Open-Ended Long-form Video Question Answering via Adaptive Hierarchical Reinforced Networks

Zhou Zhao[1], Zhu Zhang[1], Shuwen Xiao[1], Zhou Yu[2], Jun Yu[2], Deng Cai[3], Fei Wu[1], Yueting Zhuang[1]
[1] College of Computer Science, Zhejiang University
[2] College of Computer Science, Hangzhou Dianzi University
[3] State Key Lab of CAD&CG, Zhejiang University
{csezhaozhou,dengcai}@gmail.com, {zz,xsw,wufei,yzhuang}@zju.edu.cn, {yuz,yujun}@hdu.edu.cn

Abstract

Open-ended long-form video question answering is challenging problem in visual information retrieval, which automatically generates the natural language answer from the referenced long-form video content according to the question. However, the existing video question answering works mainly focus on the short-form video question answering, due to the lack of modeling the semantic representation of long-form video contents. In this paper, we consider the problem of long-form video question answering from the viewpoint of adaptive hierarchical reinforced encoder-decoder network learning. We propose the adaptive hierarchical encoder network to learn the joint representation of the long-form video contents according to the question with adaptive video segmentation. we then develop the reinforced decoder network to generate the natural language answer for open-ended video question answering. We construct a large-scale long-form video question answering dataset. The extensive experiments show the effectiveness of our method.

1 Introduction

Video question answering is the visual information delivery mechanism that enables user 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. Currently, most of the video question answering approaches mainly focus on the problem of short-form video question answering [Zeng *et al.*, 2017; Zhao *et al.*, 2017; Jang *et al.*, 2017], which learn the semantic video representation from LSTM network layer, and then generate the answer to the given question. Although the existing 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 representation of long-form video contents.

Figure 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-term semantic dependencies [Lezama *et al.*, 2011]. We illustrate a simple example in Figure 1. We show that the answer generation to question "what is the woman holding?" requires the semantic localization of the targeted segment from the long-form video contents. Thus, the simple extension of the existing video question answering works based on frame-level LSTM networks is difficult for modeling the semantic representation of long-form video contents according to the given question [Krishnan and Sitaraman, 2013]. Recently, hierarchical neural encoder [Pan *et al.*, 2016] has been proposed to learn the segment-level video semantic representation with fixed segment length. We then employ the hierarchical neural encoder with attention mechanisms to learn the joint representation of video contents according to the given question. On the other hand, although the video frames are topically consistent, they have different semantic contents [Lezama *et al.*, 2011]. Inspired by binary neurons [Bengio *et al.*, 2013], we then develop the hierarchical neural encoder that unifies the adaptive video segmentation and joint representation modeling of video content according to the given question into a joint learning framework. Thus, leverage the hierarchical neural encoder with adaptive video segmentation is critical for modeling the semantic representation of long-form video contents for video question answering.

In this paper, we study the problem of open-ended long-form video question answering from the viewpoint of adaptive hierarchical reinforced encoder-decoder network learning. We propose the hierarchical neural encoder with adaptive recurrent network to learn segment-level question-aware

video representation with adaptive video segmentation. We devise the reinforced decoder network to generate the natural language answer for open-ended video question answering. We then develop the adaptive hierarchical reinforced network learning framework, named as AHN. When a certain question is issued, AHN can generate natural language answer for it based on the referenced video contents. The main contribution of this paper are as follows:

- Unlike the previous studies, we study the problem of open-ended long-form video question answering from the viewpoint of adaptive hierarchical reinforced encoder-decoder network learning.
- We develop the adaptive hierarchical encoder to learn the segment-level question-aware video representation with adaptive video segmentation. We then devise the reinforced decoder to generate the answer for open-ended video question answering.
- We construct a large-scale dataset for open-ended long-form video question answering and validate the effectiveness of our propose method through extensive experiments.

