Video Dialog via Multi-Grained Convolutional Self-Attention Context Networks

Weike Jin
Zhejiang University, Hangzhou
weikejin@zju.edu.cn

Zhou Zhao*
Zhejiang University, Hangzhou
zhaozhou@zju.edu.cn

Mao Gu
Zhejiang University, Hangzhou
21821134@zju.edu.cn

Jun Yu
Hangzhou Dianzi University
yujun@hdu.edu.cn

Jun Xiao
Zhejiang University, Hangzhou
junx@cs.zju.edu.cn

Yueting Zhuang
Zhejiang University, Hangzhou
yzhuang@zju.edu.cn

ABSTRACT

Video dialog is a new and challenging task, which requires an AI agent to maintain a meaningful dialog with humans in natural language about video contents. Specifically, given a video, a dialog history and a new question about the video, the agent has to combine video information with dialog history to infer the answer. And due to the complexity of video information, the methods of image dialog might be ineffectively applied directly to video dialog. In this paper, we propose a novel approach for video dialog called multi-grained convolutional self-attention context network, which combines video information with dialog history. Instead of using RNN to encode the sequence information, we design a multi-grained convolutional self-attention mechanism to capture both element and segment level interactions which contain multi-grained sequence information. Then, we design a hierarchical dialog history encoder to learn the context-aware question representation and a two-stream video encoder to learn the context-aware video representation. We evaluate our method on two large-scale datasets. Due to the flexibility and parallelism of the new attention mechanism, our method can achieve higher time efficiency, and the extensive experiments also show the effectiveness of our method.

CCS CONCEPTS

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

KEYWORDS

video dialog, multi-grained self-attention, convolution

ACM Reference Format:
Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. 2019. Video Dialog via Multi-Grained Convolutional Self-Attention Context Networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331240

video dialog history

Person A	What is the boy pouring the chocolate shake into?
Person B	Plastic cup.
Person A	Where is the boy sitting?
Person B	A seat in the garden.
Person A	What is the boy doing after drinking?
Person B	He is puking.

new dialog

Person A	What is the boy pouring into the ground?
Person B	Chocolate shake.

Figure 1: A simple example of video dialog.

1 INTRODUCTION

Visual dialog can be seen as an extension of the visual question answering (VQA) problem [1, 27, 43], in which an agent is required to maintain a meaningful dialog with humans in natural language about visual contents. Different from visual question answering, where each question is asked independently, visual dialog requires the agent to answer questions that might be related to the previous dialog history. Currently, most of the existing visual dialog approaches mainly focus on the image [4, 5, 16, 17, 31]. However, video is also a kind of common visual information in our daily life [13, 38]. Thus, we extend the task of image dialog to the video domain, called video dialog. Specifically, given a video, a dialog history and a new question about the video content, the agent has to combine video information with context information from dialog history to infer the answer. Due to the complexity of video information, it's more challenging than image dialog.

Because of the inherent temporal structure of the video [36, 38], current approaches of image dialog might not be applied directly to video dialog, which lack the temporal modeling abilities. The models of video question answering [12, 26, 42] have this kind of abilities, however, they are still unsuitable to utilize directly, for the insufficiency of modeling the dialog history context. The sequential and interdependent properties of the dialog history present an additional challenge. As shown in Figure 1, in order to answer the new

[*] Zhou Zhao is the corresponding author.

SIGIR '19, July 21–25, 2019, Paris, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331240

question "what is the boy pouring into the ground?", the previous dialog context is required. Without an accurate understanding of the dialog context, it's hard for the system to figure out what kind of liquid in the cup only by the visual information. In a word, a simple extension of existing methods is difficult to provide a satisfactory result. So, a new model needs to be developed for video dialog. As we know, recurrent neural network (RNN) is widely used for its capability in capturing sequence information through recurrent computation. However, basic RNN may encounter gradient vanish problem and it's difficult to parallelize. Some improvements are made by several variants, such as LSTM [9], GRU [3] and SRU [18]. Some work also employs convolutional neural network (CNN) due to its parallelizable convolution computation. And recently, self-attention mechanism [37] has attracted great interests in the field of sequence modeling. It utilizes attention mechanism to each pair of elements from the input sequences to generate context-aware sequence representations. Because the main operation of self-attention is matrix multiplication, it's highly parallelizable and flexible to compute without any recurrent iteration.

In this paper, in order to tackle the challenges of video dialog, we propose a novel approach called multi-grained convolutional self-attention context network, which combines the video information with the context information from dialog history. On the one hand, we find both of the video and dialog data are quite variable in sequence length and the length of the video is usually quite long. Due to these characteristics, we think that RNN may not be the best choice in this case, which is more suitable to process fixed-length information due to the inherent network structure. On the other hand, the dynamic attributes (like action and state transition) of video are usually included in several frame segments instead of a single frame. However, the original self-attention mechanism only considers element-wise interaction. Thus, instead of using general RNN and self-attention directly, we design a multi-grained convolutional self-attention mechanism that could capture both element and segment level interactions, which contain multi-grained sequence information. By using this new attention mechanism, we design a hierarchical dialog history encoder to learn the context-aware question representation and a two-stream video encoder to learn the context-aware video representation. Then, we fuse the visual and textual information to answer the question in a new round of dialog. We name the overall network as MGCSACN. And the main contributions of this paper are as follows:

- Unlike the previous studies that mainly focus on static images, we extend the task of image dialog to the video domain, called video dialog, which is more challenging. And we propose a novel approach called multi-grained convolutional self-attention context network to learn joint representations of visual and textual information.
- We develop a multi-grained convolutional self-attention mechanism that could capture both element and segment level interactions, which contain multi-grained sequence information. We utilize this novel attention mechanism in both visual and textual sequence encoding processes, where RNN is usually applied.
- Our method achieves the state-of-the-art performance on two large-scale datasets. And it is also more efficient in time cost than other RNN based methods, which benefits from the flexibility and parallelism of our fully self-attention and CNN based network structure.

2 RELATED WORK

Visual dialog can be seen as an extension of visual question answering (VQA), which extends single-turn question answering to multi-turn dialog. Thus, in this section, we first introduce some related work of visual question answering. Then, we review some existing approaches of visual dialog.

