answer as the ground-truth open-ended answer. The five types of question-answering pairs of the TGIF dataset are summarized in Table III. Likewise, we count the average question and answer lengths in Table IV, where only action questions require multiple-word answers.

*2) Data Preprocessing:* For the ActivityNet dataset, we first define consecutive 16 frames as a unit and each unit overlaps 8 frames with adjacent units. We then resize frames in units to $112 \times 112$, and extract a 4,096-d visual representation for each unit by the pre-trained 3D-ConvNet [46]. We then reduce the dimensionality from 4096 to 500 dimensions using PCA and take the final results as the frame-level video features, i.e. $\mathbf{v}^{(f)}$. We next employ the pre-trained word2vec model [47] to extract the semantic representations of questions and answers. Specifically, the size of the vocabulary set is 9,028 and the dimension of word vector is set to 300. Particularly, we add four tokens <start>, <eos>, <unk> and <pad> to denote the start of the answer, the end of the answer, the out-of-vocabulary word and the padding word, respectively.

For the TGIF dataset, the average frame number is about 40 and we employed the pre-trained VGGNet [48] to extract the 4,096-d feature vector for each frame. Since the clips in TGIF are short, we do not define the frame unit and directly take the frame features as input. Other processes are similar to the ActivityNet dataset and the size of the vocabulary set is 8,467.

*3) Implementation Details:* In the training process, we employ the Adam optimizer [49] to minimize the loss for all methods, where the initial learning rate is set to 0.0001. To prevent the gradient is too large in the back propagation, a gradient clipping method is utilized to limit gradient norms within 5.0. We adopt the mini-batch strategy 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 during last five epochs.

### B. Performance Criteria

We evaluate the performance of our proposed DHRN method based on two widely-used evaluation criteria for open-ended visual question answering, i.e., Accuracy [13] and WUPS [18]. Given the testing question $\mathbf{q} \in Q$ with its corresponding ground-truth answer $\mathbf{a} = \{a_1, a_2, \ldots, a_{M_a}\}$, we denote the generated answers $\mathbf{o}$ from our DHRN method by $\mathbf{o} = \{o_1, o_2, \ldots, o_{M_o}\}$. We next introduce the evaluation criteria below.

- **Accuracy.** The Accuracy is the normalized criterion of accessing the quality of the generated answer based on the entire testing question set $Q$, given by

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

where $L$ represents the minimum of $M_a$ and $M_o$, and $\mathbf{1}[\cdot]$ is an indicator function. The accuracy score of

TABLE V
EXPERIMENTAL RESULTS ON ACCURACY, WUPS@0.0 AND WUPS@0.9 WITH ALL TYPES OF VISUAL QUESTIONS ON THE ACTIVITYNET DATASET.

| Method | Accuracy | WUPS@0.0 | WUPS@0.9 |
|---|---|---|---|
| Q-only | 0.0916 | 0.1736 | 0.4689 |
| VQA+ | 0.1687 | 0.2454 | 0.5327 |
| MN+ | 0.1992 | 0.2824 | 0.5631 |
| STAN | 0.2349 | 0.3154 | 0.5789 |
| AMU | 0.2438 | 0.3338 | 0.5863 |
| STVQA | 0.2469 | 0.3345 | 0.5878 |
| DHRN | **0.2874** | **0.369** | **0.6112** |

TABLE VI
EXPERIMENTAL RESULTS ON ACCURACY WITH DIFFERENT TYPES OF VISUAL QUESTIONS ON THE ACTIVITYNET DATASET.

| Method | Accuracy | | | | |
|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action |
| Q-only | 0.1312 | 0.2354 | 0.1491 | 0.1685 | 0.0378 |
| VQA+ | 0.2171 | 0.5972 | 0.2138 | 0.2678 | 0.0815 |
| MN+ | 0.2593 | 0.6507 | 0.2355 | 0.3126 | 0.1034 |
| STAN | 0.2848 | 0.7543 | 0.2394 | 0.3426 | 0.1443 |
| AMU | 0.3221 | 0.7465 | 0.2675 | 0.3229 | 0.1476 |
| STVQA | 0.297 | 0.7854 | 0.2684 | 0.3562 | 0.1522 |
| DHRN | **0.334** | **0.7859** | **0.2974** | **0.3655** | **0.2042** |

corresponding words is 1 only if $a_i$ and $o_i$ are identical, and 0 otherwise.

- **WUPS.** The WUPS is the soft measure based on the WUP [50] score to evaluate the quality of the generated answer. The WUP measures word similarity based on WordNet [51]. Thus, Given the entire testing $Q$, the WUPS score with the threshold $\gamma$ is measured by

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

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

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

Following the experimental setting in [18], we choose two WUPS evaluation criteria with the parameter $\gamma$ to be 0 and 0.9, denoted by WUPS@0.0 and WUPS@0.9, respectively.