2 Video Question Answering via Adaptive Hierarchical Reinforced Networks

2.1 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 $\mathbf{v}_2$ is the second frame. We then encode the word-level representation of natural language answer by $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M)$ of length M, where $\mathbf{a}_M$ is the M-th word token. We then denote the collection of video segments by $\{S_1, S_2, \ldots, S_K\}$ of size K, where S_2 is the set of segmented frames and $\mathbf{v}_i \in S_2$ means that the i-th frame belongs to segment S_2. The semantic 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 both the video and answer are sequential data with variant length, it it natural to choose the variant recurrent neural network called long-short term memory network (LSTM) [Hochreiter and Schmidhuber, 1997] to learn their feature representations by

$$\mathbf{i}_t = \delta(\mathbf{W}_i\mathbf{x}_t + \mathbf{G}_i\mathbf{h}_{t-1} + \mathbf{b}_i), \quad (1)$$

$$\hat{\mathbf{c}}_t = tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{G}_t\mathbf{h}_{t-1} + \mathbf{b}_f), \quad (2)$$

$$\mathbf{f}_t = \delta(\mathbf{W}_f\mathbf{x}_t + \mathbf{G}_f\mathbf{h}_{t-1} + \mathbf{b}_f), \quad (3)$$

$$\mathbf{c}_t = \mathbf{i}_t \cdot \hat{\mathbf{c}}_t + \mathbf{f}_t \cdot \mathbf{c}_t, \quad (4)$$

$$\mathbf{o}_t = \delta(\mathbf{W}_o\mathbf{x}_t + \mathbf{G}_o\mathbf{h}_{t-1} + \mathbf{V}_o\mathbf{c}_t + \mathbf{b}_o), \quad (5)$$

$$\mathbf{h}_t = \mathbf{o}_t \cdot tanh(\mathbf{c}_t), \quad (6)$$

where δ represents the sigmoid activation function; $\mathbf{W_s}$, $\mathbf{G_s}$ and $\mathbf{V}_o$ are the weight matrices, and $\mathbf{b_s}$ are the bias vectors. The memory cell $\mathbf{c}_t$ maintains the history of the inputs observed up to the timestep. 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 of the previous hidden state $\mathbf{h}_{t-1}$, followed by a sigmoid activation. Specifically, we denote the semantic representation of video $\mathbf{v}$ by $\mathbf{h}^{(v)} = (\mathbf{h}_1^{(v)}, \mathbf{h}_2^{(v)}, \ldots, \mathbf{h}_N^{(v)})$ and that of answer $\mathbf{a}$ by $\mathbf{h}^{(a)} = (\mathbf{h}_1^{(a)}, \mathbf{h}_2^{(a)}, \ldots, \mathbf{h}_M^{(a)})$ using LSTM networks.

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

2.2 Adaptive Encoder Network Learning

In this section, we propose the adaptive encoder network $f(\mathbf{v}, \mathbf{q})$ that unifies the hierarchical video segmentation and semantic representation learning into a common framework.

The video contents often contain a number of frames with the targeted objects to the given question that evolves over time [Yao *et al.*, 2015]. Inspired by binary neuron [Bengio *et al.*, 2013], we propose the adaptive attentional recurrent neural networks that segment the complex events in video to capture their long-term semantic dependencies, and learn the joint representation of relevant frames and segments according to the question. We first extract the frame-level video feature using ConvNet [Simonyan and Zisserman, 2014] by $\mathbf{v}^{(f)} = (\mathbf{v}_1^{(f)}, \mathbf{v}_2^{(f)}, \ldots, \mathbf{v}_N^{(f)})$, and then learn their semantic representation using LSTM networks. When the end frame of a video event is estimated, we reset the LSTM parameters of the next frame in order to segment the video. To enable the video segmentation, we define an adaptive recurrent neural networks with binary gate function, which decides whether to transfer the LSTM parameters (i.e., hidden state $\mathbf{h}_t^{(v)}$ and memory cell $\mathbf{c}_t^{(v)}$) of the current frame to update the LSTM parameters of the next frame (i.e., hidden state $\mathbf{h}_{t+1}^{(v)}$ and memory cell $\mathbf{c}_{t+1}^{(v)}$) by Equations (1), (2), (3), (4) and (5) or reinitialize them. Formally, the t-th binary gate is defined as a step function, which is computed as a non-linear combination of the feature of the $t+1$-th frame from ConvNet (i.e., $\mathbf{v}_{t+1}^{(f)}$), and the semantic representation of the t-th frame from LSTM networks (i.e., $\mathbf{h}_t^{(v)}$), given by

