Video Question Answering via Hierarchical Dual-Level Attention Network Learning

Zhou Zhao[1] Jinghao Lin[1] Xinghua Jiang[1] Deng Cai[2] Xiaofei He[2] and Yueting Zhuang[1]
[1]College of Computer Science, Zhejiang University, China
[2]State Key Lab of CAD&CG, Zhejiang University, China
{zhaozhou,yzhuang,fenixl,jiangxinghua}@zju.edu.cn,{dengcai,xiaofeihe}@gmail.com

ABSTRACT

Video question answering is a challenging task in visual information retrieval, which provides the accurate answer from the referenced video contents according to the given question. However, the existing visual question answering approaches mainly tackle the problem of static image question answering, which may be ineffectively applied for video question answering directly, due to the insufficiency of modeling the video temporal dynamics. In this paper, we study the problem of video question answering from the viewpoint of hierarchical dual-level attention network learning. We obtain the object appearance and movement information in the video based on both frame-level and segment-level feature representation methods. We then develop the hierarchical dual-level attention networks to learn the question-aware video representations with word-level and question-level attention mechanisms. We next devise the question-level fusion attention mechanism for our proposed networks to learn the question-aware joint video representation for video question answering. We construct two large-scale video question answering datasets. The extensive experiments validate the effectiveness of our method.

KEYWORDS

Video Question Answering; Hierarchical Attention Network

1 INTRODUCTION

Visual question answering is the visual information delivery mechanism that enables users to issue their queries and then collect the answers from the referenced visual contents [8]. Video question answering is the essential problem of visual question answering, which automatically returns the relevant answer from the referenced video contents according to the given question. Currently, most of the existing visual question answering approaches mainly focus on the problem of static image question answering [2, 15, 18, 20, 21, 27, 36, 42]. Several efficient binary image representations have been proposed [25, 26]. Although the existing proposed methods have achieved promising performance in the image question answering task, they may still be ineffectively extended to the problem of video question answering due to the lack of modeling the temporal dynamics of video contents [33].

Question: What is the woman slicing? **Answer**: Fish

Figure 1: Example of Video Question Answering

The visual video contents often contain the evolving complex object interactions such as object appearance and its movement information. The question-aware video information is usually scattered among the critical frames. Furthermore, a number of video frames are redundant and irrelevant to the given question. We illustrate a simple example of video question answering in Figure 1. We show that in order to provide the right result for answering the question "What is the woman slicing?", the collective visual information from multiple video frames is required for answer inference. Thus, the simple extension of the existing image question answering approaches for video question answering is difficult to provide the satisfactory results [43]. We note that the question-aware visual information is always contained in the critical frames. It is thus natural to employ the temporal attention mechanism [43] to localize the targeted video frames according to the given question and learn the effective question-aware video representations. On the other hand, we notice that the video contents often contain different information such as the object appearance from its frames and its movement in the form of motion across the frames. Fortunately, the segment-level video representation has been shown its effectiveness on the content understanding for object motion across the frames [28]. Therefore, leveraging both frame-level and segment-level feature representations is important to the effective question-aware video representation learning for video question answering.

Currently, the existing video-based question answering approaches [22, 53] mainly focus on the fill-in-the-blank task, which is to complete the missing entry in the video description by ranking candidate answers based on both visual content and contextual video description. On the other hand, the movie question answering approach [32] provide the answer ranking based on the textual movie plots. However, the video question answering based on the visual contents only may still not be well explored.

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123364

In this paper, we study the problem of video question answering based on the visual contents from the viewpoint of hierarchical dual-level attention network learning, named as DLAN, as textual information is critical to image and video [6, 41], just like auxiliary information for recommendation tasks [9, 10]. Specifically, we first obtain the object appearance information based on frame-level feature representation by 2D-ConvNet [17] and its movement information based on segment-level features representation by 3D-ConvNet [33] from the video, respectively. We then employ the word-level attention mechanism for the frame-level and segment-level features to learn the augmented word-level attended representations based on the given question. We next learn the question-aware video representations with question-level temporal attention mechanism from both the frame-level and segment-level video representations. We finally learn the question-aware joint video representation for video question answering with question-level fusion attention mechanism on the both frame-level and segment-level attended video representations. When a certain question is issued, DLAN can return the relevant answer for it based on the referenced video content. The main contributions of this paper are summarized as follows:

- Unlike the previous studies, we study the problem of video question answering from the viewpoint of hierarchical dual-level attention network learning. We introduce both word-level frame/segment attention mechanisms to learn the augmented word-level attended representations based on the given question. We then employ the question-level temporal attention mechanism to learn the question-aware video representations.
- We learn the frame-level and segment-level video representations to obtain both the object appearance information and its movement information. We then propose the question-level fusion attention mechanism to learn the joint video representation for video question answering.
- We construct two large-scale datasets for video question answering. The extensive experiments validate the effectiveness of our method.

The rest of this paper is organized as follows. In Section 2, we introduce the problem of video question answering from the viewpoint of hierarchical dual-level attention network learning method. A variety of experimental results are presented in Section 3. We provide a brief review of the related work about visual question answering and video representation learning in Section 4. Finally, we provide some concluding remarks in Section 5.

2 VIDEO QUESTION ANSWERING VIA ATTENTION NETWORKS

In this section, we first present the problem of video question answering from the viewpoint of hierarchical dual-level attention network learning framework. We then propose the question-aware video representations based on the object appearance information and its movement information with hierarchical dual-level attention networks. We next learns the question-aware joint video representation based on both frame-level and segment-level representations with question-level fusion attention attention mechanism for video question answering.

