Video Question Answering via Attribute-Augmented Attention Network Learning

Yunan Ye   Zhou Zhao   Yimeng Li   Long Chen   Jun Xiao*
Yueting Zhuang
College of Computer Science, Zhejiang University, China
{chryleo,zhaozhou,aquaird,longc,junx,yzhuang}@zju.edu.cn

ABSTRACT

Video Question Answering is a challenging problem in visual information retrieval, which provides the answer to the referenced video content according to the question. However, the existing visual question answering approaches mainly tackle the problem of static image question, which may be ineffectively for video question answering due to the insufficiency of modeling the temporal dynamics of video contents. In this paper, we study the problem of video question answering by modeling its temporal dynamics with frame-level attention mechanism. We propose the attribute-augmented attention network learning framework that enables the joint frame-level attribute detection and unified video representation learning for video question answering. We then incorporate the multi-step reasoning process for our proposed attention network to further improve the performance. We construct a large-scale video question answering dataset. We conduct the experiments on both multiple-choice and open-ended video question answering tasks to show the effectiveness of the proposed method.

CCS CONCEPTS

• **Information systems → Question answering**; • **Computing methodologies → Visual content-based indexing and retrieval**;

KEYWORDS

video question answering; visual information retrieval; attribute

1 INTRODUCTION

Visual information retrieval (VIR) is the information delivery mechanism that enables users to post their queries and then obtain the answers from visual contents [3]. As an emerging kind of recommender system, visual question answering is an important problem for VIR sites, which automatically returns the relevant answer from the referenced visual contents according to users' posted question [1, 5–7]. Currently, most of the existing visual question answering methods mainly focus on the problem of static image

*Corresponding author.

Attribute: "dog", "egg", "bowl", "woman", "plate"
Q: What is a woman boiling in a pot of water?   A:eggs

Figure 1: Video Question Answering

question answering [1, 10–12, 15, 17]. Although existing methods have achieved promising performance in image question answering task, they may still be ineffective applied to the problem of video question answering due to the lack of modeling the temporal dynamics of video contents [21, 22].

The video content often contains the evolving complex interactions and the simple extension of image question answering is thus ineffectively to provide the satisfactory answers. This is because the relevant video information is usually scattered among the entire frames. Furthermore, a number of frames in video are redundant and irrelevant to the question. We give a simple example of video question answering in Figure 1. We demonstrate that the answering for question "What is a woman boiling in a pot of water?" requires the collective information from multiple video frames. Recently, temporal attention mechanisms have been shown to its effectiveness on critical frame extraction for video representation learning [14]. Thus, we then employ the temporal attention mechanisms to model the temporal dynamics of video contents. On the other hand, the utilization of high-level semantic attributes has demonstrated the effectiveness in visual understanding tasks [13]. Furthermore, we observe that the detected attributes are able to enhance the performance of video question answering in Figure 1. Thus, leveraging both temporal dynamic modeling and semantic attributes is critical for learning effective video representation in video question answering.

In this paper, we study the problem of video question answering by modeling its temporal dynamics and semantic attributes. Specifically, we propose the attribute-augmented attention network learning framework that enables the joint frame-level attribute detection and unified video representation learning for video question answering. We then incorporate the multi-step reasoning process for our proposed attribute-augmented attention network to further improve the performance, named as r-ANL. When a certain question is issued, r-ANL can return the relevant answer for it based on the referenced video content. The main contributions of this paper are as follows:

- Unlike the previous studies, we study the problem of video question answering by modeling its temporal dynamics and
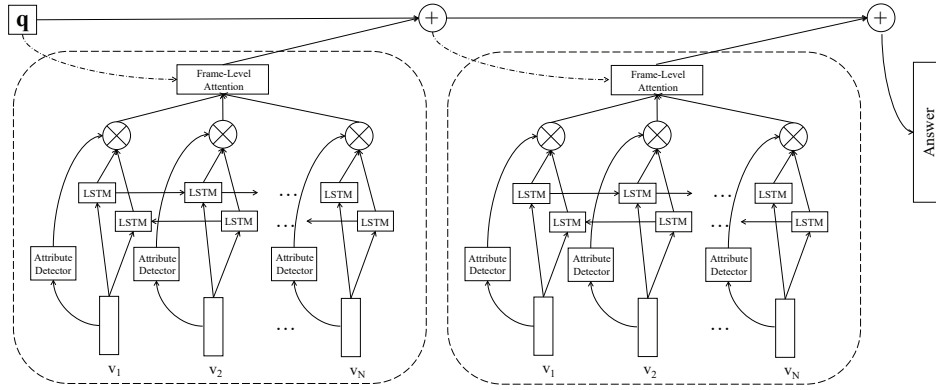
Figure 2: The Overview of Video Question Answering via Attribute-Augmented Attention Network Learning