Visual Question Answering. For image question answering, an early work is proposed by Malinowski et al. [22], which combines semantic parsing with image segmentation through a Bayesian approach. As deep learning showing great effectiveness in both computer vision and natural language processing, many neural network based approaches have been proposed [14, 23, 28]. Zhu et al. [48] add spatial attention to the standard LSTM model. Yu et al. [43] develop a semantic attention mechanism to select high-level question-related concepts. Nguyen et al. [27] propose a dense co-attention network, which employs dense symmetric interactions between the input image and question to improve the features fusion. Compared to image question answering, video question answering is relatively less explored. Jang et al. [12] propose a dual-LSTM based approach with both spatial and temporal attention. Na et al. [26] develop a read-write memory network to fuse multi-modal features.And recently, Gao et al. [8] propose a co-memory network to capture the further interaction of motion and appearance information. Liang et al. [19] propose a novel neural network called Focal Visual-Text Attention network (FVTA) for collective reasoning in visual question answering.

Visual Dialog. Visual dialog can be seen as an extension of visual question answering (VQA), which extends single-turn question answering to multi-turn dialog. Specifically, the question-answer pair in each turn may refer to the information from previous context of the dialog. As proposed in [4], visual dialog requires to predict the answer for a given question based on the image and the dialog history. While dialog system [32–34, 39, 41] has been widely explored, visual dialog is still a young task. Until recently, some different approaches are proposed. For instance, Das et al. [4] propose three models based on late fusion, attention based hierarchical LSTM, and memory networks respectively. They also propose the VisDial dataset by pairing two subjects on Amazon Mechanical Turk to chat about an image (ask questions to imagine the hidden image better). Lu et al. [21] propose a generator-discriminator architecture where the outputs of the generator are improved using a perceptual loss from a pre-trained discriminator. De Vries et al. [7] propose a GuessWhat game style dataset, where one person asks questions about an image to guess which object has been selected, and the second person answers questions. Das et al. [5] propose to use deep reinforcement learning to learn the policies of a 'Questioner-Bot' and an 'Answerer-Bot', based on the goal of selecting the right images that the two agents are talking. Seo et al. [31] resolve visual references in dialog questions based on a new attention mechanism with an attention memory, where the model indirectly resolves coreferences of expressions through the attention retrieval process. Kottur et al. [16] propose the introspective reasoning about visual coreferences, which explicitly links coreferences and grounds them in the image at a word-level, rather than implicitly or at a sentence-level, as in prior visual dialog work. Massiceti et al. [24] propose FLIPDIAL, a generative convolutional model for visual dialogue which is able to generate answers as well as generate both questions and answers based on a visual context. Lee et al. [17] propose "Answerer in Questioner's Mind" (AQM), a practical goal-oriented dialog framework using information-theoretic approach, in which, the questioner figures out the answerer's intention via selecting a plausible question by explicitly calculating the information gain of the candidate intentions and possible answers to each question. Wu et al. [40] propose a sequential co-attention generative model that can jointly reason the image, dialog history with question, and a discriminator which can dynamically access to the attention memories with an intermediate reward, and it achieves the state-of-the-art on VisDial dataset.

The aforementioned work mainly focuses on the image dialog. As for the task of video dialog, it's still less explored. One similar work is proposed by Zhao et al. [45]. They study the problem of multi-turn video question answering by employing a hierarchical attention context learning method with recurrent neural networks for context-aware question understanding and a multi-stream attention network that learns the joint embedding video representation. They also propose two large-scale multi-turn video question answering datasets from YouTubeClips [2] and TACoSMultiLevel [30]. And recently, Hori et al. [10] propose a model that incorporates technologies for multimodal attention-based video description into an end-to-end dialog system. Unlike these work, we utilize a more effective and efficient convolutional self-attention context network for video dialog task.

3 PROPOSED APPROACH

In this section, we first formulate the problem of video dialog. Then we introduce our convolutional self-attention context network for video dialog in detail.

3.1 Problem Formulation

For video dialog is still less explored, here we first introduce some basic notions and terminologies. We denote the video by $\mathbf{v} \in V$, the dialog history by $\mathbf{c} \in C$, the new question by $\mathbf{q} \in Q$ and the corresponding answer by $\mathbf{a} \in A$, respectively. For video is a sequence of static frames, the frame-level representation for video $\mathbf{v}$ is denoted by $\mathbf{v}^f = (v_1^f, v_2^f, \ldots, v_{T_1}^f)$, where T_1 is the number of frames in video $\mathbf{v}$. And the segment-level representation of video $\mathbf{v}$ is given by $\mathbf{v}^s = (v_1^s, v_2^s, \ldots, v_{T_2}^s)$, where T_2 is the number of segments and v_j^s is the representation of the j-th segment learned by pre-trained 3D-ConvNet [36]. Then the dialog history $\mathbf{c} \in C$ is given by $\mathbf{c} = (c_1, c_2, \ldots, c_N)$, where c_i is the i-th round conversation, which consists of question q_i and answer a_i. Using these notations, the task of video dialog could be formulated as follows. Given a set of video V and the associated dialog history C, the goal of video dialog task is to train a model that learns to generate human-like answers when a new question about the visual content is asked. Similar to the video question answering task, there are two types of models to produce the answer, generative and discriminative. For the generative decoder, a word sequence generator (normally a RNN) is employed to fit the ground truth answer sequences. As for discriminative decoder, an additional candidate answer vocabulary is provided and the problem is reformulated as a multi-class classification problem.

3.2 Multi-Grained Convolution Self-Attention Context Network

In this section, we introduce our multi-grained convolutional self-attention context network for video dialog, which follows the encoder-decoder network structure. As shown in Figure 2, the whole model consists of a dialog history encoder, a context-aware video encoder and an answer decoder. In the following, we will describe the detail of them individually.