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 answer by $\mathbf{a} \in A$, respectively. The word-level representation of question $\mathbf{q}$ is given by $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_N)$, where N is the length of question $\mathbf{q}$ and $\mathbf{q}_i$ is the embedding for the i-th word in the question. We then denote the frame-level representation for video $\mathbf{v}$ by $\mathbf{v}^{(f)} = (\mathbf{v}_1^{(f)}, \mathbf{v}_2^{(f)}, \ldots, \mathbf{v}_{M^{(f)}}^{(f)})$, where $M^{(f)}$ is the number of frames in video $\mathbf{v}^{(f)}$, and $\mathbf{v}_j^{(f)}$ is the embedding of the j-th frame by pre-trained 2D-ConvNet. We next denote the segment-level representation of video $\mathbf{v}$ by $\mathbf{v}^{(s)} = (\mathbf{v}_1^{(s)}, \mathbf{v}_2^{(s)}, \ldots, \mathbf{v}_{M^{(s)}}^{(s)})$, where $M^{(s)}$ is the number of segments in video $\mathbf{v}^{(s)}$, and $\mathbf{v}_k^{(s)}$ is the embedding of the k-th segment by pre-trained 3D-ConvNet.

Since the textual question, the frame-level and the segment-level video representations are sequential data with variant length, it is natural to choose the variant recurrent neural network called Long-Short Term Memory network (LSTM) [12] to learn their feature representations, given by

$$\begin{aligned}
\mathbf{f}_t &= \delta_g(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f) \\
\mathbf{i}_t &= \delta_g(\mathbf{W}_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i) \\
\mathbf{o}_t &= \delta_g(\mathbf{W}_o \mathbf{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o) \\
\mathbf{c}_t &= \mathbf{f}_t \otimes \mathbf{c}_{t-1} + \mathbf{i}_t \otimes \delta_h(\mathbf{W}_c \mathbf{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \\
\mathbf{h}_t &= \mathbf{o}_t \otimes \delta_h(\mathbf{c}_t),
\end{aligned} \tag{1}$$

where $\mathbf{W}$s and $\mathbf{U}$s are the parameter matrices, and $\mathbf{b}$s are the bias vectors. The $\mathbf{x}_t$, $\mathbf{h}_t$ and $\mathbf{c}_t$ are input vector, output vector and cell state vector, respectively. The activation $\delta_g(\cdot)$ and $\delta_h(\cdot)$ are the sigmoid function and hyperbolic tangent function, respectively. The $\otimes$ denotes the element-wise product operator. The gates in LSTM cell can modulate the interactions between the memory cell itself and its environment. The architecture structure of LSTM can be found in [12].

We thus denote the output states of frame-level video representations by $\mathbf{h}^{(f)} = (\mathbf{h}_1^{(f)}, \mathbf{h}_2^{(f)}, \ldots, \mathbf{h}_{M^{(f)}}^{(f)})$ where $\mathbf{h}_i^{(f)}$ is the output state of the i-th frame in video $\mathbf{v}$. We then consider the output states of segment-level video representations by $\mathbf{h}^{(s)} = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_{M^{(s)}}^{(s)})$ where $\mathbf{h}_j^{(s)}$ is the output state of the j-th segment in video $\mathbf{v}$. Thus, the output states for dual-level video representation based on the first-level LSTM encoding networks is denoted by $\mathbf{h}^{(v)} = (\mathbf{h}^{(f)}, \mathbf{h}^{(s)})$. We then represent the output states of question representation by $\mathbf{h}^{(q)} = (\mathbf{h}_1^{(q)}, \mathbf{h}_2^{(q)}, \ldots, \mathbf{h}_N^{(q)})$, where $\mathbf{h}_k^{(q)}$ is the output state of the k-th word in question $\mathbf{q}$. We next denote the output states of the second-level LSTM encoding networks by $\mathbf{z}^{(f)} = (\mathbf{z}_1^{(f)}, \mathbf{z}_2^{(f)}, \ldots, \mathbf{z}_{M^{(f)}}^{(f)})$ for frame-level augmented video representations and $\mathbf{z}^{(s)} = (\mathbf{z}_1^{(s)}, \mathbf{z}_2^{(s)}, \ldots, \mathbf{z}_{M^{(s)}}^{(s)})$ for segment-level augmented video representations, respectively. The output states of dual-level video representation based on the second-level LSTM encoding networks is denoted by $\mathbf{z}^{(v)} = (\mathbf{z}^{(f)}, \mathbf{z}^{(s)})$. We present the details of the hierarchical dual-level attention network learning framework in Figure 2.

Using the notations above, the problem of video question answering is formulated as follows. Given the set of videos V,