### C. Performance Comparisons

*1) Baseline Methods:* We compare our proposed method with other state-of-the-art methods for the problem of open-ended video question answering as follows:

- **Q-only** is a language-only baseline [43], which only utilize the question to generate answers by an encoder-decoder framework with an attention mechanism.
- **VQA+** method is the extension of a classical image question algorithm [18], where the mean-pooling layer is added to learn the video global representation.
- **MN+** method [1] is the extension of the end-to-end memory network algorithm, where the one-layer bi-LSTM network is added to encode the sequence of video frames for answer generation.

TABLE VII
EXPERIMENTAL RESULTS ON WUPS@0.0 WITH DIFFERENT TYPES OF VISUAL QUESTIONS ON THE ACTIVITYNET DATASET.

| Method | WUPS@0.0 | | | | |
|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action |
| Q-only | 0.1769 | 0.6373 | 0.3821 | 0.1730 | 0.1023 |
| VQA+ | 0.2562 | 0.7067 | 0.4315 | 0.2596 | 0.171 |
| MN+ | 0.3042 | 0.7636 | 0.4679 | 0.3105 | 0.1989 |
| STAN | 0.3456 | 0.8109 | 0.4599 | 0.362 | 0.2346 |
| AMU | 0.4178 | 0.8512 | 0.4807 | 0.3449 | 0.2353 |
| STVQA | 0.3906 | 0.8426 | 0.4748 | 0.378 | 0.2399 |
| DHRN | **0.4358** | **0.8872** | **0.4811** | **0.3816** | **0.2807** |

TABLE VIII
EXPERIMENTAL RESULTS ON WUPS@0.9 WITH DIFFERENT TYPES OF VISUAL QUESTIONS ON THE ACTIVITYNET DATASET.

| Method | WUPS@0.9 | | | | |
|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action |
| Q-only | 0.4375 | 0.9012 | 0.8069 | 0.4272 | 0.4057 |
| VQA+ | 0.5032 | 0.9111 | 0.8485 | 0.5054 | 0.4732 |
| MN+ | 0.5299 | 0.9355 | 0.8614 | 0.5379 | 0.5071 |
| STAN | 0.5732 | 0.956 | 0.8677 | 0.5521 | 0.5148 |
| AMU | 0.6072 | 0.9493 | 0.8742 | 0.527 | 0.5229 |
| STVQA | 0.5851 | **0.9631** | 0.8727 | 0.5568 | 0.5245 |
| DHRN | **0.6128** | 0.9539 | **0.8771** | **0.5727** | **0.5538** |

- **STAN** method [2] is based on the hierarchical spatio-temporal attention network for learning the joint representation of the dynamic video contents according to the given question.
- **AMU** method [3] is the attentional reasoning network, which is based on the attention memory unit for generating the natural language answer.
- **STVQA** method [4] is based on the spatio-temporal reasoning algorithm, which employs the spatial and temporal attention on video to answer questions.

Since some baselines are designed for the multiple-choice form, we add a conventional LSTM recurrent answer generator to the end of models for open-ended answer generation.

*2) Performance Comparisons on the ActivityNet Dataset:* Table V shows the overall experimental results of the methods on all types of questions based on three evaluation criteria on the ActivityNet dataset. Tables VI, VII and VIII illustrate the evaluation results on Accuracy, WUPS@0.0 and WUPS@0.9 with different types of questions on the ActivityNet dataset, respectively. The hyper-parameters and parameters which achieve the best performance on the validation set are chosen to conduct the testing evaluation. The experiments reveal a number of interesting points:

- the language-only method Q-only achieves the worst performance, which shows that question answering in the ActivityNet dataset depends on the sufficient understanding of video contents.
- The LSTM based methods MN+, STAN, AMU, STVQA, DHRN outperform the mean-pooling based method VQA+, which suggests that the sequential frame-level representations are critical for the problem.
- The attention based methods STAN, AMU, STVQA and DHRN achieve better performance than other baselines. This demonstrates that the joint representation learning

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2922062, IEEE Transactions on Image Processing



TABLE IX
EXPERIMENTAL RESULTS ON ACCURACY, WUPS@0.0 AND WUPS@0.9 WITH ALL TYPES OF VISUAL QUESTIONS ON THE TGIF DATASET.

| Method | Accuracy | WUPS@0.0 | WUPS@0.9 |
|---|---|---|---|
| Q-only | 0.1401 | 0.3055 | 0.6894 |
| VQA+ | 0.2652 | 0.3858 | 0.7406 |
| MN+ | 0.3078 | 0.4323 | 0.7862 |
| STAN | 0.3556 | 0.4681 | 0.7944 |
| AMU | 0.3595 | 0.4762 | 0.7934 |
| STVQA | 0.363 | 0.4757 | 0.7971 |
| DHRN | **0.3837** | **0.4902** | **0.8079** |