3.2.1 Multi-Grained Convolutional Self-Attention. Before introducing the main encoder-decoder parts, we first describe our proposed multi-grained convolutional self-attention (MGCSA) mechanism, which is utilized to capture multi-grained interaction information from the input sequence. It's used in both of our dialog history and video information encoding processes. Similar to the previous work [37], we employ the sine-cosine position encoding mechanism to add temporal information of the sequence to the self-attention process and we will ignore it for simplicity in the following description. As shown in Figure 2(c), a MGCSA unit accepts a sequence of word embeddings or video frame features as input, denoted as $X = (x_1, x_2, \ldots, x_n)$. Firstly, we split the input sequence into k segments of equal length l, given by $X = (X^1, X^2, \ldots, X^k)$, where $X^1 = (x_1, x_2, \ldots, x_l)$, $X^2 = (x_{l+1}, x_{l+2}, \ldots, x_{2l}), \ldots$ and $n = k \times l$. For an intuitive understanding, we show different segments in different grayscale. And if the input sequence cannot be split equally, the sequence will be padded to satisfy the condition. Then, we apply self-attention mechanism to each segment, in order to capture local interaction information inside each segment, given by

$$Y^i = \text{Attention}(X^i, X^i, X^i), i = 1, 2, \ldots, k \tag{1}$$

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d}})V \tag{2}$$

where d is the dimension of sequence elements and Y^i is the new i-th segment representation after self-attention. Now we obtain a new sequence $Y = (Y^1, Y^2, \ldots, Y^k)$. After that, we employ a convolutional layer with both kernel size and step size equal to segment length l. Through this convolutional layer, we get a squeezed sequence $P = (p_1, p_2, \ldots, p_k)$, which contains k elements and each element can be seen as a vector representation of its original segment. Thus, we could obtain a segment-wise interaction by employing self-attention operation on sequence P. The new sequence after self-attention is denoted as $P' = (p'_1, p'_2, \ldots, p'_k)$. Then, a fusion operation is utilized to merge the local-interacted feature sequence P and global-interacted feature sequence P'. The output sequence Z' is computed by

$$\begin{aligned} S &= \sigma(W_g^1[P; P'] + b_g^1), && (3)\\ S' &= \sigma(W_g^2[P; P'] + b_g^2), && (4)\\ Z' &= S' \odot P' + S \odot P && (5) \end{aligned}$$

where σ is sigmoid function, $[;]$ is concatenation of vectors, $\odot$ is element-wise multiplication, b_g^1, b_g^2 are bias vectors and S, S' are

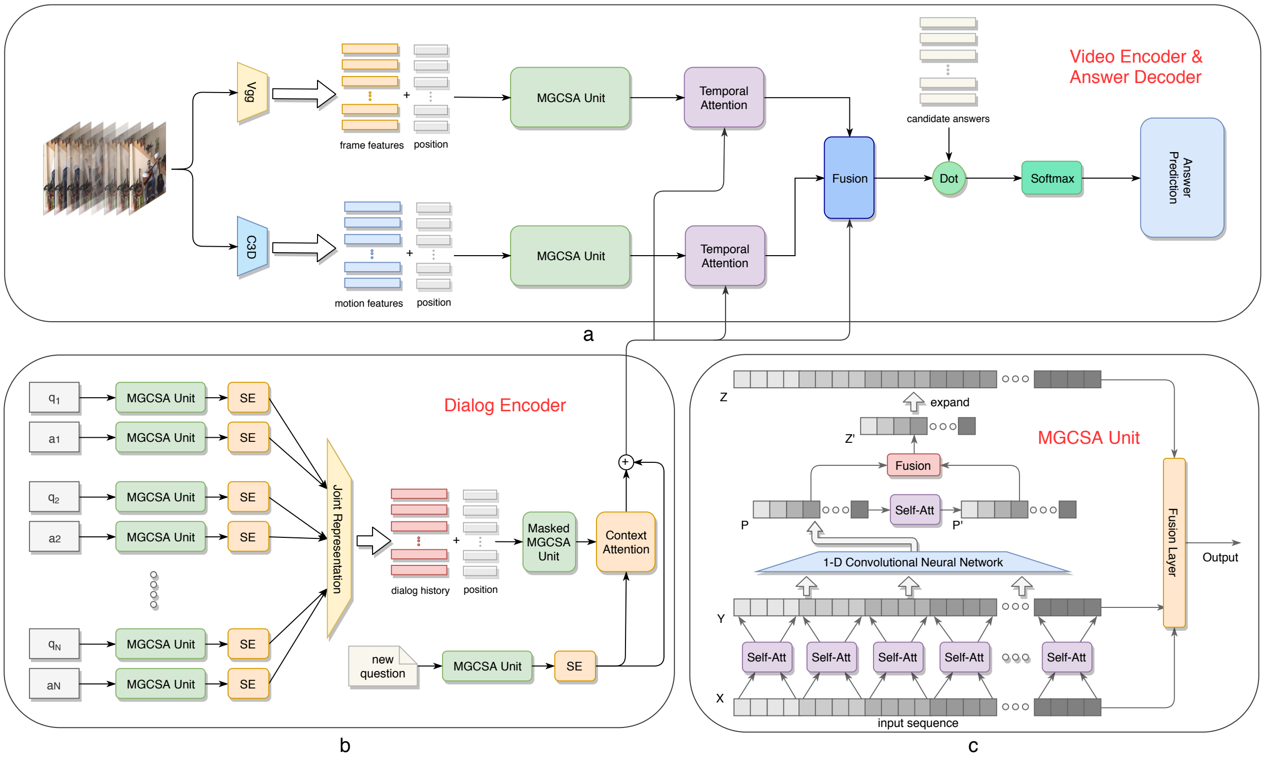


Figure 2: The Overview of Multi-Grained Convolutional Self-Attention Context Networks for Video Dialog. It consists of three parts. (a) The video encoder and answer decoder, in which we first learn the context-aware joint video representation based on multi-grained convolutional self-attention and temporal attention mechanism and then fuse multi-modal information to predict the final answer. (b) The dialog encoder, in which we encode the dialog history and the new question by utilizing same attention mechanisms. (c) The MGCSA unit, in which we show the multi-grained convolutional self-attention mechanism in detail.

score vectors of the gates whose values are between 0 and 1. Given the new sequence $Z' = (z'_1, z'_2, \ldots, z'_k)$, we duplicate each z'_i for l times to get a new sequence $Z = (z_1, z_2, \ldots, z_n)$. Finally, we employ the same feature fusion operation to combine the input sequence X, the element-level context sequence Y and the segment-level context sequence Z and generate the final multi-grained context-aware sequence representation R, given by

$$F_{yz} = Fusion(Y, Z), \tag{6}$$

$$R = Fusion(F_{yz}, X) \tag{7}$$

where $Fusion()$ includes the same operations as equations (3,4,5), with different weight parameters.