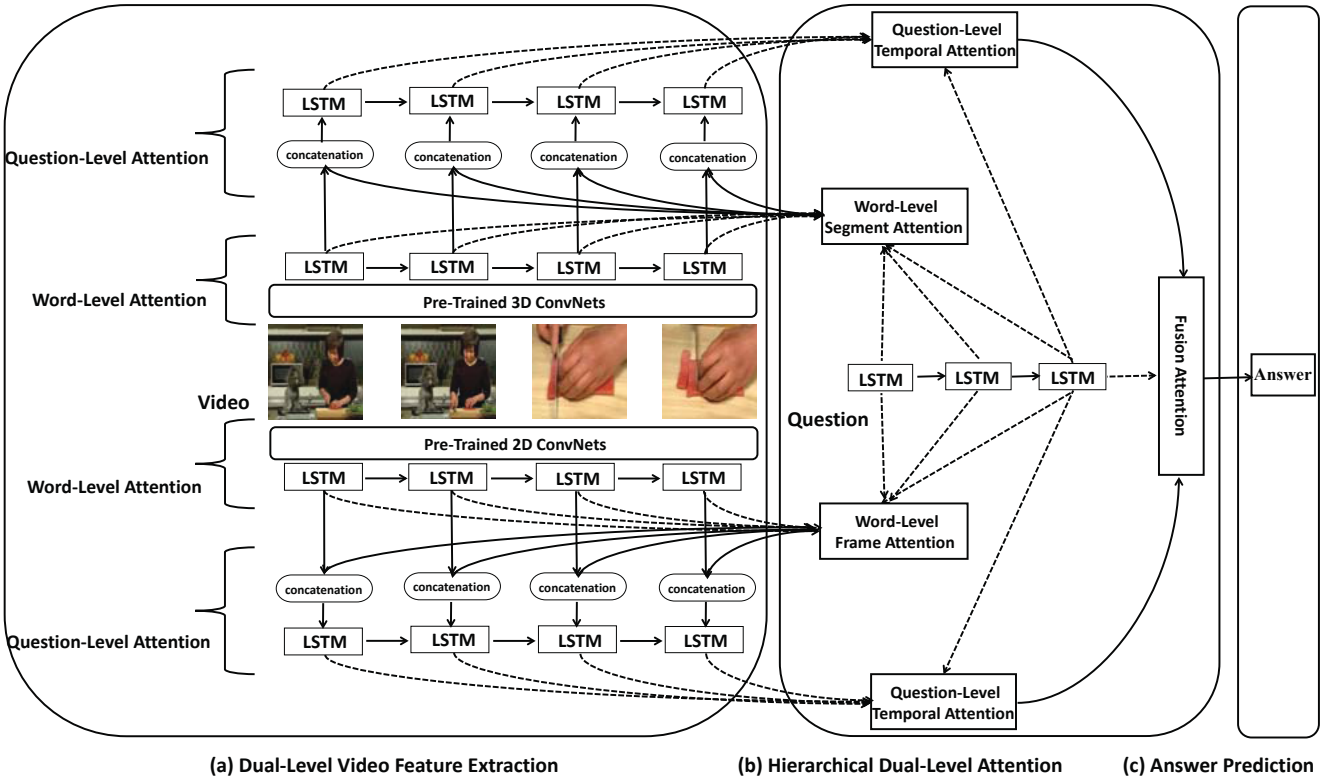


Figure 2: The Overview of Video Question Answering via Hierarchical Dual-Level Attention Network Learning. (a) We obtain the object appearance information and its movement information based on the frame-level and segment-level video representations using 2D/3D-ConvNet. (b) We learn the question-aware joint video representation based hierarchical dual-level attention network and question-level fusion attention mechanism. (c) We learn the answer prediction model based on softmax loss and question-aware joint video representation for video question answering.

questions Q and the associated answers A, our goal is to learn the hierarchical dual-level attention network such that when a certain question is issued, DLAN can return the relevant answer for it based on the reference video content.

2.2 Hierarchical Dual-Level Attention Network Learning

In this section, we present the hierarchical dual-level attention network learning framework to obtain the question-aware video representation for video question answering.

We first propose the dual-level video feature extraction methods for video representations, named as frame-level feature extraction and segment-level feature extraction. We extract the frame-level feature using 2D-ConvNet [17] by $\mathbf{v}^{(f)} = (\mathbf{v}_1^{(f)}, \mathbf{v}_2^{(f)}, \ldots, \mathbf{v}_{M^{(f)}}^{(f)})$ and the segment-level feature using 3D-ConvNet [33] by $\mathbf{v}^{(s)} = (\mathbf{v}_1^{(s)}, \mathbf{v}_2^{(s)}, \ldots, \mathbf{v}_{M^{(s)}}^{(s)})$, respectively. Thus, the dual-level video features consists of frame-level and segment-level features expressed as $\mathbf{v} = (\mathbf{v}^{(f)}, \mathbf{v}^{(s)})$. We then learn the dual-level video representation using LSTM networks, denoted by $\mathbf{h}^{(v)} = (\mathbf{h}^{(f)}, \mathbf{h}^{(s)})$, where vector $\mathbf{h}^{(f)} = (\mathbf{h}_1^{(f)}, \mathbf{h}_2^{(f)}, \ldots, \mathbf{h}_{M^{(f)}}^{(f)})$ is composed of all the output states for frame-level features, and vector $\mathbf{h}^{(s)} = (\mathbf{h}_1^{(s)}, \mathbf{h}_2^{(s)}, \ldots, \mathbf{h}_{M^{(s)}}^{(s)})$ consists of all the output states for segment-level features.

We next develop the hierarchical dual-level attention networks to learn the question-aware joint video representation based on the dual-level video representations and the given question. We first employ the word-level frame attention networks to learn the augmented video frame representations. Given the question representation $\mathbf{h}^{(q)} = (\mathbf{h}_1^{(q)}, \mathbf{h}_2^{(q)}, \ldots, \mathbf{h}_N^{(q)})$, the word-level frame attention score $\mathbf{s}_{ik}^{(f,w)}$ for the k-th question word and the i-th video frame is given by

$$s_{ik}^{(f,w)} = \mathbf{P}^{(f,w)} tanh(\mathbf{W}_{hs}^{(f,w)} \mathbf{h}_i^{(f)} + \mathbf{W}_{qs}^{(w)} \mathbf{h}_k^{(q)} + \mathbf{b}_s^{(w)}), \quad (2)$$

where $\mathbf{W}_{hs}^{(f,w)}$, $\mathbf{W}_{qs}^{(w)}$ are parameter matrices and $\mathbf{b}_s^{(w)}$ is the bias vector. The $\mathbf{P}^{(f,w)}$ is the parameter vector for computing the word-level frame attention score. The $\mathbf{h}_i^{(f)}$ is the output state of the i-th frame in video $\mathbf{v}$ and $\mathbf{h}_k^{(q)}$ is the output state of the k-th word in question $\mathbf{q}$. For each word $\mathbf{q}_k$ in question $\mathbf{q}$, its activation for the

i-th frame by the softmax function is given by $\alpha_{i,k}^{(f)} = \frac{\exp(s_{ik}^{(f,w)})}{\sum_k \exp(s_{ik}^{(f,w)})}$, which is the normalization of the word-level frame attention scores. Thus, the word-level frame attended representation is then given by $\mathbf{h}_i^{(f,w)} = \sum_k \alpha_{i,k}^{(f)} \mathbf{h}_k^{(q)}$. Therefore, the augmented frame-level representation of the i-th frame is given by the concatenation between the i-th output state by LSTM network and the output states of relevant words in question by $\hat{\mathbf{h}}_i^{(f)} = (\mathbf{h}_i^{(f)}, \mathbf{h}_i^{(f,w)})$. Similarly, the word-level segment attention vector score for the j-th segment is given by