semantic attributes. We propose the attribute-augmented attention network learning framework that jointly detects frame-level attribute and learns the unified video representation for video question answering.

- We incorporate the multi-step reasoning process for the proposed attention networks to enable the progressive joint representation learning of multimodal temporal attentional video with semantic attributes and textual question to further improve the performance of video question answering.
- We construct a large-scale dataset for video question answering. We evaluate the performance of our method on both multiple choice and open-ended video question answering tasks.

2 VIDEO QUESTION ANSWERING VIA ATTENTION NETWORK LEARNING

2.1 Problem Formulation

Before presenting our method, we first introduce some basic notions and terminologies. We denote the question by $\mathbf{q} \in Q$, the video by $\mathbf{v} \in V$ and the attributes by $\mathbf{a} \in A$, respectively. The frame-level representation of video $\mathbf{v}$ is given by $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N)$, where N is the length of video $\mathbf{v}$. We then denote the frame-level representation of attribute $\mathbf{a}_v$ for video $\mathbf{v}$ by $\mathbf{a}_v = (a_{v,1}, a_{v,2}, \ldots, a_{v,N})$, where $a_{v,i}$ is the set of the attributes for the i-th frame. We then denote W as the vocabulary set or dictionary, where $w_i \in R^{|W|}$ is the one-hot word representation. Since both video and question content are sequential data with variant length, it is natural to choose the variant recurrent neural network called long-short term memory network (LSTM) [8].

Specifically, we learn the feature representation of both video and question by bidirectional LSTM, which consists of a forward LSTM and a backward LSTM [18]. The backward LSTM has the same network structure with the forward one while its input sequence is reversed. We denote the hidden state of the forward LSTM at time t by $\mathbf{h}_t^f$, and the hidden state of the backward LSTM by $\mathbf{h}_t^b$. Thus, the hidden state of video $\mathbf{v}$ at time t from bidirectional layer is denoted by $\mathbf{h}_t = [\mathbf{h}_t^f, \mathbf{h}_{N-t+1}^b]$. The hidden states of video $\mathbf{v}$ is given by

$\mathbf{h}_v = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N)$. We then denote the latent representation of question $\mathbf{q}$ from bidirectional layer by $\mathbf{h}_q$.

Using the notations above, the problem of video question answering is formulated as follows. Given the set of videos V, questions Q and attributes A, our goal is to learn the attribute-augmented attention network such that when a certain question is issued, r-ANL can return the relevant answer for it based on the referenced video content. We present the details of the attribute-augmented attention network learning framework in Figure 2.

2.2 Attribute-Augmented Attention Network Learning

In this section, we propose the attribute-augmented attention network to learn the joint representation of multimodal video content and detected attributes according to the question for both multiple choice and open-ended video question answering tasks.

We first employ a set of pre-trained attribute detectors to obtain the visual attributes for each frame in video $\mathbf{v}$, denoted as $a_{v,i}$ [9, 16, 19]. Each attribute $\mathbf{w}_j \in a_{v,i}$ corresponds to one entry in the vocabulary set W. We then obtain the representation for attribute set by $f(a_{v,i}) = \frac{1}{|a_{v,i}|} \sum_{\mathbf{w}_j \in a_{v,i}} \mathbf{T}_w \mathbf{w}_j$, where $\mathbf{T}_w$ is the embedding matrix for attribute representation and $|a_{v,i}|$ is the size of attribute set for the i-th frame. We thus learn the joint representation of multimodal attributes and frame representation by $g_{\mathbf{v}_i}(a_{v,i}) = \mathbf{h}_i \otimes f(a_{v,i})$, where $\otimes$ is the element-wise product and $\mathbf{h}_i$ is from bidirectional layer at time i.

