Multi-Speaker Video Dialog with Frame-Level Temporal Localization*

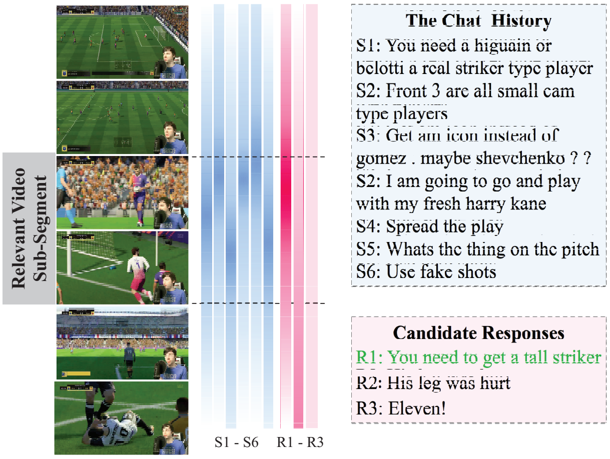
Qiang Wang,[1] Pin Jiang,[1] Zhiyi Guo,[1] Yahong Han,[1]† Zhou Zhao[2]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]College of Computer Science, Zhejiang University, Hangzhou, China
{qiangw, jpin, guo_zhiyi, yahong}@tju.edu.cn, zhaozhou@zju.edu.cn

Abstract

To simulate human interaction in real life, dialog systems are introduced to generate a response to previous chat utterances. There have been several studies for two-speaker video dialogs in the form of question answering. However, more informative semantic cues might be exploited via a multi-rounds chatting or discussing about the video among multiple speakers. So multi-speakers video dialogs are more applicable in real life. Besides, speakers always chat about a sub-segment of the long video fragment for a period of time. Current video dialog systems require to be directly given the relevant video sub-segment which speakers are chatting about. However, it is always hard to accurately spot the corresponding video sub-segment in practical applications. In this paper, we introduce a novel task of Multi-Speaker Video Dialog with frame-level Temporal Localization (MSVD-TL) to make video dialog systems more applicable. Given a long video fragment and a set of chat history utterances, MSVD-TL targets to predict the following response and localize the relevant video sub-segment in frame level, simultaneously. We develop a new multi-task model with a response prediction module and a frame-level temporal localization module. Besides, we focus on the characteristic of the video dialog generation process and exploit the relation among the video fragment, the chat history, and the following response to refine their representations. We evaluate our approach for both the Multi-Speaker Video Dialog without frame-level temporal localization (MSVD w/o TL) task and the MSVD-TL task. The experimental results further demonstrate that MSVD-TL enhances the applicability of video dialog in real life.

Introduction

With the advance of artificial intelligence, dialog systems are introduced to simulate human interaction in real life and generate a response to previous chat utterances. Early dialog systems (Sordoni et al. 2015; Lowe et al. 2015; Serban et al. 2016) are just limited to the application of the textual dialog in natural language understanding. However, there are more and more multimedia data in real life. To expand the applicable scope of dialog systems, Das et al. (Das et al. 2017) proposed Visual Dialog to help two speakers understand the static visual information of the image. Alamri et al. (Alamri et al. 2018) introduced Scene-Aware Dialog as an application of dialog systems for videos with both static visual information and dynamic temporal information. Although Visual Dialog and Scene-Aware Dialog successfully introduce visual understanding of images and videos to dialog systems, they just consider two-speaker dialogs in the form of question answering. More informative semantic cues might be exploited via a multi-rounds chatting or discussing about the video among multiple speakers. As shown in Figure 1, multiple speakers could provide more various views for video understanding with more detailed information. Besides, Figure 1 also shows that speakers always chat about a sub-segment of the long video fragment for a period of time. Scene-Aware Dialog is directly given the relevant video sub-segment which speakers are chatting about beforehand. However, it is always hard to accurately spot the corresponding video sub-segment in practical applica-



The Chat History
S1: You need a higuain or belotti a real striker type player
S2: Front 3 are all small cam type players
S3: Get am icon instead of gomez . maybe shevchenko ? ?
S2: I am going to go and play with my fresh harry kane
S4: Spread the play
S5: Whats the thing on the pitch
S6: Use fake shots

Candidate Responses
R1: You need to get a tall striker
R2: His leg was hurt
R3: Eleven!