3.2.2 Dialog History Encoder. In this section, we introduce the hierarchical dialog history encoder in detail, which learns the coherent question representation with dialog history. As shown in Figure 2(b), given a dialog history $\mathbf{c} = (c_1, c_2, \ldots, c_N)$ where c_i is the i-th round conversation that consists of the question q_i and the answer a_i, we first employ the multi-grained convolutional self-attention (MGCSA) to learn the representation of the question q_i and the answer a_i in the i-th round conversation. Because the output of the MGCSA unit is still a sequence of word embeddings, we use a sentence embedding (SE) mechanism similar to [20] to get the sentence representation. The following equations are used to compute the output of sentence embedding.

$$f(x_i) = \text{softmax}(W_1 \tanh(W_2 x_i + b_1)), \tag{8}$$

$$O = \sum_{i=1}^{n} f(x_i) \odot x_i \tag{9}$$

where W_1, W_2 are weight parameters, b_1 is the bias, n is the length of the input sequence and x_i is one element of the input. For i-th round of dialog history, we obtain the sentence-level representation of question r_i^q and answer r_i^a after sentence embedding. Then, we perform the joint representation mechanism [47] on this question-answer pair to learn the representation of the i-th round dialog history c_i, given by

$$c_i = \tanh(W_c^1 r_i^q + W_c^2 r_i^a) \tag{10}$$

where $W_c^1 \in \mathbb{R}^{d_c \times d}$ and $W_c^2 \in \mathbb{R}^{d_c \times d}$ are projection matrixes used for the fusion of question and answer representations. The d is the dimension of r_i^q and r_i^a, and d_c is the dimension of joint representations. The tanh is element-wise hyperbolic tangent function, which performs well in multi-modal representation fusion. Different from single-turn question answering, the context of the video dialog is related. Thus, we utilize a masked multi-grained convolutional self-attention to encode the context interaction. Here, the mask is used to avoid the current round of dialog seeing the later round of dialog, which conforms to our common sense. The interacted dialog context representations based on the joint question-answer representations $\mathbf{c} = (c_1, c_2, \ldots, c_N)$ are denoted by $\mathbf{u} = (u_1, u_2, \ldots, u_N)$, where $u_i \in \mathbb{R}^{d_c}$.

Next, given a new question, a same process as previous round of dialog is employed to learn the question representation q. We use it to further filter out the contextual information related to the new question by attention mechanism. The attention score s_i^{qu} is given by

$$s_i^{qu} = w_{qu}^{\top} tanh(W_{qu}^1 q + W_{qu}^2 u_i + b_{qu}) \tag{11}$$

where the $W_{qu}^1 \in \mathbb{R}^{d_m \times d}$, $W_{qu}^2 \in \mathbb{R}^{d_m \times d}$ are parameter matrices and the $w_{qu}^{\top} \in \mathbb{R}^{d_m}$ is parameter vector. The $b_{qu} \in \mathbb{R}^{d_m}$ is bias vector and the d_m is the middle dimension. We then apply softmax function to generate the attention distribution over dialog context, which is given by

$$\alpha_i^{qu} = \frac{\exp(s_i^{qu})}{\sum_i^N \exp(s_i^{qu})} \tag{12}$$

Thus, the attended dialog context representation is given by $u^q = \sum_i^N \alpha_i^{qu} u_i$. And the final context-aware question representation is given by

$$q^u = q + u^q \tag{13}$$

3.2.3 Context-Aware Video Encoder. In this section, we describe the part of video encoder, which learns context-aware video representations for answer prediction. As we know that video contains rich visual information, such as appearance, objects, motions, etc. A lot of work [8, 12] has proven that it's necessary to take such information into account for high-quality video understanding. Here, we take the appearance and motion information into consideration to capture both frame-level and segment-level video representations. Specifically, we utilize a pre-trained VGGNet [35] to extract the appearance features and a 3D-ConvNet [36] to obtain the motion features. As denoted in Section 3.1, the appearance sequence is $\mathbf{v}^f = (v_1^f, v_2^f, \ldots, v_{T_1}^f)$ and the motion sequence is $\mathbf{v}^s = (v_1^s, v_2^s, \ldots, v_{T_2}^s)$.

First, we also utilize the proposed multi-grained convolutional self-attention mechanism to learn the multi-grained representations of both appearance information and motion information, denoted as $\mathbf{v'}^f = ({v'}_1^f, {v'}_2^f, \ldots, {v'}_{T_1}^f)$ and $\mathbf{v'}^s = ({v'}_1^s, {v'}_2^s, \ldots, {v'}_{T_2}^s)$, separately. Then, given the context-aware question representation q^u, we develop a temporal attention process to localize the relevant frames or segments with the targeted information according to the question and dialog history context, due to the abundant redundancy of video information. For the i-th frame ${v'}_i^f$, its temporal attention score s_i^{qf} is given by

$$s_i^{qf} = w_{qf}^{\top} \tanh(W_{qf}^1 q^u + W_{qf}^2 {v'}_i^f + b_{qf}) \tag{14}$$

where the $W_{qf}^1 \in \mathbb{R}^{d_n \times d}$, $W_{qf}^2 \in \mathbb{R}^{d_n \times d_f}$ are parameter matrices and the $w_{qf}^{\top} \in \mathbb{R}^{d_n}$ is parameter vector. The $b_{qf} \in \mathbb{R}^{d_n}$ is bias vector. The d_n is the middle dimension and d_f is the appearance feature dimension. Softmax function is still used to generate the attention distribution over video frames, which is given by

$$\alpha_i^{qf} = \frac{\exp(s_i^{qf})}{\sum_i^{T_1} \exp(s_i^{qf})} \tag{15}$$

Thus, the context-aware video appearance representation is given by

$$v^{qf} = \sum_i^{T_1} \alpha_i^{qf} {v'}_i^f \tag{16}$$

On the other hand, the same operation is applied to the video motion information. We could obtain the context-aware video motion representation v^{qs}. Therefore, the final context-aware video representation is calculated by

$$v_q^{fs} = v^{qf} \odot v^{qs} \tag{17}$$

where $\odot$ is the element-wise product operator.

Finally, we fuse the context-aware video representation v_q^{fs} with the context-aware question representation q^u for the following question prediction, given by

$$f_{quv} = \text{Concat}(g(v_q^{fs}), g(q^u)) \tag{18}$$

where Concat$(\cdot)$ is a function that concatenates the two input vectors and $g(\cdot)$ is the gated hyperbolic tangent activations [6].