Inspired by the temporal attention mechanism, we introduce the attribute-augmented attention network to learn the attribute-augmented video representation according to the question for video question answering. Given the question $\mathbf{q}$ and the i-th frame of video $\mathbf{v}$, the temporal attention score s_{qi} is given by:

$$s_{qi} = tanh(\mathbf{W}_q \mathbf{h}_q + \mathbf{W}_v g_{\mathbf{v}_i}(a_{v,i}) + \mathbf{b}_t), \qquad (1)$$

where $\mathbf{W}_q$ and $\mathbf{W}_v$ are parameter matrices and $\mathbf{b}_t$ is bias vector. The $\mathbf{h}_q$ denotes the latent representation of question $\mathbf{q}$ and $g_{\mathbf{v}_i}(a_{v,i})$ is attribute-augmented latent representation of the i-th frame from bidirectional LSTM networks, respectively. For each frame $\mathbf{v}_i$, the activations in temporal dimension by the softmax function is given by $\alpha_{v,i} = \frac{\exp(s_{qi})}{\sum_{i=1}^{N} \exp(s_{qi})}$, which is the normalization of the temporal

attention score. Thus, the temporally attended video representation according to question $\mathbf{q}$ is given by $m_{\mathbf{q}}(\mathbf{v}) = \sum_{i=1}^{N} \alpha_{v,i} g_{\mathbf{v}_i}(a_{v,i})$.

We then incorporate the multi-step reasoning process for the proposed attribute-augmented attention networks to further improve the performance of question-oriented video representation for video question answering. Given the attribute-augmented attention network $m_{\mathbf{q}}(\mathbf{v})$, video $\mathbf{v}$ and question $\mathbf{q}$, the attribute-augmented attention network learning with multi-step reasoning process is given by:

$$\begin{aligned} \mathbf{z}_r &= \mathbf{z}_{r-1} + m_{\mathbf{z}_{r-1}}(\mathbf{v}), \\ \mathbf{z}_0 &= \mathbf{q}, \end{aligned} \tag{2}$$

which is recursively updated. The joint question-oriented video representation is then returned after the R-th reasoning process update, given by $\mathbf{z}_R$. The learning process of reasoning attribute-augmented attention networks in case of $r = 2$ is illustrated in Figure 2.

We next present the objective function of our method for both multiple-choice and open-ended video question answering tasks. For training the model for multiple-choice task, we model video question answering as a classification problem with pre-defined classes. Given the updated joint question-oriented video representation $\mathbf{z}_R$, a softmax function is then employed to classify $\mathbf{z}_R$ into one of the possible answers as

$$p_y = softmax(W_y \mathbf{z}_R + b_y), \tag{3}$$

where $\mathbf{W}_y$ is the parameter matrix and $\mathbf{b}_y$ is the bias vector. On the other hand, for training the model for open-ended video question answering, we employ the LSTM decoder $d(\cdot)$ to generate free-form answers based on the updated joint question-oriented video representation $\mathbf{z}_R$. Given video $\mathbf{v}$, question $\mathbf{q}$ and ground-truth answer $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)$ and the generated answer $\mathbf{o} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_M)$, the loss function $\mathcal{L}(d(\mathbf{z}_R), \mathbf{y})$ is given by:

$$\mathcal{L}(d(\mathbf{z}_R), \mathbf{y}) = \sum_{i=1}^{M} \mathbf{1}[\mathbf{y}_i \neq \mathbf{o}_i], \tag{4}$$

where $\mathbf{1}[\cdot]$ is the indicator function. We denote all the model coefficients including neural network parameters and the result embeddings by Θ. Therefore, the objective function in our learning process is given by

$$\min_{\Theta} \mathcal{L}(\Theta) = \mathcal{L}_{\Theta} + \lambda \|\Theta\|^2, \tag{5}$$

where λ is the trade-off parameter between the training loss and regularization. To optimize the objective function, we employ the stochastic gradient descent (SGD) with the diagonal variant of AdaGrad.