TABLE X
EXPERIMENTAL RESULTS ON ACCURACY WITH DIFFERENT TYPES OF VISUAL QUESTIONS ON THE TGIF DATASET.

| Method | Accuracy | | | | |
|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action |
| Q-only | 0.0827 | 0.2865 | 0.1643 | 0.1032 | 0.1176 |
| VQA+ | 0.1222 | 0.7518 | 0.2463 | 0.1824 | 0.2073 |
| MN+ | 0.1388 | 0.7866 | 0.2621 | 0.2296 | 0.2807 |
| STAN | 0.222 | 0.7974 | 0.3461 | 0.2337 | 0.3078 |
| AMU | 0.2364 | 0.7985 | 0.3463 | 0.2375 | 0.3057 |
| STVQA | 0.2559 | **0.7986** | 0.3511 | 0.249 | 0.2943 |
| DHRN | **0.2716** | 0.7974 | **0.3738** | **0.2718** | **0.3276** |

of video and question can also improve the performance of open-ended video question answering.

- Overall, our DHRN method achieves the best performance on the three evaluations. From the viewpoint of different types of questions, our DHRN method achieves higher performance improvement on action-related questions than other methods. These facts show that our DHRN method can learn the long-range semantic information from long-form video contents effectively and improve the performance of long-form video question answering.

*3) Performance Comparisons on the TGIF Dataset:* Table IX shows the overall experimental results of the methods on all types of questions based on three evaluation criteria on the TGIF dataset. Tables X, XI and XII illustrate the evaluation results on Accuracy, WUPS@0.0 and WUPS@0.9 with different types of questions on the TGIF dataset, respectively. Similarly, the best hyper-parameters and parameters are chosen to conduct the testing evaluation. There are a number of interesting points in the experiment results:

- Overall, all methods achieve better performance on the TGIF dataset than on the Activity dataset. This suggests that the short-form video question answering is easier than the long-form task.
- Similar to the results on the ActivityNet dataset, the LSTM based methods MN+, STAN, AMU, STVQA, DHRN achieve better performance than the mean-pooling based method VQA+, and the attention based methods STAN, AMU, STVQA and DHRN outperform other baselines. This demonstrates that the sequence modeling and the joint representation learning of video and question are also critical for short-form video question answering.
- Our DHRN method still achieves the best performance on the three evaluations. However, the performance is only slightly improved than other baselines. This fact shows our DHRN method is also effective for short-form video question answering but more suitable for long-form videos.

TABLE XI
EXPERIMENTAL RESULTS ON WUPS@0.0 WITH DIFFERENT TYPES OF VISUAL QUESTIONS ON THE TGIF DATASET.

| Method | WUPS@0.0 | | | | |
|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action |
| Q-only | 0.1182 | 0.8323 | 0.5290 | 0.1527 | 0.1208 |
| VQA+ | 0.1797 | 0.8756 | 0.5831 | 0.2355 | 0.2512 |
| MN+ | 0.2086 | 0.9074 | 0.6058 | 0.3697 | 0.3236 |
| STAN | 0.2937 | 0.9172 | 0.6129 | 0.3665 | 0.3502 |
| AMU | 0.3136 | 0.9182 | 0.6132 | 0.373 | 0.3561 |
| STVQA | 0.3269 | **0.9183** | 0.6141 | 0.3805 | 0.3406 |
| DHRN | **0.3447** | 0.9172 | **0.6232** | **0.3981** | **0.3622** |

TABLE XII
EXPERIMENTAL RESULTS ON WUPS@0.9 WITH DIFFERENT TYPES OF VISUAL QUESTIONS ON THE TGIF DATASET.

| Method | WUPS@0.9 | | | | |
|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action |
| Q-only | 0.5743 | 0.9504 | 0.8732 | 0.5692 | 0.5764 |
| VQA+ | 0.6424 | 0.9704 | 0.8916 | 0.6257 | 0.6464 |
| MN+ | 0.7043 | 0.9747 | 0.8938 | 0.7803 | 0.709 |
| STAN | 0.7128 | 0.9761 | 0.8914 | 0.7824 | 0.7276 |
| AMU | 0.7203 | 0.976 | 0.8916 | 0.7814 | 0.7178 |
| STVQA | 0.7272 | **0.9762** | 0.8964 | 0.7851 | 0.7193 |
| DHRN | **0.7394** | 0.9761 | **0.9056** | **0.7969** | **0.7343** |

All in all, the dynamic hierarchical reinforced encoder-decoder networks, composed of the dynamic hierarchical encoder and reinforced decoder, are effective for video question answering, especially for long-form video question answering.