$$s_{jk}^{(s,w)} = \mathbf{P}^{(s,w)} tanh(\mathbf{W}_{hs}^{(s,w)} \mathbf{h}_j^{(s)} + \mathbf{W}_{qs}^{(w)} \mathbf{h}_k^{(q)} + \mathbf{b}_s^{(w)}), \quad (3)$$

where $\mathbf{P}^{(s,w)}$ is the parameter vector for computing the word-level segment attention score. The activation for the j-th segment with the k-th question word is given by $\alpha_{j,k}^{(s)} = \frac{\exp(s_{jk}^{(s,w)})}{\sum_k \exp(s_{jk}^{(s,w)})}$. The word-level segment attended representation for the j-th segment is then given by $\mathbf{h}_j^{(s,w)} = \sum_k \alpha_{j,k}^{(s)} \mathbf{h}_k^{(q)}$. Thus, the augmented representation for the jth segment is given by $\hat{\mathbf{h}}_j^{(s)} = (\mathbf{h}_j^{(s)}, \mathbf{h}_j^{(s,w)})$.

On the other hand, a number of video frames/segments are redundant and irrelevant to the given question. It is thus important to learn the targeted frames/segments with question-level temporal attention mechanism. We first encode the frame-level and segment-level augmented representations by the second-level LSTM networks, and then learn the question-aware video representations with question-level temporal attention mechanism. Given the frame-level augmented video representations $\hat{\mathbf{h}}^{(f)} = (\hat{\mathbf{h}}_1^{(f)}, \hat{\mathbf{h}}_2^{(f)}, \ldots, \hat{\mathbf{h}}_{M^{(f)}}^{(f)})$ and the segment-level augmented video representations $\hat{\mathbf{h}}^{(s)} = (\hat{\mathbf{h}}_1^{(s)}, \hat{\mathbf{h}}_2^{(s)}, \ldots, \hat{\mathbf{h}}_{M^{(s)}}^{(s)})$, the output states of the second-level LSTM encoding networks by $\mathbf{z}^{(f)} = (\mathbf{z}_1^{(f)}, \mathbf{z}_2^{(f)}, \ldots, \mathbf{z}_{M^{(f)}}^{(f)})$ for frame-level augmented video representations and $\mathbf{z}^{(s)} = (\mathbf{z}_1^{(s)}, \mathbf{z}_2^{(s)}, \ldots, \mathbf{z}_{M^{(s)}}^{(s)})$ for segment-level augmented video representations, respectively. Given the encoded frame-level augmented video representations $\mathbf{z}^{(f)} = (\mathbf{z}_1^{(f)}, \mathbf{z}_2^{(f)}, \ldots, \mathbf{z}_{M^{(f)}}^{(f)})$ and the last output state of given question $\mathbf{h}_N^{(q)}$, the question-level frame temporal attention score for the i-th encoded augmented frame $s_i^{(f,q)}$ is given by

$$s_i^{(f,q)} = \mathbf{P}^{(f,q)} tanh(\mathbf{W}_{zs}^{(f,q)} \mathbf{z}_i^{(f)} + \mathbf{W}_{qs}^{(q)} \mathbf{h}_N^{(q)} + \mathbf{b}_s^{(q)}), \quad (4)$$

where $\mathbf{W}_{zs}^{(f,q)}$, $\mathbf{W}_{qs}^{(q)}$ are parameter matrices and $\mathbf{b}_s^{(q)}$ is the bias vector. The $\mathbf{P}^{(f,q)}$ is the parameter vector for computing the question-level frame temporal attention score. We choose the non-linear function tanh that controls the amount of information from each frame. For each encoded frame $\mathbf{z}_i^{(f)}$, its activation in temporal dimension by the softmax function is given by $\beta_i^{(f)} = \frac{\exp(s_i^{(f,q)})}{\sum_i \exp(s_i^{(f,q)})}$, which is the normalization of the question-level frame attention scores. Thus, the question-level frame attended representation is given by $\mathbf{z}^{(f,q)} = \sum_i \beta_i^{(f)} \mathbf{z}_i^{(f)}$. Similarly, given the segment-level augmented representations $\mathbf{z}^{(s)} = (\mathbf{z}_1^{(s)}, \mathbf{z}_2^{(s)}, \ldots, \mathbf{z}_{M^{(s)}}^{(s)})$, the question-level segment temporal attention score for the j-th encoded augmented segment $s_j^{(s,q)}$ is given by

$$s_j^{(s,q)} = \mathbf{P}^{(s,q)} tanh(\mathbf{W}_{zs}^{(s,q)} \mathbf{z}_j^{(s)} + \mathbf{W}_{qs}^{(q)} \mathbf{h}_N^{(q)} + \mathbf{b}_s^{(q)}), \quad (5)$$

where $\mathbf{P}^{(s,q)}$ is the parameter vector for computing the question-level segment temporal attention score. Therefore, the activation of the j-th segment is given by $\beta_j^{(s)} = \frac{\exp(s_j^{(s,q)})}{\sum_j \exp(s_j^{(s,q)})}$. The question-level segment attended representation is given by $\mathbf{z}^{(s,q)} = \sum_j \beta_j^{(s)} \mathbf{z}_j^{(s)}$. We then consider the question-aware dual-level temporal attended video representation by $(\mathbf{z}^{(f,q)}, \mathbf{z}^{(s,q)})$.