3.2.4 Answer Decoder. Here, we model the problem of video dialog as a classification task with pre-defined candidate answer sets following the existing visual question answering models [1, 45]. Given the final context-aware video-textual fusion representation f_{quv} defined in the above section, we now develop a multiple-choice method for the task of video dialog. We first learn the semantic representation a_i^c of each candidate answer in the answer sets by employing another multi-grained convolutional self-attention unit. Then, we could obtain the answer representation matrix $A = [a_1^c; a_2^c; \cdots; a_{T3}^c] \in \mathbb{R}^{T3 \times d_h}$ for an answer set, where $T3$ is the number of candidate answers and the dimension d_h of the answer representation is as same as the final fusion representation f_{quv}. Finally, we calculate the similarity between the final fusion representation $f_{quv} \in \mathbb{R}^{d_h}$ and each candidate answer, and a softmax function is employed to classify the possible answers as

$$p_a = \text{softmax}(A \times f_{quv}^{\top}) \tag{19}$$

where $p_a \in \mathbb{R}^{T3}$ is the probability distribution for candidate answers. And it has to be noted that video dialog can also be seen as a generation task, which will be discussed in our future work. For instance, instead of using a softmax function for answer classification, it's also possible to utilize a LSTM answer generator to generate free-form natural language answers by taking the final fusion representation f_{quv} as input.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Datasets. The datasets that we use here are proposed in [45], which are extracted from YouTubeClips[2] and TACoS-MultiLevel[30]. These two datasets consist of 1,987 and 1,303 videos individually.

Each YouTubeClips video is composed of 60 frames and each TACoS-MultiLevel video consists of 80 frames. In both datasets, each video has five different dialogs generated by five pairs of crowdsourcing workers from the professional company. There are 6515 video dialogs in YouTubeClips dataset and 9935 video dialogs in TACoS-MultiLevel dataset. And the numbers of dialog question-answer pairs are 66,806 and 37,228 in YouTubeClips dataset and TACoS-MultiLevel dataset correspondingly. Statistically, there are five turns of conversation in most of the video dialogs in TACoS-MultiLevel dataset and the number of question-answer pairs for each video dialog in YouTubeClips dataset is mostly between three and twelve. The percentages of training data, validation data and testing data for both datasets are 90%, 5%, and 5% respectively according to the number of constructed video dialogs.

Additionally, the semantic similarities between its ground-truth answer of each video dialog and all other answers based on the Euclidean distance with the pre-trained glove[29] embedding are calculated to rank the top 50 answers as candidate answers for each video dialog.

4.1.2 Evaluation Metrics. We choose three generally employed evaluation criteria MRR, P@K, and MeanRank to evaluate the performance of the proposed MGCSACN method and other baseline models. Given a new question $q \in Q$ with its ground-truth answer a^t, we denote the rank of the ground-truth answer for question q by $r^q_{a^t}$. We now introduce the evaluation criteria below.

- **MRR.** The MRR measures the ranking quality for the ground-truth answer, which is given by

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{r^q_{a^t}},$$

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

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

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

$$MeanRank = \frac{\sum_{q \in Q} r^q_{a^t}}{|Q|}.$$

4.1.3 Implementation. The pre-process of constructing the video dialog system is shown as follows. Firstly, we used the pre-trained word2vec model[25] to obtain the semantic representations of the dialogs. The dimension of the word vector is 100 and the size of the total vocabulary table is 6,500. Secondly, we resize each frame to 224 × 224 and employ the pre-trained VGGNet[35] to get the feature representation of each frame. At the same time, 4,096-dimensional motion feature representations of the video are extracted by the pre-trained 3D-ConvNet [36]. Specifically, each motion segment contains 16 frames and overlaps 8 frames with adjacent segments for both datasets.

In MGCSA units, the length of the segment is an important hyper-parameter, which influences the segment-level sequence interaction. We set the segment length to 5 for video sequences and 3 for questions and answers. In fact, the selection of the segment length is dynamic according to different kinds of input sequences. And an improper segment length will affect the performance and occupy redundant memory. On the one hand, a too small segment length may decrease the impact of segment-level interaction, which makes it no difference with element-level interaction. On the other hand, a too large segment length may influence the diversity of segment-level sequence information. We do a lot of experiments to obtain the final setting. The output dimension of MGCSA unit and the dimension of the joint representation of the question-answer pairs in the dialog history are also two essential parameters and some experiments are conducted to get the optimal condition. Specifically, we vary the MGCSA dimension from 128 to 1024, and the joint representation dimension from 64 to 512. As a result, we set both the MGCSA dimension and the joint representation dimension to 128.

As for the training process, we minimize the cross-entropy loss by using the Adam optimizer[15], where the initial learning rate is set to 0.0005 and the exponential decay rate is set to 0.8. Additionally, we apply a gradient clipping method that controls gradient norms within 1.0 to prevent the occurrence of a large gradient in the process of backpropagation. Moreover, the mini-batch strategy is utilized in the training process and we set the batch size to 32. To stop the training process automatically when the performance of the model is no longer enhanced in a certain step, we employ an early-stopping technology to prevent unnecessary consumption of time and computer resources.

4.2 Performance Comparisons

Because video dialog is a new task, we extend some existing image dialog and video question answering models (similar to [45]) as baseline models for video dialog, which are introduced in the following:

- **ESA** method[44] simply uses attention mechanisms to generate the fusion representation of the given question and the video features without using the dialog history.
- **ESA+** method extends ESA model [44] by adding a hierarchical LSTM network to model the dialog history. In this method, we fuse the given question with the dialog history to construct a more complicated joint representation.
- **STVQA+** method extends STVQA model[12] by adding a hierarchical LSTM network to model the dialog history and employing the dual-LSTM network to fuse the dialog history and video features.
- **STAN+** method extends STAN model[46] by adding a hierarchical LSTM network to model the dialog history and utilizing spatio-temporal attention with the dialog history.
- **CDMN+** method extends CDMN model[8] by adding a hierarchical LSTM network to model dialog history and use a motion-appearance co-memory network to simultaneously learn the motion and appearance features.
- **LF+, HRE+ and MN+** methods extend three models[4] by utilizing a LSTM network to encode the video information, which are based on late fusion, attention based hierarchical LSTM, and memory networks respectively.

Table 1: Experimental results on TACoS-MultiLevel dataset.