TABLE XIII
EXPERIMENTAL RESULTS ON KEYRECALL WITH DIFFERENT TYPES OF VISUAL QUESTIONS ON THE ACTIVITYNET DATASET.

| Method | KeyRecall | | | | |
|---|---|---|---|---|---|
| | Object | Number | Color | Location | Action |
| AMU | 0.3321 | 0.7531 | 0.2740 | 0.3314 | 0.1521 |
| STVQA | 0.3075 | 0.7851 | 0.2691 | 0.3589 | 0.1643 |
| DHRN | **0.3443** | **0.7856** | **0.2968** | **0.3749** | **0.2175** |

*4) Performance Comparisons About Key Words:* The reasonable answers of given questions often have key words, which decide whether the answer is acceptable. For example, verbs in answers are crucial for questions about actions, numeral is critical for number questions, nouns decide the answers of object questions, color questions require correct color prediction, and location questions need both prepositions and nouns. But Accuracy and WUPS give the same weight to all words, ignoring the key ones. Here we adopt the KeyRecall as another criterion, which means the recall rate of key words in the generated answers. But for one-word answers, KeyRecall is equivalent to Accuracy, so it is meaningless for the TGIF dataset. To verify the ability of our DHRN, we compare it with two excellent baselines AMU and STVQA on the ActivityNet dataset, shown in Table XIII.

We can find that our DHRN method still achieves the best performance, demonstrating the effective modeling of key information. Moreover, compared with the Accuracy performance, the KeyRecall is relatively higher without the aligning

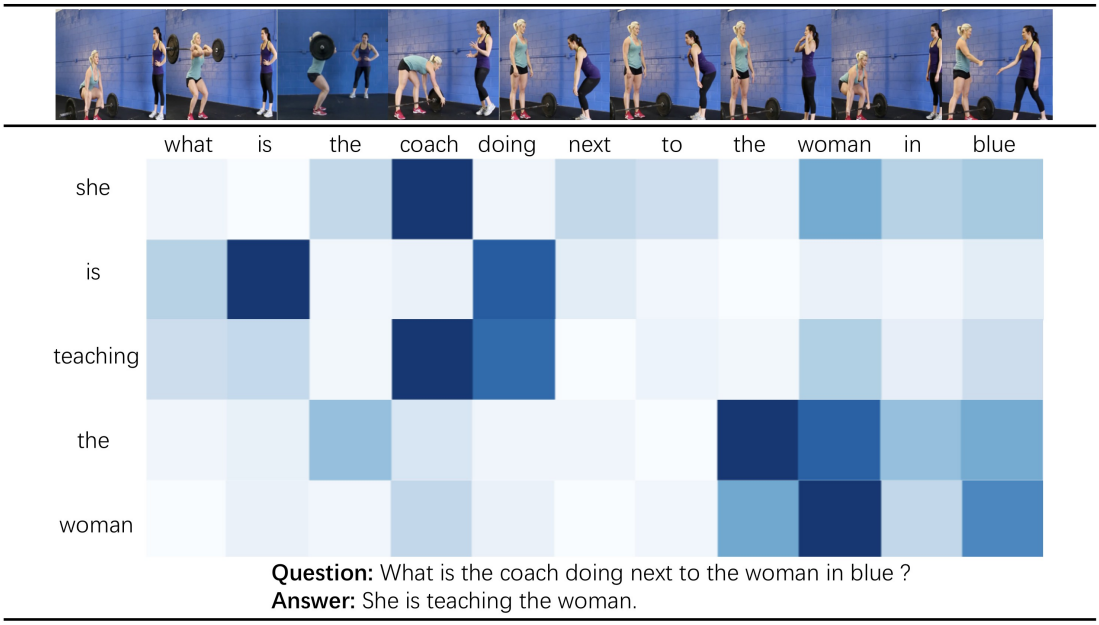



Fig. 5. The Answer-to-Question Attention Results of Hierarchical Attention Mechanism in Reinforced Decoder Networks

TABLE XIV
MULTIPLE-CHOICE RESULTS ON THE MOVIEQA DATASET.

| Method | SSCB | MemN2N | LMN | DHRN |
|---|---|---|---|---|
| Accuracy | 21.6 | 23.1 | 38.3 | **38.7** |

limitation of word pairs, i.e., the $i$-th word in the generated answer corresponds to the $i$-th word in the ground truth.