We then learn the question-aware joint representation for video question answering with question-level fusion attention mechanism. Given the dual-level temporal attended video representation $\mathbf{z} = (\mathbf{z}^{(f,q)}, \mathbf{z}^{(s,q)})$, the question-level fusion attention score for frame-level representations $s^{(f)}$ is given by

$$s^{(f)} = \mathbf{p} tanh(\mathbf{W}_{zs}^{(f)} \mathbf{z}^{(f,q)} + \mathbf{W}_{qs}^{(q)} \mathbf{h}_N^{(q)} + \mathbf{b}_s), \quad (6)$$

and the fusion attention score for segment-level representations $s^{(s)}$ is given by

$$s^{(s)} = \mathbf{p} tanh(\mathbf{W}_{zs}^{(s)} \mathbf{z}^{(s,q)} + \mathbf{W}_{qs}^{(q)} \mathbf{h}_N^{(q)} + \mathbf{b}_s), \quad (7)$$

where $\mathbf{W}_{zs}^{(f)}$, $\mathbf{W}_{zs}^{(s)}$, $\mathbf{W}_{qs}^{(q)}$ are parameter matrices and $\mathbf{b}_s$ is the bias vector. $\mathbf{p}$ is a parameter vector for computing the fusion attention score. Therefore, the question-aware joint video representation based on dual-level attended representations $(\mathbf{z}^{(f,q)}, \mathbf{z}^{(s,q)})$ with fusion attention mechanism is given by

$$\mathbf{z} = \frac{\exp(s^{(f)})}{\exp(s^{(f)}) + \exp(s^{(s)})} \mathbf{z}^{(f,q)} + \frac{\exp(s^{(s)})}{\exp(s^{(f)}) + \exp(s^{(s)})} \mathbf{z}^{(s,q)}. \quad (8)$$

Following the existing visual question answering models [2, 18], we model the problem of video question answering as a classification task with pre-defined classes. Given the question-aware joint video representation $\mathbf{z}$, a softmax function is employed to classify $\mathbf{z}$ into one of the possible answers as

$$p_a = softmax(\mathbf{W}_z \mathbf{z} + \mathbf{b}_z),$$

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

Table 1: Summary of YouTube2Text Dataset

Data Splitting	Question Type				Total
	Object	Count	Location	Person	
Train	13,766	758	99	9,686	24,309
Valid	4,220	169	45	3,200	7,634
Test	8,966	573	88	6,562	16,189

3 EXPERIMENTS

In this section, we first construct two video question answering datasets, and then conduct several experiments on them, to show the effectiveness of our approach DLAN for video question answering.

Table 2: Summary of VideoClip Dataset

Data Splitting	Question Type				Total
	Object	Count	Color	Location	
Train	18,951	7,278	10,676	2,479	39,384
Valid	3,972	1,597	2,698	675	8,942
Test	2,540	961	1,376	332	5,209

3.1 Data Preparation

We construct two video question answering datasets from the YouTube2Text data [7] and the VideoClip data [19] with natural language descriptions. The YouTube2Text data consists of 1,987 videos and 122,708 natural language descriptions. The videoclip data is composed of 201,068 video clips and 287,933 descriptions. Following the state-of-the-art question generation method [11], we generate the question-answer pairs from the video descriptions. Following the existing visual question answering approaches [2, 15, 20], we first generate four types of questions for YouTube2Text data, which are related to the Object, Count, Location and Person queries for the video, and then generate four types of questions for VideoClip data, which are about the Object, Count, Color and Location queries. We split the generated dataset into three parts: the training, the validation and the testing sets. The four types of video question-answering pairs based on YouTube2Text data is summarized in Table 3 and the question-answering pairs based on VideoClip data is illustrated in Table 2. The video question answering datasets will be provided later.

We preprocess the video question answering datasets as follows. We sample 60 frames from each video in YouTube2Text data and 20 frames from each video in VideoClip data for frame-level video representation. For both dataset, we resize each frame to 224×224 and extract the frame representation by the pretrained VGGNet [29], and take the 4,096-dimensional feature vector for each frame. To obtain the segment-level video representation, we employed the pretrained C3DNet [33] to obtain 45 segments from each video in YouTube2Text data and 20 segments from each video in VideoClip data. Each segment contains 16 frames for both datasets. We employ the pretrained word2vec model to learn the semantic representation of questions and answers. Specifically, the size of vocabulary set is 6,500 and the dimension of word vector is set to 256. Following the existing approaches [2, 20], we set the number of pre-defined answer classes to 300 in VideoClip data and 500 in YouTube2Text data.

3.2 Performance Comparisons

We evaluate the performance of our proposed DLAN method using the evaluation criteria of *Accuracy*. Given the testing question $\mathbf{q} \in Q_t$ and video $\mathbf{v} \in V_t$ with the ground-truth answer $\mathbf{a}$, we denote the predicted answer by our DLAN method by $\mathbf{o}$. We then introduce the evaluation criteria of *Accuracy* below:

$$Accuracy = \frac{1}{|Q_t|} \sum_{\mathbf{q} \in Q_t, \mathbf{v} \in V_t} (1 - \mathbf{1}[\mathbf{a} \neq \mathbf{o}]),$$

where $Accracy = 1$ (best) means that the predicted answer and the ground-truth ones are exactly the same, while $Accuracy = 0$ means the opposite. The $\mathbf{1}[\mathbf{a} \neq \mathbf{o}]$ is the indicator function.

We extend the existing image question answering and video caption methods as the baseline algorithms for the problem of video question answering following the strategies in [45].

- **VQA+** method is the extension of image question answering algorithm [2], where one-layer LSTM network is added to encode the video. Both video and question features are fused into the joint representation with element-wise multiplication for answer prediction.
- **SS+** method is the extension of sequence-to-sequence algorithm [34], where the LSTM network first encodes the video, and then encodes the question, finally predicts the answer.
- **SA+** method is the extension of video caption algorithm [43], where temporal attention mechanism is employed for learning the question-aware video representation.
- **MN+** method is the extension of end-to-end memory network algorithm [31], where one-layer bi-LSTM network is added to encode the sequence of video frames for answer prediction.