Figure 1: The chat history utterances and the following response for a dialog usually correspond to a sub-segment of the long video fragment, as shown in the middle bars.

*This work is supported by the NSFC (under Grant 61876130, 61932009, U1509206).

†Corresponding author.

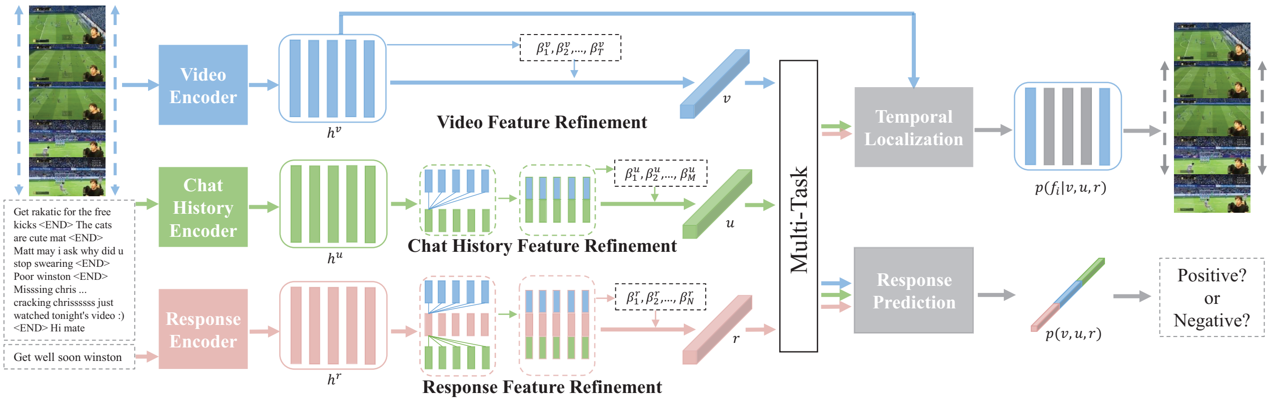

Video Encoder
h^v
$\beta_1^v, \beta_2^v, ..., \beta_T^v$
v
Video Feature Refinement
Chat History Encoder
h^u
$\beta_1^u, \beta_2^u, ..., \beta_M^u$
u
Chat History Feature Refinement
Response Encoder
h^r
$\beta_1^r, \beta_2^r, ..., \beta_N^r$
r
Response Feature Refinement
Get rakatic for the free kicks <END> The cats are cute mat <END> Matt may i ask why did u stop swearing <END> Poor winston <END> Misssing chris ... cracking chrissssss just watched tonight's video :) <END> Hi mate
Get well soon winston
Multi-Task
Temporal Localization
$p(f_i|v,u,r)$
Response Prediction
$p(v,u,r)$
Positive? or Negative?

Figure 2: The framework of our approach. Firstly, the video encoder, the chat history encoder, and the response encoder extract frame features for the video fragment, word features for the chat history, and word features for candidate responses, respectively. Then, we employ the designed selective refinement module to refine these features with the relation among them and produce the corresponding global features. Finally, the response prediction module and the frame-level temporal localization module judge the correctness of the input response and localize the relevant video sub-segment in frame level, respectively.

tions, which will cause some limitations of the video dialog system for practical application. Therefore, adding a temporal localization function to automatically search the relevant video sub-segment is desirable for video dialog systems.

In this paper, we introduce a novel task of Multi-Speaker Video Dialog with frame-level Temporal Localization (MSVD-TL). To evaluate this task, we extend the Twitch-FIFA dataset (Pasunuru and Bansal 2018) which provides collected soccer game videos along with multiple users' live chat conversations about the game. Given a long video fragment and a set of chat history utterances, MSVD-TL targets to predict the following response and localize the relevant video sub-segment in frame level, simultaneously. The two parallel sub-tasks of response prediction and temporal localization motivate us to develop a new multi-task model. As shown in Figure 2, the response prediction module judges whether the triple of the video fragment, the chat history, and the candidate response is positive or negative. The frame-level temporal localization module predicts the relevance score of each frame and localizes the relevant video sub-segment in frame level. Besides, we focus on the characteristic of the video dialog generation process and exploit the relation among the video fragment, the chat history, and the following response to refine their representations. Finally, we evaluate our approach for both the multi-speaker video dialog without frame-level temporal localization (MSVD w/o TL) task and the MSVD-TL task. The experimental results demonstrate that video dialog with frame-level temporal localization could achieve perfect performance. It also illustrates that MSVD-TL solves the limitation of pre-given video fragments in practical applications. Besides, the comparison experiments of our model illustrate that considering the relation among inputs of the video fragment, the chat history, and the following response could enhance their representations.