*5) Performance Comparisons on Multiple-Choice:* Compared with open-ended video question answering, multiple-choice video question answer provides a clearer evaluation to select a correct answer from candidates. To further verify the performance of our DHRN method, we conduct the multiple-choice experiments on a long-form movie question answering dataset MovieQA [31]. The original dataset not only offers movie clips, related question and corresponding answers, but also give the plots, subtitles, scripts, and DVS of the movies. Since our method is designed for pure video question answering, we only utilize the clip contents to choose the candidate answers. To guarantee fairness, all baselines [31], [37] also only use the clip contents as clues. Concretely, we replace the reinforced decoder with a simple multiple selector. We first learn each answer representation $\mathbf{h}_i^{(u)}$ by a LSTM network, then use the last question representation $\mathbf{h}_R^{(q)}$ to attend the segment-level video representation $\mathbf{h}^{(s)}$ and obtain global video representation $\mathbf{h}^{(g)}$, and finally compute the scores for all candidates, given by $score_i = (\mathbf{h}_R^{(q)} + \mathbf{h}^{(g)})^\top \mathbf{h}_i^{(u)}$. Following previous approaches [31], we evaluate our DHRN method on the $val$ set of the MovieQA dataset, shown in Table XIV. From the evaluation results, we note that our DHRN method achieves better performance than other baselines while only considering the clip contents and ignoring subtitles, scripts, and so on.

TABLE XV
EXPERIMENTAL RESULTS OF ABLATION STUDY FOR THE DYNAMIC HIERARCHICAL REINFORCED NETWORKS ON THE ACTIVITYNET DATASET.

| Configuration | Accuracy | WUPS@0.9 | WUPS@0.0 |
|---|---|---|---|
| DHRN(w/o. segment) | 0.2653 | 0.3492 | 0.5963 |
| DHRN(w/o. highway) | 0.2788 | 0.3609 | 0.606 |
| DHRN(w/o. hier) | 0.2765 | 0.3583 | 0.6049 |
| DHRN(w/o. rl) | 0.2683 | 0.3514 | 0.5977 |
| DHRN(SCST) | 0.2854 | **0.3693** | 0.6074 |
| DHRN(full) | **0.2874** | 0.369 | **0.6112** |

## D. In-Depth Analysis of the Proposed Framework

In this section, we discuss the proposed framework in details. The ablation study shows the effectiveness of each component of our DHRN method. The qualitative analysis visually presents the experimental results. And the hyper-parameter analysis demonstrates the influence of important hyper-parameters on the performance of our DHRN method.

*1) Ablation Study:* To validate the effectiveness of each component, we further conduct ablation studies on the frame-level dynamic LSTM networks, segment-level highway LSTM networks and reinforced decoder networks. Concretely, we discard or replace one component at a time to generate an ablation model as follows.

- **DHRN(w/o. segment)**: We first remove the binary segmentation gate from the frame-level dynamic LSTM network and take frame semantic representations as the input of segment-level LSTM.
- **DHRN(w/o. highway)**: We then discard the question-aware highway gate from the segment-level highway network.
- **DHRN(w/o. hier)**: We next replace the hierarchical attention mechanism with a normal attention strategy from the reinforced decoder network.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2922062, IEEE Transactions on Image Processing