Unlike the previous works, our DLAN method learns the question-aware joint video representation with hierarchical dual-level attention network learning for the problem of video question answering. That is, the proposed DLAN method learns the question-aware joint video representation for video question answering based on both frame-level and segment-level video features with word-level and question-level attention mechanisms. To exploit the effect of frame-level video representation and segment-level video representation to the performance of video question answering, we denote our method with frame-level video representation only by $DLAN_{(f)}$ and our method with segment-level video representation by $DLAN_{(s)}$. The weights of LSTMs are randomly sampled by a Gaussian distribution with zero mean. For other experiment settings, we employed the initial learning rate of 1e-3, and a dropout rate of 0.6 after every LSTM layer. The early stopping approach with a limit of 5 iterations is also used for trainning. The batch size is set to 100, and ADAM gradient descent is used [16].

Tables 3 and 2 show the overall experimental results of the methods on all types of questions based on the evaluation criteria of *Accuracy* using YouTube2Text data and VideoClip data, respectively. The hyperparamters 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 the evaluation criteria of *Accuracy*.

The experimental results reveal a number of interesting points:

- The method based on temporal attention, SA+, outperforms other baseline methods VQA+, SS+ and MN+, which suggests that the temporal attention mechanism is critical for the problem of video question answering.
- Our method $DLAN_{(f)}$ achieves better performance than other baselines. The method $DLAN_{(f)}$ leverages both the word-level frame attention and question-level temporal attention to learn the question-aware video representation for video question answering. This suggests that the hierarchical dual-level attention mechanism can also improve the performance for the problem.

Table 3: Experimental results on YouTube2Text Dataset.

Model	Overall Accuracy	Question Type				Video Representation Level	
		Object	Count	Location	Person	Frame-Level	Segment-Level
VQA+	0.3139	0.1682	0.7661	0.2614	0.4675	✓	
SS+	0.3147	0.1973	0.7714	0.1023	0.4313	✓	
SA+	0.3203	0.1721	0.7853	0.1932	0.4776	✓	
MN+	0.2944	0.1522	0.6911	0.0909	0.4541	✓	
$\text{DLAN}_{(f)}$	0.3265	0.1854	0.7976	0.1818	0.4738	✓	
$\text{DLAN}_{(s)}$	0.3417	0.2129	0.7818	**0.2955**	0.4736		✓
DLAN	**0.3633**	**0.2308**	**0.8202**	0.2617	**0.4995**	✓	✓

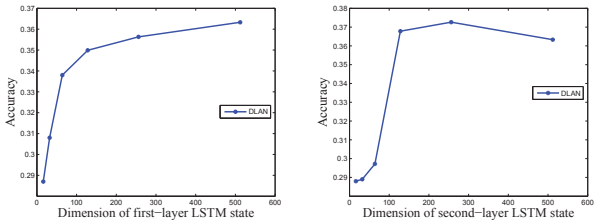
Table 4: Experimental results on VideoClip Dataset.

Model	Overall Accuracy	Question Type				Video Representation Level	
		Object	Count	Color	Location	Frame-Level	Segment-Level
VQA+	0.5061	0.41	0.7867	0.5218	0.3645	✓	
SS+	0.4117	0.31	0.7476	0.396	0.2827	✓	
SA+	0.5144	0.4233	0.7956	0.527	0.3464	✓	
MN+	0.5015	0.4088	0.7749	0.5117	0.3727	✓	
$\text{DLAN}_{(f)}$	0.5446	0.4539	**0.8044**	**0.556**	**0.4398**	✓	
$\text{DLAN}_{(s)}$	0.5208	0.435	0.795	0.5182	0.3946		✓
DLAN	**0.5506**	**0.4732**	0.795	0.5509	0.4337	✓	✓

- Our method *DLAN* obtains the best performance on the evaluation criteria of overall accuracy. This fact shows that the question-aware joint video representation learning based on both frame-level and segment-level representations can further improve the performance of video question answering.

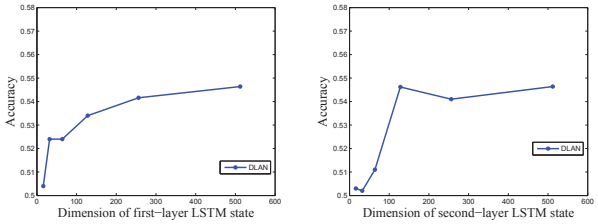
In our approach, there are two essential parameters, which are the dimension of hidden states in the first-layer LSTM networks, and the dimension of hidden states in the second-layer LSTM networks. We employ the word-level attention mechanism for the output states of the first-layer LSTM networks to learn the question-aware frame/segment representations. We then leverage the question-level attention mechanism for the output states of the second-layer LSTM networks to learn the question-aware video representations. We investigate the effect of the dimension parameters in our method by varying either the dimension of the first-layer LSTM hidden states or the dimension of the second-layer LSTM hidden states from 16 to 512, with another layer fixed to 256. We first illustrate the performance of our method DLAN by varying the dimensions of the first-layer LSTM hidden states and the second-layer LSTM hidden states using YouTube2Text data in Figures 3(a) and 3(b). We then show the performance of our method DLAN by varying the dimensions of the first-layer LSTM hidden states and the second-layer LSTM hidden states using VideoClip data in Figures 4(a) and 4(b). The x-axis denotes the dimension of LSTM hidden states and the y-axis show the accuracy of our method in all figures. Our method achieves the stable performance when the dimension of the first-layer LSTM hidden states is set to 256 and the dimension of the second-layer LSTM hidden states is set to 256.

Figures 5(a) and 5(b) show the convergence of our method and Figures 6(a) and 6(b) illustrate the running time of our method



(a) First-layer LSTM state (b) Second-layer LSTM state

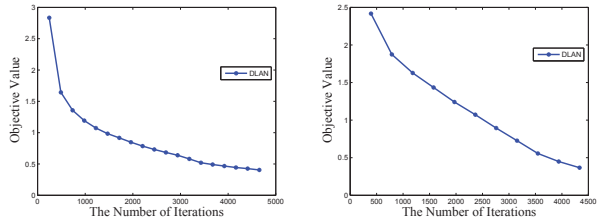
Figure 3: Effect of dimension of hidden LSTM state on YouTube2Text data.