Method	MRR	P@1	P@5	MeanRank
ESA	0.411	0.298	0.515	11.964
ESA+	0.411	0.300	0.507	10.435
STVQA+	0.427	0.305	0.540	9.762
STAN+	0.452	0.319	0.594	8.401
CDMN+	0.454	0.317	0.597	8.376
LF+	0.434	0.281	0.625	6.438
HRE+	0.454	0.313	0.594	8.813
MN+	0.467	0.343	0.625	7.688
SFQIH+	0.468	0.375	0.656	6.313
HACRN$_{(w/o.m)}$	0.444	0.319	0.579	8.726
HACRN$_{(w/o.f)}$	0.452	0.324	0.583	8.622
HACRN$_{(w/o.r)}$	0.512	0.391	0.643	6.625
HACRN$_{(r)}$	0.526	0.386	0.682	**5.804**
MGCSACN	**0.542**	**0.437**	**0.717**	5.875

Table 2: Experimental results on YoutubeClip dataset.

Method	MRR	P@1	P@5	MeanRank
ESA	0.333	0.224	0.418	11.571
ESA+	0.396	0.252	0.541	8.412
STVQA+	0.411	0.266	0.578	7.284
STAN+	0.418	0.274	0.577	7.258
CDMN+	0.422	0.278	0.584	7.074
LF+	0.389	0.250	0.563	8.531
HRE+	0.413	0.281	0.594	7.688
MN+	0.422	0.313	0.563	8.594
SFQIH+	0.441	0.313	0.656	6.781
HACRN$_{(w/o.m)}$	0.443	0.283	0.635	6.149
HACRN$_{(w/o.f)}$	0.454	0.295	0.636	6.042
HACRN$_{(w/o.r)}$	0.469	0.315	0.661	5.792
HACRN$_{(r)}$	0.470	0.306	0.670	**5.496**
MGCSACN	**0.481**	**0.344**	**0.687**	6.969

- **SFQIH+** method extends SF-QIH-se model[11] by employing a LSTM network to encode the video information, which concatenates all of the input embeddings for each of the possible answer options and employs a similarity network to predict a probability distribution over the possible answers.

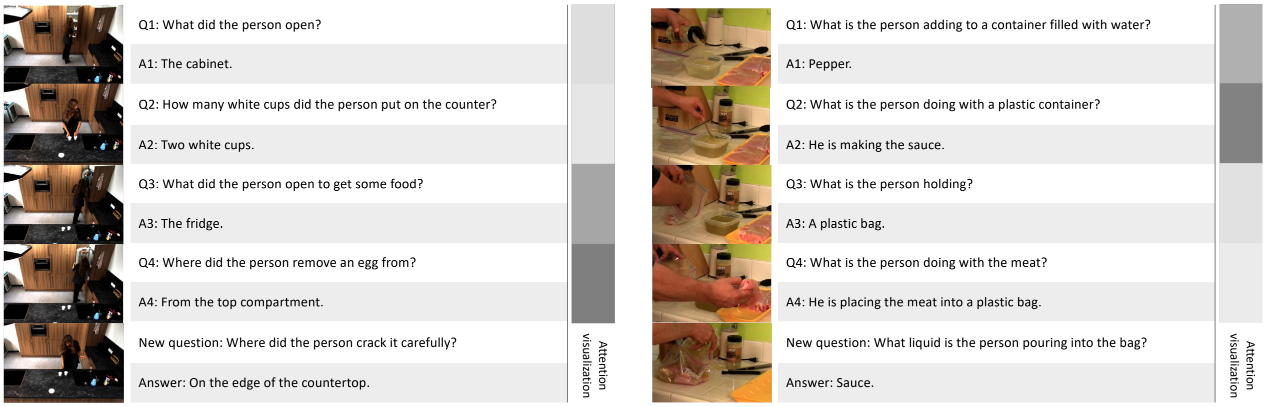
Besides the baseline models, we also compare our method with the latest work. Currently, there is little existing work about video dialog. Zhao et al. [45] propose a similar work called multi-turn video question answering. They use LSTM and attention mechanism to encode the dialog history and the given question to get the joint question representation. Then it combines this joint representation with video features by multi-stream attention network. And a multi-step reasoning strategy is applied to enhance the reasoning ability. There are some variants of this model. The method without frame channel hierarchical spatio-temporal attention context network is marked by HACRN$_{(w/o.f)}$ and the method without the motion channel temporal attention context network is marked by HACRN$_{(w/o.m)}$. Moreover, the method without multi-step reasoning process is denoted by HACRN$_{(w/o.r)}$, and the method with multi-step reasoning process is denoted by HACRN$_{(r)}$.

Unlike the above work, instead of using general RNN and attention directly, our method (MGCSACN) utilizes a multi-grained convolutional self-attention mechanism to encode both visual and textual information, which captures multi-grained sequence information. Table 1 shows the experimental results on TACoS-MultiLevel dataset and table 2 demonstrates the experimental results on YoutubeClip dataset. By analyzing these experimental results, we come into several interesting conclusions:

- The first part of the result tables which includes extended video QA models shows the impact of video information for video dialog. The models that have a better understanding of video content could achieve better performance in video dialog task.
- The second part of the result tables which includes extended image dialog models shows the significance of history dialog information for video dialog. The models that have a better understanding of dialog context perform better in video dialog task.
- Our method MGCSACN performs better than all baseline models and the latest HACRN model in almost all criteria. This fact shows the effectiveness of our fully attention and CNN based network structure, which employs a multi-grained convolutional self-attention mechanism that could capture both element and segment level interactions from the two-stream video representations and dialog history.

4.3 Qualitative Analysis

To have an intuitive understanding of our model, we analyze parts of the attention mechanisms in our method. In Figure 3, we show the visualization of the dialog context attention in dialog encoder, in which the question-answer pairs get various scores according to their correlations with the question. As shown in the left part of Figure 3, the 4-th round question-answer pair attracts more attention for it pointing out the 'it' in the new question is the 'egg'. And in the right part, it's the second round which attracts more attention. Moreover, Figure 4 shows the visualization of video attention in both frame-Level and segment-Level. As we can see, there are two important actions in the question, cracking the egg and frying. Thus, the histogram of segment-level attention shows high attention on the segments that include frying butter, cracking eggs and frying eggs, respectively. Obviously, it contains interfering information. So, we need more detailed frame-level information. And the frame-level attention successfully focuses on corresponding frames with the eggs. Combining these two attention mechanisms, the video dialog system could obtain a better understanding of the video according to the question. Besides, an example of the experimental results on YoutubeClip dataset is shown in Figure 5. Compared with the baseline models, our model performs better to get an accurate answer in this sample.