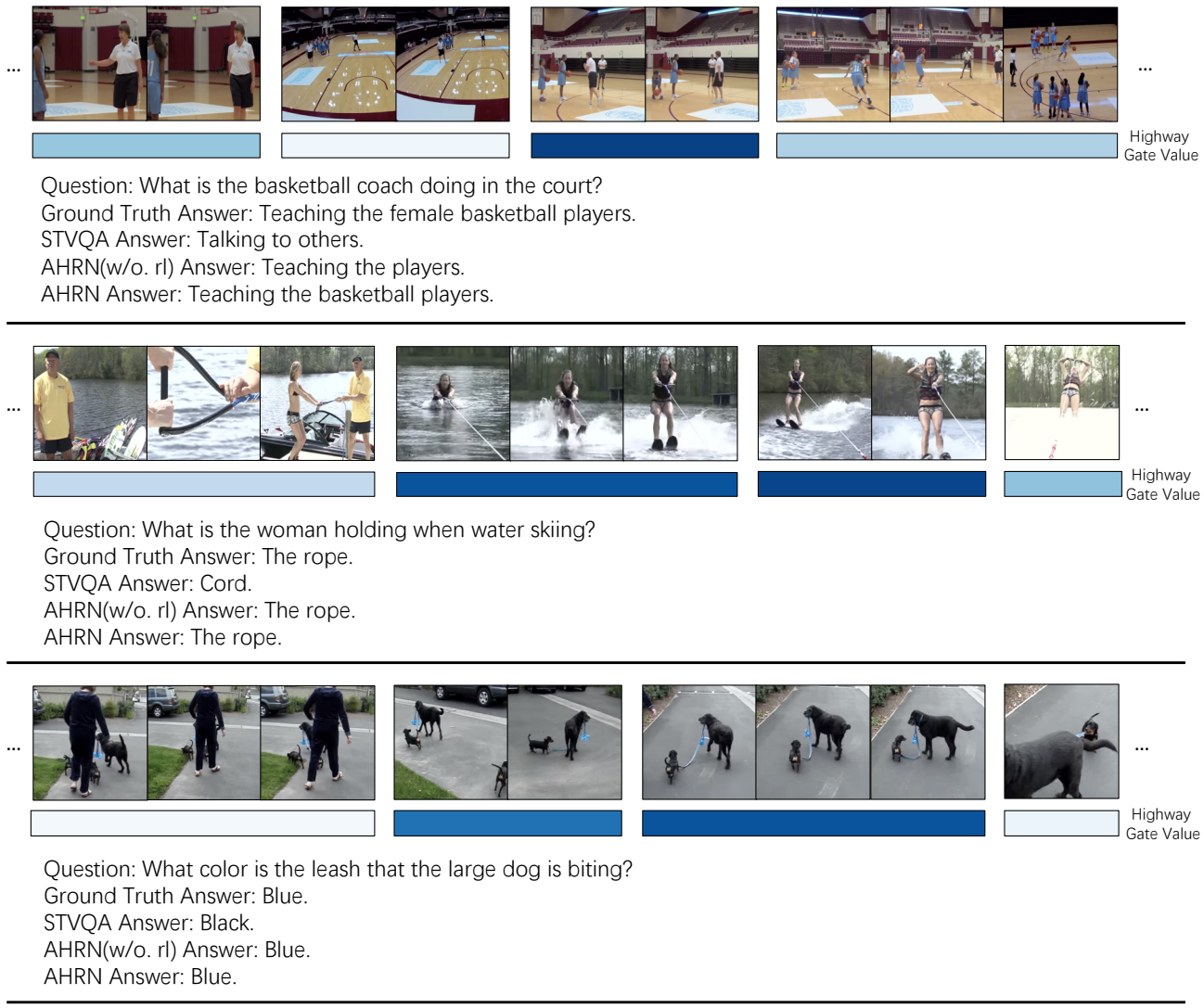



Fig. 6. Examples of Open-Ended Long-Form Video Question Answering on the ActivityNet Dataset

- **DHRN(w/o. rl)**: To evaluate the benefits of the reinforcement learning framework, we remove the reinforcement training from the decoding process.
- **DHRN(SCST)**: We further replace the proposed reinforcement training with another reinforcement algorithm SCST [52].

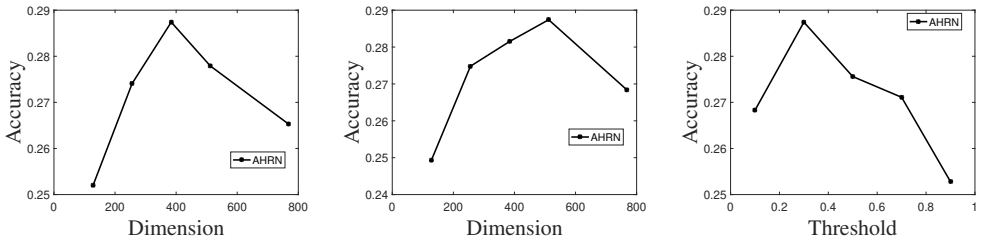
The ablation results on the ActivityNet dataset are shown in Table XV. According to the ablation results, we find several interesting points:

- The DHRN(full) outperforms all ablation models on the ActivityNet dataset, which demonstrates each component of our DHRN method is helpful for open-ended long-form video question answering.
- The DHRN(w/o. segment) and DHRN(w/o. rl) achieve the worse performance on other baselines, indicating the binary segmentation gate and reinforcement learning framework play relatively important roles in our DHRN method. The binary segmentation gate can effectively segment the successive video stream and the reinforcement learning framework is beneficial for high-quality answer generation
- Compared with SCST algorithm, our DHRN(full) method achieves a slight boost. This fact demonstrates our reinforcement learning method is suitable for this task. But the SCST has a faster convergence speed in the practical training. Concretely, after the maximum likelihood training, our reinforcement method requires 12 epochs to fine-tune the model but the SCST algorithm only need 8 epochs, where one epoch takes about 5 minutes.

*2) Qualitative Analysis:* To demonstrate how the hierarchical attention mechanism works, we display the answer-to-question attention results in Figure 5, where each word generated from the decoder LSTM networks depends on the relevant words in the question. At each step of answer generation, the decoder network learns the context vector from segment-level semantic representations based on the representation of relevant words in the question. As we can see in Figure 5, the answer-to-question attention results are visualized using thermodynamic diagram and the darker color represents the

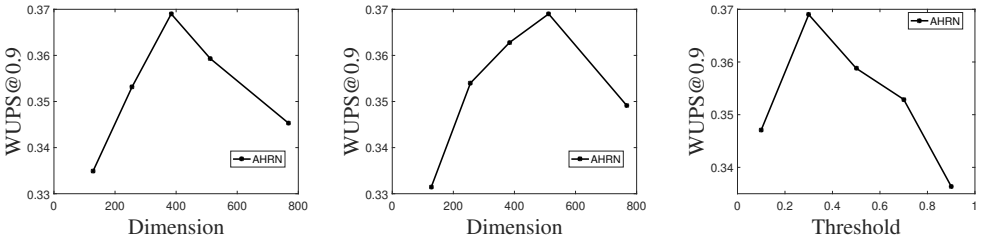
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2922062, IEEE Transactions on Image Processing






(a) Frame-Level LSTM State (b) Segment-level LSTM State (c) Gate Threshold

Fig. 7. Effect of frame-level LSTM hidden state dimension, segment-level LSTM hidden state dimension and binary segmentation gate threshold using Accuracy on the ActivityNet dataset.