(a) First-layer LSTM state (b) Second-layer LSTM state

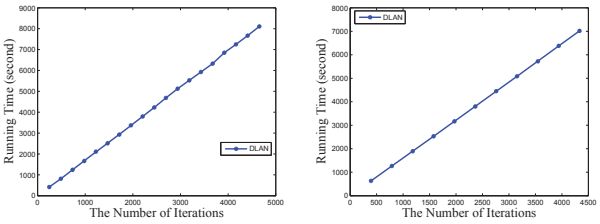
Figure 4: Effect of dimension of hidden LSTM state on VideoClip data.

using both datasets. The x-axis denotes the number of iterations in all figures. The y-axis in Figures 5(a) and 5(b) shows the objective value and the y-axis in Figures 6(a) and 6(b) illustrates the running time of our proposed method. We report that the training time of



(a) YouTube2Text (b) VideoClip

Figure 5: Objective value versus iterations.



(a) YouTube2Text (b) VideoClip

Figure 6: Running time versus iterations.

our method is 8,108 seconds using YouTube2Text data and 7,015 seconds using VideoClip data. This study validates the efficiency of our method.

To understand how the hierarchical dual-level attention network learns the question-aware joint video representation for video question answering, we present the attentional results of several video question answering examples in Figures 7(a), 7(b), 7(c) and 7(d). The set of video question answering examples covers the questions of the Object, Count and Color types. We show the attentional results of our method on these examples in Figure 8. The segment-level temporal attentional results for all examples are visualized using thermodynamic diagram [23]. The frame-level and segment-level fusion attention scores for question-aware joint video representation learning are shown by numerical scores. We observe that both question-level temporal attention scores and fusion attention scores vary according to different types of questions and videos.

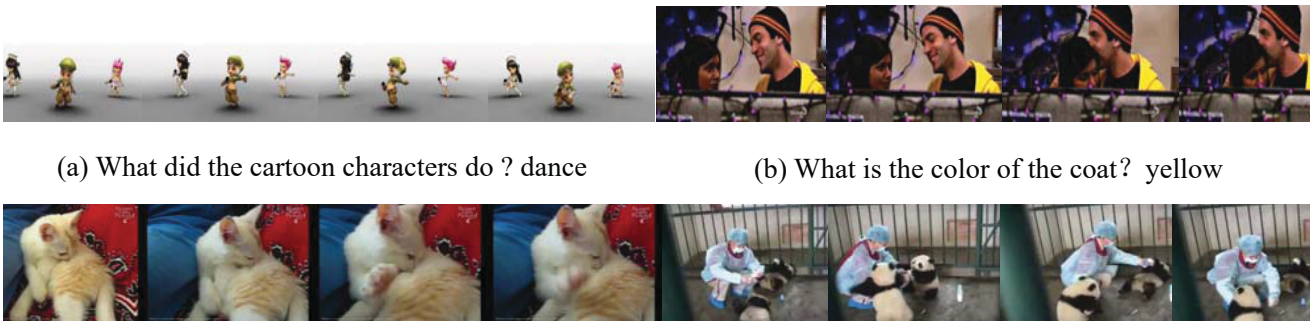
4 RELATED WORK

In this section, we briefly review some related work on visual question answering and video representation learning.

The existing approaches for visual question answering can be categorized into image-based question answering methods [2, 15, 18, 20, 21, 27, 36, 42, 44] and video-based question answering ones [22, 32, 45, 51, 53]. Given an image and a natural language question for the image, the task of image-based question answering is to provide the accurate answer for the given question [2], which indeed is a special type of content-based image retrieval[48]. Malinowski et. al. [21] develop the multi-world probabilistic approach for open-ended image question answering. With the development of attention mechanism, Shih et. al. [27] propose the spatial-attention mechanism that selects the relevant image regions to the given question. Lu et. al. [20] devise the co-attention mechanism. Ye et. al. [44] used the attributed-augmented attention mechanism and Yang et. al. [42] develop the stacked attention method for image question answering. To exploit the complex image question answering task, QRU method [18] is proposed with reasoning process that iteratively selects the relevant image regions and updates the question representation. Xiong et. al. [38] propose the dynamic memory networks and Zhao et. al. [49, 52] propose graph-based methods for both image and textual question answering. Kim et. al. [15] employ the multimodal residual network for image question answering. Johnson et. al. [14] study the composition image question answering. A survey of existing image question answering methods can be found in [36].

As a natural extension of image-based question answering, the video-based question answering has been proposed as a more challenging task [45]. The fill-in-the-blank approaches [22, 53] complete the missing entry in the video description by ranking candidate answers based on both visual content and contextual video description. Tapaswi et. al. [32] 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. Hong et. al. [13] study the multimedia question answering. Unlike the previous studies, we study the problem of video question answering only based on the visual contents.

Video representation learning is important for understanding its evolving complex object interactions, which has attracted considerable attentions recently [3, 24, 30, 33, 37, 47]. Tran et. al. [33] propose to learn spatio-temporal video features using deep 3D ConvNet with $3 \times 3 \times 3$ convolution kernels. Ballas et. al. [3] introduce a recurrent convolutional network architecture with different spatial resolutions for video representation learning. Wu et. al. [37] introduce the DNN framework that explores both inter-feature and inter-class relationships to achieve video classification. Pan et. al. [24] propose hierarchical recurrent neural encoder that exploits temporal information for video representation learning. Xu et. al. [40] propose the CNN network with latent concept descriptor for learning video representation. Acar et. al. [1] learn the affective video representation with audio modality. Zhao et. al. [50] study a multi-modal sparse coding representation for multi-modal data like videos. Zhang et. al. [47] study a binary video representation for self-supervised temporal hashing. Cui et. al. [4] propose to learn the video recommendation from its content attributes for recommendation. Xu et. al. [39] propose a deep shared video representation learning architecture for multimodal fusion of multi-timescale temporal data with music and video modalities. Simonyan et. al. [28] propose a two-stream ConvNet architecture which incorporates spatial and temporal networks for video action recognition. Feichtenhofer et. al. [5] present the spatiotemporal ResNet architecture for video-based feature representation. Wang et. al. [35] propose the video representation learning method based on Siamese-triplet network supervised by visual tracking information. Srivastava et. al. [30] develop the sequence-to-sequence learning framework for video representation learning. Zhang et. al [46] use an embedding network to capture the relation between visual objects under certain language descriptions. Unlike the previous



(a) What did the cartoon characters do ? dance

(b) What is the color of the coat? yellow

(c) What is a cat waving ? paw

(d) How many baby pandas is a person feeding ? two

Figure 7: The Examples of Video Question Answering Results.