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

The $1[\cdot]$ is a step function and $\delta(\cdot)$ is a sigmoid function. The $\mathbf{w}_\gamma$ is a learnable row vector, $\mathbf{W}_{\gamma v}$, $\mathbf{W}_{\gamma h}$ and $\mathbf{b}_\gamma$ are the learnable weights and bias. The τ is the threshold parameter of the step function. For example, the inputs of the 3-th binary gate γ_3 are the $\mathbf{h}_3^{(v)}$ (i.e., the hidden state of the 3-th frame from LSTM networks) and $\mathbf{v}_4^{(f)}$ (i.e., the video feature of the 4-th frame) in Figure 2. That is, the binary gate function γ_t decides whether to transfer the parameters of the t-th frame from LSTM networks by $\mathbf{h}_t^{(v)} \leftarrow \mathbf{h}_t^{(v)} \cdot (1 - \gamma_t)$ and $\mathbf{c}_t^{(v)} \leftarrow \mathbf{c}_t^{(v)} \cdot (1 - \gamma_t)$, respectively.

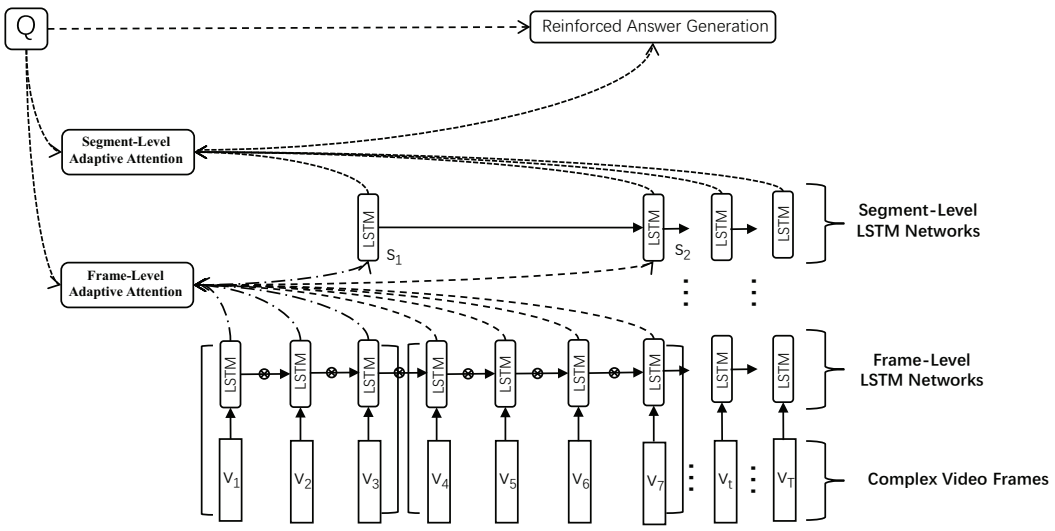


Figure 2: The Framework of Adaptive Hierarchical Reinforced Networks for Open-ended Long-form Video Question Answering. (a) The hierarchical encoder networks learn the joint representation of multimodal attentional video and textual question with adaptive video segmentation. (b) The reinforced decoder networks then generate the natural language answers for open-ended video question answering.

Given the semantic representation of frames $(\mathbf{h}_1^{(v)}, \mathbf{h}_2^{(v)}, \ldots, \mathbf{h}_N^{(v)})$ with binary gate values $(\gamma_1, \gamma_2, \ldots, \gamma_{N-1})$, we then learn the joint question-aware video 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 LSTM networks. Given the question representation $\mathbf{h}^{(q)}$, the frame-level attention for the t-th frame $\mathbf{v}_t \in S_k$ is given by

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

where $\mathbf{W}_h^{(f)}$, $\mathbf{W}_q^{(f)}$ are parameter matrices and $\mathbf{b}^{(v)}$ is the bias vector. The $\mathbf{P}^{(f)}$ is the parameter vector for computing the frame-level attention score. For each frame $\mathbf{v}_t \in S_k$, its activation by the softmax function is given by $\beta_t^{(v)} = \frac{\exp(\alpha_t^{(v)})}{\sum_{\mathbf{v}_t \in S_k} \alpha_t^{(v)}}$. Thus, the attentional semantic representation for segment S_k is then given by $\mathbf{s}_k = \sum_{\mathbf{v}_t \in S_k} \beta_t^{(v)} \mathbf{h}_t^{(v)}$. Given the representation of segments $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$, we devise the segment-level LSTM networks to learn their semantic representation, denoted by $(\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_K^{(s)})$. Therefore, the adaptive encoder network is given by $f(\mathbf{v}, \mathbf{q}) = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_K^{(s)})$.