(a) Frame-Level LSTM State (b) Segment-Level LSTM State (c) Gate Threshold

Fig. 8. Effect of frame-level LSTM hidden state dimension, segment-level LSTM hidden state dimension and binary segmentation gate threshold using WUPS@0.9 on the ActivityNet dataset.

higher correlation between the word semantic representations and the current state of decoder LSTM networks. For example, the word "teaching" in the answer is more relevant to word "coach" and "doing" in the question. This demonstrates the hierarchical attention mechanism is helpful for high-quality answer generation.

Furthermore, we display some typical examples of long-form video question answering from the ActivityNet dataset in Figure 6. According to long-form video contents and given questions, we show the generated answers of our DHRN method and STVQA method which has the best performance in all baselines. And to qualitatively explore the performance of the reinforced decoder, we also generate answers without the reinforcement training. Moreover, our frame-level dynamic LSTM segments the frame sequence into variable-length units and segment-level highway LSTM gives a gate value to decide whether a segment should be filtered out. In these examples, we display the segmentation of video sequences and the highway gate values, where the darker color represents the larger gate value and the higher relevance between this segment and the given question. Compared with the ground-truth answer, we intuitively observe that our DHRN method achieves better performance than the STVQA method and the reinforced decoder has a further improvement than DHRN(no_rl). We note that our frame-level segmentation can effectively aggregate the frames with similar contents and gate values are able to reflect the correlation.

*3) Hyper-Parameter Analysis:* In our DHRN approach, there are three essential hyper-parameters, which are the dimension of hidden state in frame-level LSTM networks, the dimension of hidden state in segment-level LSTM networks and the threshold $\tau$ of binary segmentation gates. We investigate the effect of these hyper-parameters on our method by varying both the dimension of hidden state in frame-level LSTM networks and segment-level LSTM networks from 128 to 768, and the threshold $\tau$ of binary segmentation gates from 0.1 to 0.9 on Accuracy in Figures 7(a), 7(b) and 7(c). We then vary these hyper-parameters to show their effect on our method using WUPS@0.9 in Figures 8(a), 8(b) and 8(c). While a hyper-parameter is varied, the other hyper-parameters are set to their best value.

From these figures, we can see that our DHRN method achieves the best performance when the dimension of hidden state in frame-level LSTM networks is set to 384, the dimension of hidden state in segment-level LSTM networks is set to 512 and the threshold $\tau$ is set to 0.3. Moreover, the performance changes on Accuracy and WUPS@0.9 are basically consistent, showing the stable influence of three hyper-parameters for two criteria.

## V. Conclusion

In this paper, we present the problem of open-ended long-form video question answering from the viewpoint of dynamic hierarchical reinforced encoder-decoder network learning. We first employ the dynamic hierarchical encoder to segment long-form video contents and learn the jointly video semantic representations according to the given question. We then develop the reinforced decoder network with a hierarchical attention mechanism to generate the natural language answer for open-ended long-form video question answering. We construct a large-scale long-form video question answering dataset and evaluate the effectiveness of our proposed method through extensive experiments on the long-form dataset and another public short-form dataset.

## VI. Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant No.61602405, No.61751209 and No.61836002, Joint Research Program of ZJU and Hikvision Research Institute. This project is also supported by Alibaba Innovative Research and Microsoft Research Asia.

## 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 *AAAI*, 2017, pp. 4334–4340.

[2] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang, "Video question answering via hierarchical spatio-temporal attention networks." in *IJCAI*, 2017, pp. 3518–3524.

[3] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *ACM MM*. ACM, 2017, pp. 1645–1653.

[4] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *CVPR*, 2017, pp. 2680–2688.

[5] J. Lezama, K. Alahari, J. Sivic, and I. Laptev, "Track to the future: Spatio-temporal video segmentation with long-range motion cues," in *CVPR*. IEEE, 2011, pp. 3369–3376.

[6] S. S. Krishnan and R. K. Sitaraman, "Understanding the effectiveness of video ads: a measurement study," in *IMC*. ACM, 2013, pp. 149–162.

[7] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *CVPR*, 2016, pp. 1029–1038.