Example	Visualization of Temporal Attention	Fusion Attention	
		Frame-Level	Segment-Level
(a)		0.0405	0.9595
(b)		0.7308	0.2692
(c)		0.0942	0.9058
(d)		0.8459	0.1541

Figure 8: The Attentional Results of Hierarchial Dual-Level Attention Network.

studies, we study the question-aware joint video representation learning with hierarchical dual-level attention network based on both frame-level and segment-level features for video question answering.

5 CONCLUSIONS

In this paper, we study the problem of video question answering from the viewpoint of hierarchical dual-level attention network learning. We first propose the frame-level and segment-level video feature representation methods to obtain both the object appearance information and its movement information in videos. We then employ the hierarchical dual-level attention networks to learn the question-aware video representations with word-level and question-level attention mechanisms. We next incorporate the question-level fusion attention mechanism for our proposed network to learn the question-aware joint video representation for video question answering. We construct two large-scale video question answering datasets and evaluate the effectiveness of our proposed method through extensive experiments.

ACKNOWLEDGEMENTS

This work was supported by National Basic Research Program of China (973 Program) under Grant 2013CB336500, and National Natural Science Foundation of China under Grant 61602405 and Grant U1611461, Fundamental Research Funds for the Central Universities 2016QNA5015 and the China Knowledge Centre for Engineering Sciences and Technology. The Project is also Supported by the Key Laboratory of Advanced Information Science and Network Technology of Beijing (XDXX1603).

REFERENCES

[1] E. Acar. Learning representations for affective video understanding. In *ACM Multimedia*, pages 1055–1058. ACM, 2013.
[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.
[3] N. Ballas, L. Yao, C. Pal, and A. Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015.
[4] P. Cui, Z. Wang, and Z. Su. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *ACM Multimedia*, pages 597–606. ACM, 2014.
[5] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In *NIPS*, pages 3468–3476, 2016.
[6] C. Gan, Y. Yang, L. Zhu, D. Zhao, and Y. Zhuang. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision*, 120(1):61–77, 2016.
[7] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, pages 2712–2719, 2013.
[8] A. Gupta and R. Jain. Visual information retrieval. *Communications of the ACM*, 40(5):70–79, 1997.
[9] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017.
[10] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 549–558. ACM, 2016.

[11] M. Heilman and N. 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.

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

[13] R. Hong, M. Wang, G. Li, L. Nie, Z.-J. Zha, and T.-S. Chua. Multimedia question answering. *IEEE MultiMedia*, 19(4):72–78, 2012.

[14] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, 2016.

[15] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. *NIPS*, 2016.

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

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

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

[19] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*, pages 4641–4650, 2016.

[20] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297, 2016.

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

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

[23] C.-A. Palma, J. Björk, F. Klappenberger, E. Arras, D. Kühne, S. Stafström, and J. V. Barth. Visualization and thermodynamic encoding of single-molecule partition function projections. *Nature communications*, 6, 2015.

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

[25] F. Shen, Y. Mu, Y. Yang, W. Liu, L. Liu, J. Song, and H. T. Shen. Classification by retrieval: Binarizing data and classifier. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2017.

[26] F. Shen, Y. Yang, L. Liu, W. Liu, D. Tao, and H. T. Shen. Asymmetric binary coding for image search. *IEEE Transactions on Multimedia*, 2017.

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

[28] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

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

[30] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, pages 843–852, 2015.

[31] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In *NIPS*, pages 2440–2448, 2015.

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

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

[34] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *ICCV*, pages 4534–4542, 2015.

[35] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, pages 2794–2802, 2015.

[36] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel. Visual question answering: A survey of methods and datasets. *arXiv preprint arXiv:1607.05910*, 2016.

[37] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *ACM Multimedia*, pages 167–176. ACM, 2014.

[38] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. *ICML*, 1603, 2016.

[39] B. Xu, X. Wang, and X. Tang. Fusing music and video modalities using multi-timescale shared representations. In *ACM Multimedia*, pages 1073–1076. ACM, 2014.

[40] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. In *CVPR*, pages 1798–1807, 2015.

[41] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu. Image classification by cross-media active learning with privileged information. *IEEE Transactions on Multimedia*, 18(12):2494–2502, 2016.

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

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

[44] Y. Ye, Z. Zhao, Y. Li, J. Xiao, and Z. Yueting. Video question answering via attributed-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2017.

[45] 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. *arXiv preprint arXiv:1611.04021*, 2016.

[46] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017.

[47] H. Zhang, M. Wang, R. Hong, and T.-S. Chua. Play and rewind: Optimizing binary representations of videos by self-supervised temporal hashing. In *ACM Multimedia*, pages 781–790. ACM, 2016.

[48] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 33–42. ACM, 2013.

[49] Z. Zhao, X. He, D. Cai, L. Zhang, W. Ng, and Y. Zhuang. Graph regularized feature selection with data reconstruction. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):689–700, 2016.

[50] Z. Zhao, H. Lu, D. Cai, X. He, and Y. Zhuang. Partial multi-modal sparse coding via adaptive similarity structure regularization. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 152–156. ACM, 2016.

[51] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, 2017.

[52] Z. Zhao, L. Zhang, X. He, and W. Ng. Expert finding for question answering via graph regularized matrix completion. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):993–1004, 2015.

[53] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann. Uncovering temporal context for video question and answering. *International Journal of Computer Vision*, 2017.