3 EXPERIMENTS

3.1 Data Preparation

We construct the dataset of video question-answering from the YouTube2Text data [2] with natural language descriptions, which consists of 1,987 videos and 122,708 descriptions. Following the state-of-the-art question generation method, we generate the question-answer pairs from the video descriptions. Following the existing visual question answering approaches [1], we generate three types of questions, which are related to the what, who and other queries

Table 1: Summary of Dataset

Data Splitting	Question Types		
	What	Who	Other
train	57,385	27,316	3,649
valid	3,495	2,804	182
test	2,489	2,004	97

for the video. We split the generated dataset into three parts: the training, the validation and the testing sets. The three types of video question-answering pairs used for the experiments are summarized in Table 1. The dataset will be provided later.

We then preprocess the video question-answering dataset as follows. We first sample 40 frames from each video and then re-size each frame to 300×300. We extract the visual representation of each frame by the pretrained ResNet [4], and take the 2,048-dimensional feature vector for each frame [20]. We employ the pretrained word2vec model to extract the semantic representation of questions and answers [23]. Specifically, the size of vocabulary set is 6,500 and the dimension of word vector is set to 256. For training model for open-ended video question answering task, we add a token <eos> to mark the end of the answer phrase, and take the token <Unk> for the out-of-vocabulary word.

3.2 Performance Comparisons

We evaluate the performance of our proposed r-ANL method on both multiple-choice and open-ended video question answering tasks using the evaluation criteria of Accuracy. Given the testing question $\mathbf{q} \in Q_t$ and video $\mathbf{v} \in V_t$ with the groundtruth answer $\mathbf{y}$, we denote the predicted answer by our r-ANL method by $\mathbf{o}$. We then introduce the evaluation criteria of *Accuracy* below:

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

where *Accuracy* = 1 (best) means that the generated answer and the ground-truth ones are exactly the same, while *Accuracy* = 0 means the opposite. When we performance the multiple-choice video question answering task, we set the value of K to 1.
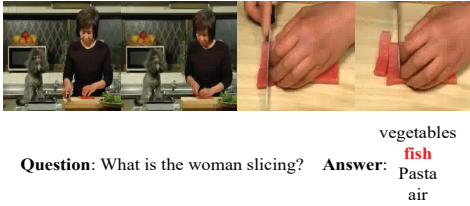
We extend the existing image question answering methods as the baseline algorithms for the problem of video question answering.

- **VQA+** method is the extension of VQA algorithm [1], where we add the mean-pooling layer that obtains the joint video representation from ResNet-based frame features, and then computes the joint representation of question embedding and video representation by their element-wise multiplication for generating open-ended answers.
- **SAN+** method is the incremental algorithm based on stacked attention networks [17], where we add the LSTM network to fuse the sequential representation of video frames for video question answering.

Unlike the previous visual question answering works, our r-ANL method learns the question-oriented video question with multiple reasoning process for the problem of video question answering. To study the effectiveness of attribute-augmented mechanism in our attention network, we evaluate our method with the one without

Table 2: Experimental results on both open-ended and multiple-choice video question answering tasks.

Method	Open-ended VQA task question type				Multiple-choice VQA task question type			
	What	Who	Other	Total accuracy	What	Who	Other	Total accuracy
VQA+	0.2097	0.2486	0.7010	0.386	0.5998	0.3071	0.8144	0.574
SAN+	0.168	0.224	0.722	0.371	0.582	0.288	0.804	0.558
r-ANL$_{(-a)}$	0.164	0.231	0.784	0.393	0.550	0.288	0.825	0.554
r-ANL$_{(1)}$	0.179	0.235	0.701	0.372	0.582	0.261	0.825	0.556
r-ANL$_{(2)}$	0.158	0.249	0.794	0.400	0.603	0.285	0.825	0.571
r-ANL$_{(3)}$	**0.216**	**0.294**	**0.804**	**0.438**	**0.633**	**0.364**	**0.845**	**0.614**

Question: What is the woman slicing? **Answer**: vegetables **fish** Pasta air

Figure 3: Experimental results of both open-ended and multiple-choice video question answering

attributes, denoted as r-ANL$_{(-a)}$. To exploit the effect of reasoning process, we denote our r-ANL method with r reasoning steps by r-ANL$_{(r)}$. The input words of our method are initialized by pre-trained word embeddings with size of 256, and weights of LSTMs are randomly by a Gaussian distribution with zero mean.