[8] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[9] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[10] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," in *EMNLP*, 2016, pp. 1192–1202.

[11] N. Dethlefs and H. Cuayáhuitl, "Hierarchical reinforcement learning for adaptive text generation," in *INLG*. ACL, 2010, pp. 37–45.

[12] Z. Zhao, Z. Zhang, S. Xiao, Z. Yu, J. Yu, D. Cai, F. Wu, and Y. Zhuang, "Open-ended long-form video question answering via adaptive hierarchical reinforced networks," in *IJCAI*, 2018, pp. 3683–3689.

[13] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015, pp. 2425–2433.

[14] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *TIP*, vol. 23, no. 5, pp. 2019–2032, 2014.

[15] J. Yu, X. Yang, F. Gao, and D. Tao, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE transactions on cybernetics*, vol. 47, no. 12, pp. 4014–4024, 2016.

[16] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *CVPR*, 2016, pp. 4995–5004.

[17] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *NIPS*, 2015, pp. 2953–2961.

[18] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *NIPS*, 2014, pp. 1682–1690.

[19] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual qa," in *NIPS*, 2016, pp. 361–369.

[20] R. Li and J. Jia, "Visual question answering with question representation update (qru)," in *NIPS*, 2016, pp. 4655–4663.

[21] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *CVPR*, 2016, pp. 4613–4621.

[22] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NIPS*, 2016, pp. 289–297.

[23] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016, pp. 21–29.

[24] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018, pp. 6077–6086.

[25] B. Patro and V. P. Namboodiri, "Differential attention for visual question answering," in *CVPR*, 2018, pp. 7680–7688.

[26] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *CVPR*, 2017, pp. 326–335.

[27] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Computer Vision and Image Understanding*, vol. 163, pp. 21–40, 2017.

[28] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE transactions on neural networks and learning systems*, no. 99, pp. 1–13, 2018.

[29] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *IJCV*, vol. 124, no. 3, pp. 409–421, 2017.

[30] A. Mazaheri, D. Zhang, and M. Shah, "Video fill in the blank with merging lstms," *arXiv preprint arXiv:1610.04062*, 2016.

[31] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *CVPR*, 2016, pp. 4631–4640.

[32] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in *CVPR*, 2018, pp. 6576–6585.

[33] H. Xue, Z. Zhao, and D. Cai, "Unifying the video and question attentions for open-ended video question answering," *TIP*, vol. 26, no. 12, pp. 5656–5666, 2017.

[34] Z. Zhao, J. Lin, X. Jiang, D. Cai, X. He, and Y. Zhuang, "Video question answering via hierarchical dual-level attention network learning," in *ACM MM*. ACM, 2017, pp. 1050–1058.

[35] Y. Ye, Z. Zhao, Y. Li, L. Chen, J. Xiao, and Y. Zhuang, "Video question answering via attribute-augmented attention network learning," in *SIGIR*. ACM, 2017, pp. 829–832.

[36] K. Gao and Y. Han, "Spatio-temporal context networks for video question answering," in *Pacific Rim Conference on Multimedia*. Springer, 2017, pp. 108–118.

[37] B. Wang, Y. Xu, Y. Han, and R. Hong, "Movie question answering: remembering the textual cues for layered visual contents," in *AAAI*, 2018.

[38] Y. Xu, Y. Han, R. Hong, and Q. Tian, "Sequential video vlad: training the aggregation locally and temporally," *TIP*, vol. 27, no. 10, pp. 4933–4944, 2018.

[39] Z. Zhao, Z. Zhang, X. Jiang, and D. Cai, "Multi-turn video question answering via hierarchical attention context reinforced networks," *TIP*, 2019.

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

[41] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *ICCV*, 2015, pp. 4507–4515.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[43] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio, "An actor-critic algorithm for sequence prediction," *arXiv preprint arXiv:1607.07086*, 2016.

[44] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *CVPR*, 2017, pp. 290–298.

[45] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *ICCV*, 2017, pp. 706–715.

[46] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[47] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

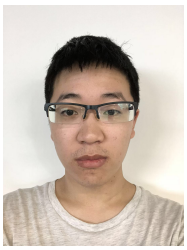
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[50] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *ACL*. ACL, 1994, pp. 133–138.

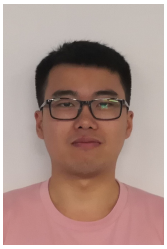
[51] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.

[52] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017, pp. 7008–7024.

**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 degree in computer science. His research interests include machine learning and computer vision.



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

**Xiaohui Yan** is a researcher in Poisson Lab, Huawei Technologies, Beijing, China. His main research interests include text understanding, dialog system, question answer and information retrieval. He has published more than ten papers in prestigious journals and conferences, including TKDE, WWW, AAAI, EMNLP, NAACL and CIKM.

**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.