Our contributions of this work are as follows: (1) We introduce a novel task MSVD-TL to enhance the applicability of video dialog systems in real life. (2) We develop a multi-task framework to predict the following response and localize the relevant video sub-segment in frame level, simultaneously. (3) We focus on the characteristic of the video dialog generation process and exploit the relation among the video fragment, the chat history, and the following response to refine their representations.

Related Work

Dialog systems have been studied with the outstanding achievement in the field of artificial intelligence. However, early dialog systems (Shaikh et al. 2010; Vinyals and Le 2015; Serban et al. 2017; Luan, Ji, and Ostendorf 2016) are almost based on textual dialogs, which just focus on semantic understanding of natural language sentences. As more and more multimedia data appear in real life, dialog systems for multi-modal comprehension have attracted more researchers. Recently, Das et al. (Das et al. 2017) proposed the Visual Dialog task, which simulates the communication process about an image between a questioner and an answerer. The goal of Visual Dialog is to predict the response based on both the visual information of the image, the semantic information of the dialog history, and the semantic information of the current question. To focus on the characteristic of the dialog generation process, AMAE (Seo et al. 2017), HCIAE (Lu et al. 2017), SF (Jain, Lazebnik, and Schwing 2018), Coatt (Wu et al. 2018), and CorefNMN (Kottur et al. 2018) all utilize the relation among inputs of the image, the dialog history, the current question, and candidate answers to enhance their representations. Although the task of Visual Dialog really extends the application scope of dialog systems with visual understanding, images just contain the static visual information while videos own additional dynamic temporal information. Based on it, Alamr et al. (Alamri et al. 2018) introduced the task of Scene-Aware Dialog for videos. Similar to Visual Dialog, Scene-Aware Dialog aims to predict the response to current question based on the visual information of the video and the semantic information of the dialog history. There are also several methods to solve this task. Nguyen et al. (Nguyen et al. 2018) used the FiLM block (Perez et al. 2018) to extract features for inputs, which ultimately leads to the extraction of more relevant features. Besides, Pasunuru et al. (Pasunuru and Bansal 2019) and Hori et al. (Hori et al. 2018) both employed attention mechanisms to explore the relation among inputs of the video, the dialog history, and the question.

Although Visual Dialog and Scene-Aware Dialog achieved the application of dialog systems for images and videos, they just considers the two-speaker dialog in the form of questioning answering. Whereas more informative semantic cues might be exploited via a multi-rounds chatting or discussing about the video among multiple speakers. Besides, compared with the Visual Dialog task which needs to automatically search the relevant regions based on semantic information, the Scene-Aware Dialog task is directly given the relevant video sub-segment which speakers are chatting about beforehand. However, it is always hard to accurately spot the corresponding video sub-segment in practical applications. The requirement of the relevant video sub-segment beforehand will cause some limitations of video dialog systems for practical application. Therefore, a video dialog system with the function of automatical temporal localization is desirable to enhance its practical applicability.

The Proposed Approach

Formally, given a video fragment $V = \{f_t\}_{t=1}^{T}$ and a set of chat history utterances $U = \{u_k\}_{k=1}^{K}$, the goal of MSVD-TL is to predict the following response R and localize the relevant video sub-segment $V_s = \{f_t\}_{t=S}^{E}$ in frame level. In detail, T is the total frame number of the video fragment. K is the utterance number of the chat history. S and E are the start frame index and the end frame index of the relevant video sub-segment, respectively. We develop a new multi-task model with a response prediction module and a frame-level temporal localization module to solve the two sub-tasks respectively. Beside, we focus on the characteristic of the video dialog generation process and design a selective refinement module to exploit the relation among the video fragment, the chat history, and the following response to refine their representations. In this section, we firstly introduce the data preprocessing process to extract features of the video fragment, the chat history, and the following response. Then, we explain how the selective refinement module works. Finally, we devise the response prediction module and the frame-level temporal localization module to predict the following response and localize the relevant video sub-segment in frame level.

Data Preprocessing

During the phase of data preprocessing, we firstly extract frame features for the video fragment, word features for the chat history, and word features for the following response. Then, we process these features through three independent recurrent neural networks (RNN) to learn the temporal characteristic, respectively.