Q1: What did the person open?

A1: The cabinet.

Q2: How many white cups did the person put on the counter?

A2: Two white cups.

Q3: What did the person open to get some food?

A3: The fridge.

Q4: Where did the person remove an egg from?

A4: From the top compartment.

New question: Where did the person crack it carefully?

Answer: On the edge of the countertop.

Q1: What is the person adding to a container filled with water?

A1: Pepper.

Q2: What is the person doing with a plastic container?

A2: He is making the sauce.

Q3: What is the person holding?

A3: A plastic bag.

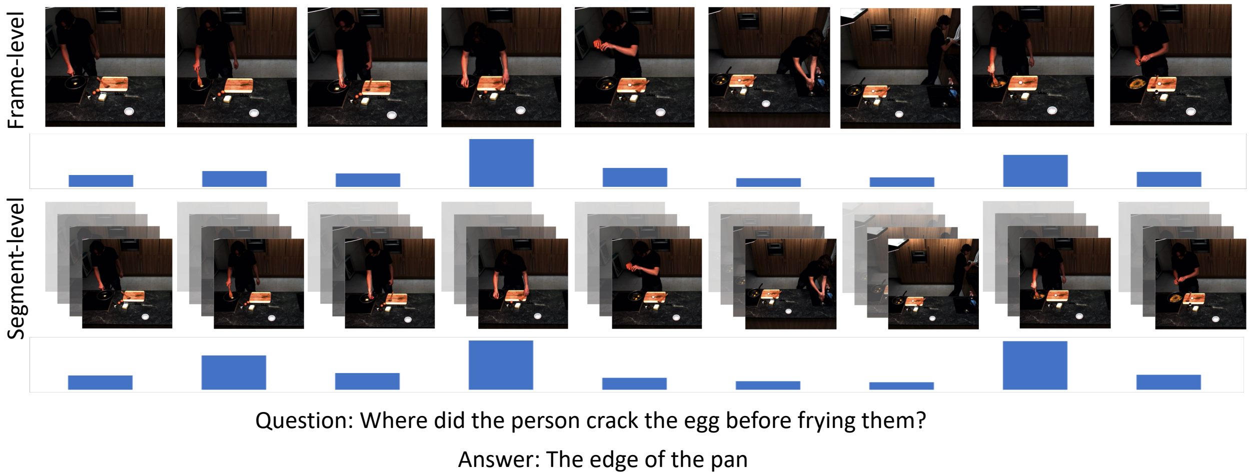
Q4: What is the person doing with the meat?

A4: He is placing the meat into a plastic bag.

New question: What liquid is the person pouring into the bag?

Answer: Sauce.

Figure 3: The visualization of dialog context attention.



Question: Where did the person crack the egg before frying them?

Answer: The edge of the pan

Figure 4: The visualization of video attention in both frame level and segment level.

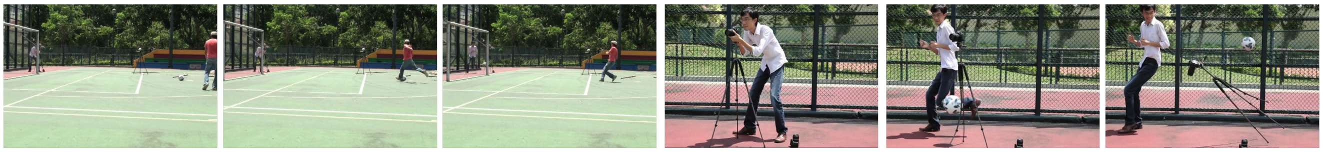
Table 3: Time cost of different methods.

Method	Training Time(s)	Inference Time(s)
HACRN$_{(w/o.r)}$	0.519	0.166
HACRN$_{GRU}$	0.485	0.169
NSACN	0.151	0.050
MGCSACN	0.153	0.089

4.4 Time Analysis

In this section, we compare the time efficiency of our method with some RNN based models. In order to efficiently control variables, we keep all the similar hyper-parameters the same, such as batch size, feature dimension. Moreover, the process of jointly fusing the given question with dialog history and video features should be guaranteed to be consistent. As for HACRN model, we choose the variant without multi-step reasoning mechanism to avoid its influence. We also replace the LSTM unit of the HACRN model with GRU unit, noted by HACRN$_{GRU}$. And the NSACN model replaces the MGCSA unit of our MGCSACN with a normal self-attention structure, which is relatively simple and only takes element-level interaction into consideration. All of these models are evaluated in the same hardware environment (E5-2678 v3, 1080ti GPU and 128GB memory).

The experimental results are shown in Table 3. The inference time per batch only contains the time of forward-propagation and the training time per batch consists of both backpropagation and forward propagation. Compared with HACRN$_{(w/o.r)}$, our method has a decrease of 70.5% in training time and a decrease of 46.4% in inference time. The HACRN$_{GRU}$ is faster than HACRN$_{(w/o.r)}$, however, it’s still much slower than our method. Although our method is not as fast as NSACN, the performance of answer prediction of our method is much better, as shown in Table 4 and 5. These facts indicate the high time efficiency of our method.



Conversation Context	Question	Answer
Q1: Where were the two persons?	What was the man in white doing before he dodged the football?	MN+: He was standing on the ground.
A1: On the playground.		SFQIH+: He was jumping away.
Q2: Who kicked a football?		STVQA+: He was jumping away
A2: The man in red.		STAN+: He was staring at the ground.
Q3: Where was the man in white?		CDMN+: He was laughing.
A3: He was behind the camera.		HACRN: He was standing behind the camera.
Q4: What knocked down the camera?		MGCSACN: He was taking photos.
A4: The football.		Ground Truth: He was taking photos.

Figure 5: An example of video dialog experimental results on YoutubeClip dataset.

Table 4: Ablation study results on TACoS-MultiLevel dataset.