2.3 Reinforced Decoder Network Learning

In this section, we propose the reinforced decoder network $g(\cdot)$ based on segment-level LSTM networks to generate the open-ended answer for video question answering.

Given the segment-level semantic representation of video contents $f(\mathbf{v}, \mathbf{q}) = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_K^{(s)})$ from segment-level LSTM networks, the question-aware attentional LSTM networks predict the next word in answer by sampling $a_t \sim p_\theta(a_t | \mathbf{a}_{1:t-1}, f(\mathbf{v}, \mathbf{q})) = g_t(f(\mathbf{v}, \mathbf{q}), \mathbf{h}_t^{(a)}, \mathbf{e}_t)$, where $g_t(\cdot)$ is the recurrent answer generator at time t. The $\mathbf{h}_t^{(a)}$ is the decoder state and $\mathbf{e}_t$ is the context vector at time t. Given question representation $\mathbf{h}^{(q)}$, the segment-level attention score for the k-th segment and t-th decoder state for answer generation is given by

$$\alpha_k^{(s)} = \mathbf{P}^{(s)} tanh(\mathbf{W}_h^{(s)} \mathbf{h}_k^{(s)} + \mathbf{W}_q^{(s)} \mathbf{h}^{(q)} + \mathbf{W}_h^{(a)} \mathbf{h}_t^{(a)} + \mathbf{b}^{(s)}),$$

where $\mathbf{W}_h^{(s)}$, $\mathbf{W}_q^{(s)}$, $\mathbf{W}_h^{(a)}$ are parameter matrices and $\mathbf{b}^{(s)}$ is the bias vector. The $\mathbf{P}^{(s)}$ is the parameter vector for computing the question-aware segment-level attention score. For each segment $\mathbf{h}_k^{(s)}$ with decoder state $\mathbf{h}_t^{(a)}$, its activation by the softmax function is given by $\beta_k^{(s)} = \frac{\exp(\alpha_k^{(s)})}{\sum_k \exp(\alpha_k^{(s)})}$. Thus, the attentional question-aware semantic video representation at decoder state $\mathbf{h}_t^{(a)}$ is given by $\mathbf{e}_t = \sum_k \beta_k^{(s)} \mathbf{h}_k^{(s)}$. One common approach to train the proposed decoder network is under the framework of maximum likelihood estimation, given by

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

However, the training based on maximum likelihood estimation makes the learnt decoder network suboptimal [Bahdanau *et al.*, 2016]. 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 $p_\theta(a_t | \mathbf{a}_{1:t-1}, f(\mathbf{v}, \mathbf{q}))$ as the policy. Following the existing visual-semantic embedding works [Ren *et al.*, 2017], 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 $R_{\mathbf{a}}(\hat{\mathbf{a}}) = \|\mathbf{h}^{(a)} - \mathbf{h}^{(\hat{a})}\|^2$. Specifically, we define the expected cumulative reward at each decoding step using value function by

Data	Question Types				
	Object	Number	Color	Location	Action
All	13,588	2,960	5,273	8,959	5,105
Train	10,121	2,228	3,642	6,303	3,705
Valid	1,335	287	547	336	509
Test	2,132	445	1,084	1,754	891

Table 1: Summary of Dataset

$Q(\hat{a}_t|\hat{\mathbf{a}}_{1:t-1}, f(\mathbf{v},\mathbf{q})) = E_{p_\theta(\hat{\mathbf{a}}_{t+1:M}|\hat{\mathbf{a}}_{1:t}, f(\mathbf{v},\mathbf{q}))} R_{\mathbf{a}}(\hat{\mathbf{a}})$. The value function $Q(\hat{a}_t|\hat{\mathbf{a}}_{1:t-1}, f(\mathbf{v},\mathbf{q}))$ is then estimated by aggregating the Monte-Carlo simulation at each decoding step, given by