Table 2 shows the overall experimental results of the methods on both open-ended and multiple-choice video question answering tasks with different types of questions. The hyperparameters and parameters which achieve the best performance on the validation set are chosen to conduct the testing evaluation. We report the average value of all the methods on three evaluation criteria. We give an example of the experimental results by our method in Figure 3.

4 CONCLUSION

In this paper, we study the problem of video question answering from the viewpoint of attribute-augmented attention network learning. We first propose the attribute-augmented method that learns the joint representation of visual frame and textual attributes. We then develop the attribute-augmented attention network to learn the question-oriented video representation for question answering. We next incorporate the multi-step reasoning process to our proposed attention network that further improve the performance of the method for the problem. We construct a large-scale video question answering dataset and evaluate the effectiveness of our proposed method through extensive experiments.

ACKNOWLEDGMENTS

REFERENCES

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

[2] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*. 2712–2719.

[3] Amarnath Gupta and Ramesh Jain. 1997. Visual information retrieval. *Commun. ACM* 40, 5 (1997), 70–79.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[5] Xiangnan He, Ming Gao, Min-Yen Kan, and Dingxian Wang. 2017. BiRank: Towards Ranking on Bipartite Graphs. *IEEE Trans. Knowl. Data Eng.* (2017), 57–71.

[6] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. 173–182.

[7] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast Matrix Factorization for Online Recommendation with Implicit Feedback. In *SIGIR*. 549–558.

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

[9] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *CVPR*.

[10] Changzhi Luo, Bingbing Ni, Shuicheng Yan, and Meng Wang. 2016. Image Classification by Selective Regularized Subspace Learning. *IEEE Trans. Multimedia* (2016), 40–50.

[11] Liqiang Nie, Meng Wang, Yue Gao, Zheng-Jun Zha, and Tat-Seng Chua. 2013. Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information. *IEEE Trans. Multimedia* (2013), 426–441.

[12] Liqiang Nie, Meng Wang, Zheng-Jun Zha, Guangda Li, and Tat-Seng Chua. 2011. Multimedia answering: enriching text QA with media information. In *SIGIR*. 695–704.

[13] Liqiang Nie, Shuicheng Yan, Meng Wang, Richang Hong, and Tat-Seng Chua. 2012. Harvesting visual concepts for image search with complex queries. In *ACM MM*. 59–68.

[14] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. 2012. Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. *IEEE Trans. Multimedia* (2012), 975–985.

[15] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. 2012. Multimodal Graph-Based Reranking for Web Image Search. *IEEE Trans. Image Processing* (2012), 4649–4661.

[16] Meng Wang, Xueliang Liu, and Xindong Wu. 2015. Visual Classification by ℓ-Hypergraph Modeling. *IEEE Trans. Knowl. Data Eng.* (2015), 2564–2574.

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

[18] Hanwang Zhang, Xindi Shang, Huanbo Luan, Meng Wang, and Tat-Seng Chua. 2016. Learning from collective intelligence: Feature learning using social images and tags. *IEEE Trans. Multimedia* 13 (2016).

[19] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *ACM MM*. ACM, 33–42.

[20] Zhou Zhao, Xiaofei He, Deng Cai, Lijun Zhang, Wilfred Ng, and Yueting Zhuang. 2016. Graph Regularized Feature Selection with Data Reconstruction. *IEEE Trans. Knowl. Data Eng.* (2016), 689–700.

[21] Zhou Zhao, Hanqing Lu, Vincent W. Zheng, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Community-Based Question Answering via Asymmetric Multi-Faceted Ranking Network Learning. In *AAAI*. 3532–3539.

[22] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2016. Expert Finding for Community-Based Question Answering via Ranking Metric Network Learning. In *IJCAI*. 3000–3006.

[23] Zhou Zhao, Lijun Zhang, Xiaofei He, and Wilfred Ng. 2015. Expert Finding for Question Answering via Graph Regularized Matrix Completion. *IEEE Trans. Knowl. Data Eng.* (2015), 993–1004.