Method	MRR	P@1	P@5	MeanRank
MGCSACN$_{(o.v)}$	0.473	0.344	0.625	7.531
MGCSACN$_{(o.a)}$	0.469	0.344	0.594	6.281
MGCSACN$_{(o.m)}$	0.501	0.313	0.656	7.438
NSACN	0.499	0.366	0.650	6.950
MGCSACN	**0.542**	**0.437**	**0.717**	**5.875**

Table 5: Ablation study results on YoutubeClip dataset.

Method	MRR	P@1	P@5	MeanRank
MGCSACN$_{(o.v)}$	0.365	0.250	0.500	9.125
MGCSACN$_{(o.a)}$	0.404	0.281	0.531	9.375
MGCSACN$_{(o.m)}$	0.416	0.313	0.563	8.781
NSACN	0.459	0.312	0.593	7.437
MGCSACN	**0.481**	**0.344**	**0.687**	**6.969**

4.5 Ablation Study

We perform an ablation study to evaluate the contribution of each component and analyze the optimal design options in our model. Firstly, we remove the dialog history and check the performance to illustrate its significance. Specifically, we use the normal question representation to replace the context-aware question representation. Similarly, we respectively remove the appearance features and motion features of the video to demonstrate the importance of different video information. The experimental results on TACoS-MultiLevel dataset are listed in Table 4 and the results on YoutubeClip dataset are showed in Table 5, where MGCSACN$_{(o.v)}$, MGCSACN$_{(o.m)}$, MGCSACN$_{(o.a)}$ represents the models without the participation of dialog history, appearance features and motion features correspondingly. As we can see from these results, the fact that the full model MGCSACN performs much better than all of the MGCSACN$_{(o.v)}$, MGCSACN$_{(o.m)}$ and MGCSACN$_{(o.a)}$ models proves the effectiveness of the dialog history, appearance and motion information of the video. And the fact that the MGCSACN model outperforms the NSACN model shows the effectiveness of our multi-grained convolutional self-attention mechanism.

5 CONCLUSION

In this paper, we extend the task of image dialog to video domain, called video dialog. In order to tackle this problem, we propose a novel approach called multi-grained convolutional self-attention context network, which combines attention mechanisms with convolutional neural networks to encode both of the video and dialog context. Specifically, we design a multi-grained convolutional self-attention mechanism that could capture both element and segment level interactions, which contains multi-grained sequence information. By using this new attention mechanism, we employ a hierarchical dialog history encoder to learn the context-aware question representation and a two-stream video encoder to learn the video representation. Then, we fuse the visual and textual information to generate corresponding answers for the new question. It's a fully self-attention and CNN based model so as to get rid of the limitation of general RNN networks. And due to the flexibility and parallelism of the network structure, our method could achieve high time efficiency. The extensive experiments based on two large-scale datasets also show the effectiveness of our method. And in future work, we will explore the more generative video dialog system.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China under Grant No.61602405, No.61836002, No.U1611461 and No.61751209, Zhejiang Natural Science Foundation (LR19F020002), Key R&D Program of Zhejiang Province (2018C01006), the Fundamental Research Funds for the Central Universities, Chinese Knowledge Center for Engineering Sciences and Technology, Joint Research Program of ZJU & Hikvision Research Institute and Alibaba Innovative Research.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[2] David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. 190–200.

[3] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 103–111.

[4] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1080–1089.

[5] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *International Conference on Computer Vision*. 2970–2979.

[6] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2016. Language modeling with gated convolutional networks. *arXiv preprint arXiv:1612.08083* (2016).

[7] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C Courville. 2017. GuessWhat?! Visual object discovery through multimodal dialogue.. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4466–4475.

[8] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018).

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

[10] Chiori Hori, Huda Alamri, Jue Wang, Gordon Winchern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. 2018. End-to-End Audio Visual Scene-Aware Dialog using Multimodal Attention-Based Video Features. *arXiv preprint arXiv:1806.08409* (2018).

[11] Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. 2018. Two can play this game: visual dialog with discriminative question generation and answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5754–5763.

[12] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2680–8.

[13] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[14] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*. 361–369.

[15] DP Kingma and J Ba. [n.d.]. DP Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv: 1412.6980. *Adam: A Method for Stochastic Optimization* ([n. d.]).

[16] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 153–169.

[17] Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. 2018. Answerer in Questioner's Mind: Information Theoretic Approach to Goal-Oriented Visual Dialog. In *Advances in Neural Information Processing Systems*. 2580–2590.

[18] Tao Lei, Yu Zhang, and Yoav Artzi. 2017. Training rnns as fast as cnns. *arXiv preprint arXiv:1709.02755* (2017).

[19] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander Hauptmann. 2018. Focal Visual-Text Attention for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6135–6143.

[20] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[21] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*. 314–324.

[22] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*. 1682–1690.

[23] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. 1–9.

[24] Daniela Massiceti, N Siddharth, Puneet K Dokania, and Philip HS Torr. 2018. FlipDial: A generative model for two-way visual dialogue. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6097–6105.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[26] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. 2017. A read-write memory network for movie story understanding. In *International Conference on Computer Vision*.

[27] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6087–6096.

[28] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 30–38.

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[30] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*. 184–195.

[31] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*. 3719–3729.

[32] Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron C Courville. 2017. Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation.. In *AAAI*. 3288–3294.

[33] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.. In *AAAI*. 3776–3784.

[34] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues.. In *AAAI*. 3295–3301.

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

[36] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[38] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*. 4534–4542.

[39] Jason E Weston. 2016. Dialog-based language learning. In *Advances in Neural Information Processing Systems*. 829–837.

[40] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2018. Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. *IEEE Conference on Computer Vision and Pattern Recognition* (2018), 6106–6115.

[41] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 496–505.

[42] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. Video question answering via attribute-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 829–832.

[43] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017. Multi-level attention networks for visual question answering. In *Conf. on Computer Vision and Pattern Recognition*, Vol. 1. 8.

[44] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging Video Descriptions to Learn Video Question Answering.. In *AAAI*. 4334–4340.

[45] Zhou Zhao, Xinghua Jiang, Deng Cai, Jun Xiao, Xiaofei He, and Shiliang Pu. 2018. Multi-Turn Video Question Answering via Multi-Stream Hierarchical Attention Context Network.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 3690–3696.

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

[47] Xiaoqiang Zhou, Baotian Hu, Qingcai Chen, Buzhou Tang, and Xiaolong Wang. 2015. Answer sequence learning with neural networks for answer selection in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. 713–718.

[48] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4995–5004.