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

The $\{\hat{\mathbf{a}}^{(1)}_{t+1:M}, \hat{\mathbf{a}}^{(2)}_{t+1:M}, \ldots, \hat{\mathbf{a}}^{(J)}_{t+1:M}\}$ 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 of the reinforced decoder network according to the policy gradient theorem is given by

$$\begin{aligned} & \nabla_\theta \mathcal{L}_{RL}(g(f(\mathbf{v},\mathbf{q}))) \\ = & \sum_{t=1}^{M} \nabla_\theta \log p_\theta(a_t|\mathbf{a}_{1:t-1}, f(\mathbf{v},\mathbf{q})) Q(\hat{a}_t|\hat{\mathbf{a}}_{1:t-1}, f(\mathbf{v},\mathbf{q})). \end{aligned}$$

3 Experiments

3.1 Data Preparation

We construct the long-form video question answering dataset from the ActivityNet data [Heilbron *et al.*, 2015] with natural language descriptions, which contains 20,000 videos amounting to 849 hours and 100,000 descriptions. The average time duration of videos is around 180 seconds and the longest video runs for over 10 minutes. Following the state-of-the-art question generation method [Heilman and Smith, 2010], we generate the question-answer pairs from the video descriptions. Following the existing visual question answering approaches [Antol *et al.*, 2015; Shih *et al.*, 2016; Zhao *et al.*, 2017], we generate five types of questions, which are related to the object, number, color, location and action for the video. We split the generated dataset into three parts: the training, the validation and the testing sets. The five types of long-form video question-answering pairs used for the experiments are summarized in Table 1. The dataset will be provided later.

We then preprocess the long-form video question answering dataset as follows. We first resize each frame to 224×224 and then extract the visual feature of each frame by pretrained VGGNet [Simonyan and Zisserman, 2014], and take the 4096-dimensional feature vector for each frame. We employ the pretrained word2vec model to extract the semantic representation of questions and answers. Specifically, the size of vocabulary set is 8,500 and the dimension of word vector is set to 256. Note that we add a token <eos> to mark the end of the answer, and take the token <unk> for the out-of-vocabulary word.

Method	Accuracy	WUPS@0.0	WUPS@0.9
VQA+	0.254	0.4831	0.3546
MM+	0.2657	0.5162	0.3665
STAN	0.2641	0.5151	0.3696
STVQA	0.3022	0.5376	0.3879
AHN$_{(ml)}$	0.3317	0.5696	0.4103
AHN$_{(rl)}$	**0.3429**	**0.6081**	**0.4313**

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

3.2 Performance Criteria

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

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

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

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

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

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

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

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

  Following the experimental setting in [Malinowski and Fritz, 2014], we choose two WUPS evaluation criteria with the parameter γ to be 0 and 0.9, denoted by WUPS@0.0 and WUPS@0.9, respectively.

3.3 Performance Comparisons

We compare our proposed method with other four state-of-the-art methods for the problem of open-ended video question answering as follows:

Method	Accuracy				
	Object	Number	Color	Location	Action
VQA+	0.2469	0.7563	0.3089	0.133	0.1954
MM+	0.2537	0.7851	0.3217	0.143	0.2011
STAN	0.2515	0.7995	**0.3452**	0.125	0.2169
STVQA	0.3063	0.7918	0.3327	0.166	0.2829
AHN$_{(ml)}$	0.3596	0.8021	0.3417	0.1899	**0.3014**
AHN$_{(rl)}$	**0.3735**	**0.8055**	0.3241	**0.2285**	0.2905

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

Method	WUPS@0.0				
	Object	Number	Color	Location	Action
VQA+	0.5489	0.9478	0.8311	0.1647	0.3017
MM+	0.5513	0.9602	0.8215	0.2715	0.3178
STAN	0.5612	0.9517	0.8421	0.251	0.321
STVQA	0.5847	0.9689	0.8632	0.3001	0.3657
AHN$_{(ml)}$	0.6154	**0.9701**	0.8712	0.3217	0.3874
AHN$_{(rl)}$	**0.6234**	0.9657	**0.9042**	**0.4192**	**0.4091**

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

- **VQA+** method is the extension of image question algorithm [Malinowski and Fritz, 2014], where the mean-pooling layer is added to encode the video.
- **MM+** method [Zeng *et al.*, 2017] is the extension of end-to-end memory network algorithm, which one-layer bi-LSTM network is added to encode the sequence of video frames for answer generation.
- **STAN** method [Zhao *et al.*, 2017] is based on the hierarchical spatio-temporal attention network for learning the joint representation of the dynamic video contents according to the given question.
- **STVQA** method [Jang *et al.*, 2017] is based on the spatio-temporal reasoning algorithm, which employ the spatial and temporal attention on video to answer questions.

Unlike the previous video question answering works, our AHN method learns the hierarchical attentional video representation with adaptive recurrent encoder networks, and then generates the natural language answer with reinforced decoder networks for open-ended video question answering. To exploit the effect of reinforced decoder networks, we denote the our AHN method with reinforced decoder networks by AHN$_{(rl)}$ and the one without reinforced decoder networks by AHN$_{(ml)}$. The weights of both frame-level LSTM networks and segment-level LSTM networks are randomly by a Gaussian distribution with zero mean.

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

Method	WUPS@0.9				
	Object	Number	Color	Location	Action
VQA+	0.3614	0.8165	0.4947	0.2039	0.2378
MM+	0.3741	0.8411	0.5201	0.2187	0.2412
STAN	0.3716	0.8315	0.5142	0.2101	0.254
STVQA	0.3959	0.8478	0.5514	0.2287	0.2578
AHN$_{(ml)}$	0.4214	0.8643	0.5743	0.2462	0.2847
AHN$_{(rl)}$	**0.45**	**0.8762**	**0.5844**	**0.2782**	**0.2937**

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

average value of all the methods on three evaluation criteria. The experiments reveal a number of interesting points:

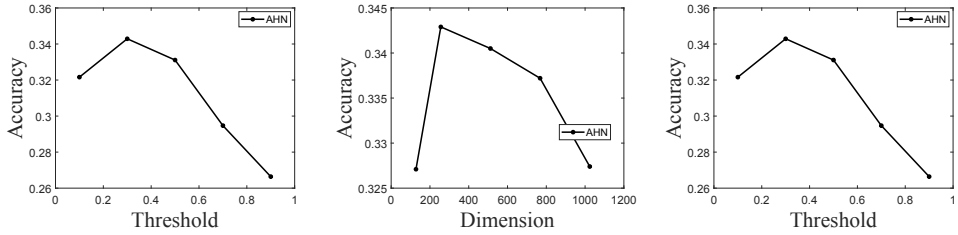
- The methods based on LSTM networks, MM+, STAN, STVQA, AHN outperform the mean-pooling based method VQA+, which suggests that the sequential frame-level representation is critical for the problem.
- The attention based methods STAN, STVQA and AHN achieves better performance than other baselines. This suggests that the joint representation learning of video and question can also improve the performance of open-ended video question answering.
- In all the cases, our AHN method achieves the best performance. This fact shows that the hierarchical attentional video representation with adaptive recurrent encoder networks, and reinforced decoder networks are effective for the problem.

In our approach, there are three essential 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 τ of the step function . We investigate the effect of these 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 1024, and the threshold τ of step function $\gamma_t(\mathbf{v}_{t+1}^{(f)}, \mathbf{h}_t^{(v)})$ from 0.1 to 0.9 on Accuracy in Figures 3(a), 3(b) and 3(c). We then vary these parameters to show their effect on our method using WPUS@0.9 in Figures 4(a), 4(b) and 4(c). Our method achieves the best performance when the dimension of hidden state in frame-level LSTM networks is set to 512, the dimension of hidden state in segment-level LSTM networks is set to 256 and the threshold is set to 0.3.

4 Related Work

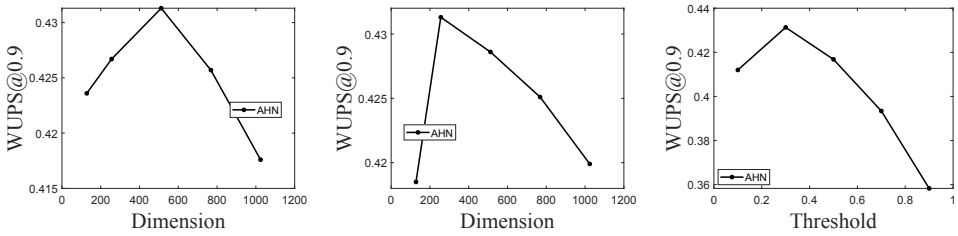
In this section, we briefly review some related work on visual question answering.

The existing approaches for visual question answering can be categorized into image-based question answering methods [Antol *et al.*, 2015; Lu *et al.*, 2016; Li and Jia, 2016; Malinowski and Fritz, 2014; Shih *et al.*, 2016] and video-based question answering ones [Mazaheri *et al.*, 2016; Tapaswi *et al.*, 2016; Zhu *et al.*, 2015; Zeng *et al.*, 2016; Zhao *et al.*, 2017]. Given a natural-language question for the image, the task of image-based question answering is to provide the accurate answer for the given question [Antol *et al.*,



(a) Frame-level LSTM state (b) Segment-level LSTM state (c) Threshold

Figure 3: Effect of frame-level LSTM hidden state dimension, segment-level LSTM hidden state dimension and step function threshold on Accuracy.



(a) Frame-level LSTM state (b) Segment-level LSTM state (c) Threshold

Figure 4: Effect of frame-level LSTM hidden state dimension, segment-level LSTM hidden state dimension and step function threshold on WUPS@0.9.

2015]. Malinowski et. al. [Malinowski and Fritz, 2014] develop the multi-world probabilistic approach for open-ended image question answering. With the development of attention mechanism, Shih et. al. [Shih *et al.*, 2016] propose the spatial-attention mechanism that selects the relevant image regions to the given question. Lu et. al. [Lu *et al.*, 2016] devise the co-attention mechanism and Yang et. al. [Yang *et al.*, 2016] develop the stacked attention method for image question answering. To exploit the complex image question answering task, QRU method [Li and Jia, 2016] is proposed with reasoning process that iteratively selects the relevant image regions and updates the question representation. A survey of existing image question answering methods can be found in [Wu *et al.*, 2016].

As a natural extension of image-based question answering, the video-based question answering has been proposed as a more challenging task [Zeng *et al.*, 2016]. The fill-in-the-blank approaches [Zhu *et al.*, 2015; Mazaheri *et al.*, 2016] complete the missing entry in the video description by ranking candidate answers based on both visual content and contextual video description. Tapaswi et. al. [Tapaswi *et al.*, 2016] propose the three-way scoring function for movie question answering based on both the relevance between given question and textual movie subtitles, and textual movie subtitles and answers. Zhao et. al. [Zhao *et al.*, 2017] propose the hierarchical spatio-temporal attention networks for video question answering. Jang et al. [Jang *et al.*, 2017] develop the spatio-temporal reasoning algorithm, which employ the spatial and temporal attention on video to answer questions. Zeng et al. [Zeng *et al.*, 2017] extend the end-to-end memory network with additional LSTM layer for video question answering. 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 representation of long-form video contents. Unlike the previous studies, we study the problem of open-ended long-form video question answering from the viewpoint of adaptive hierarchical network learning.

5 Conclusion

In this paper, we present the problem of open-ended long-form video question answering from the viewpoint of adaptive hierarchical reinforced encoder-decoder network learning. We first propose the adaptive hierarchical encoder network to learn the video representation from the critical segments of the targeted objects with adaptive video segmentation. We then develop the reinforced decoder network to generate the natural language answer for open-ended 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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No.61602405, No.61702143, No.61622205 and No.61472110, Sponsored by CCF-Tencent Open Research Fund and the China Knowledge Centre for Engineering Sciences and Technology. This work is also Supported by Zhejiang Natural Science Foundation(LZ17F020001) and Key R&D Program of Zhejiang Province(2015C01027).

References

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

[Bahdanau *et al.*, 2016] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*, 2016.

[Bengio *et al.*, 2013] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

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

[Heilbron *et al.*, 2015] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970. IEEE, 2015.

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

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

[Jang *et al.*, 2017] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR 2017*, pages 2680–8, 2017.

[Krishnan and Sitaraman, 2013] S Shunmuga Krishnan and Ramesh K Sitaraman. Understanding the effectiveness of video ads: a measurement study. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 149–162. ACM, 2013.

[Lezama *et al.*, 2011] José Lezama, Karteek Alahari, Josef Sivic, and Ivan Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, pages 3369–3376. IEEE, 2011.

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

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

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

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

[Pan *et al.*, 2016] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*, pages 1029–1038, 2016.

[Ren *et al.*, 2017] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. *CVPR*, 2017.

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

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

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

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

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

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

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

[Zeng *et al.*, 2016] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. *arXiv preprint arXiv:1611.04021*, 2016.

[Zeng *et al.*, 2017] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *AAAI*, pages 4334–4340, 2017.

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

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