Video Encoder For the video fragment V, we firstly extract the static feature for each frame with pre-trained Inception-v3 (Szegedy et al. 2016) convolutional neural network (CNN) model. Then we add the dynamic temporal information with a RNN. The output video fragment feature is denoted as $H^v = \{h_i^v\}_{t=1}^{T}$, where T is the total frame number of the video fragment and h_i^v is the i-th frame feature with both static frame information and dynamic temporal information.

Chat History Encoder For the set of chat history utterances U, we firstly join all chat history utterances with an *END* token to generate a sequence of words. After that, we use GloVe (Pennington, Socher, and Manning 2014) as the embedding matrix to embed each independent word in the sequence. Then, we employ a RNN to add the sequential information to word features. The output chat history feature is denoted as $H^u = \{h_i^u\}_{i=1}^{M}$, where M is the maximum word number of the joint chat history and h_i^u is the i-th word feature.

Response Encoder For the following response R, we firstly use GloVe (Pennington, Socher, and Manning 2014) as the embedding matrix to embed each independent word in the sentence. Then, we employ a RNN to add the sequential information to word features. The output response feature is denoted as $H^r = \{h_i^r\}_{i=1}^{N}$, where N is the maximum word number of the following response and h_i^r is the i-th word feature.

Selective Refinement Module

In consideration of the characteristic of video dialog generation process, we design a selective refinement module to enhance representation of the video fragment, the chat history, and the following response by adding supplement information for them. Specifically, the video fragment is the reference of the chat history and the following response. Therefore, the video fragment does not need supplement information. However, the visual information of the video fragment could be the supplement information for the chat history and the following response. Besides, the following response is not only based on the video fragment but also relevant to the chat history. Therefore, the supplement information of the following response also contains the semantic information of the chat history.

Video Feature Refinement Although the video fragment contains complete information, different frames play different roles in video understanding. We apply the self-attention mechanism to select important frames and produce the global video feature. Firstly, we calculate the importance score α_i^v for the i-th frame in the video fragment, and normalize the importance score over all frames to produce the corresponding important weight β_i^v. Then, the global video feature v is formulated as the dynamic weighted summation of all frame features.

$$\alpha_i^v = W_v^T h_i^v \tag{1}$$

$$\beta_i^v = \frac{exp(\alpha_i^v)}{\sum_{i=1}^{T} exp(\alpha_i^v)} \tag{2}$$

$$v = \sum_{i=1}^{T} \beta_i^v \cdot h_i^v \tag{3}$$

where W_v is a trainable parameter, and T is the total number of frames in the video fragment.

Chat History Feature Refinement Since the visual information of the video fragment could be the supplement information for the chat history, we employ a cross-attention mechanism to add relevant visual information to word representations of the chat history. Besides, different words in the chat history also play different roles in semantic understanding. Therefore, we apply a self-attention mechanism to select important words to product the global chat history feature, respectively.

Firstly, we calculate the relevance score $\alpha_{i,j}^{v\to u}$ for the pair of i-th word in the chat history and the j-th frame in the video fragment, and normalize the relevance score over all video frames to produce the relevant weight $\beta_{i,j}^{v\to u}$. Then, the weighted summation of all video frame features are regarded as the supplement feature for the i-th word in the chat history.

$$\alpha_{i,j}^{v\to u} = W_{u_1}^T tanh(W_{uu} h_i^u + W_{uv} h_j^v) \tag{4}$$

$$\beta_{i,j}^{v\to u} = \frac{exp(\alpha_{i,j}^{v\to u})}{\sum_{j=1}^{T} exp(\alpha_{i,j}^{v\to u})} \tag{5}$$

$$e_i^{v\to u} = \sum_{j=1}^{T} \beta_{i,j}^{v\to u} \cdot h_j^v \tag{6}$$

where W_{u_1}, W_{uu}, and W_{uv} are trainable parameters. T is the total frame number of the video fragment.

Secondly, we fuse the supplement feature $e_i^{v\to u}$ and the original feature h_i^u by a concatenation operation. The result is regarded as the new feature for the i-th word of the chat history. After that, we calculate the importance score α_i^u for the i-th word in the chat history, and normalize the importance score over all words to produce the important weight β_i^u. Then, the global chat history feature u is formulated as the dynamic weighted summation of all new word features.

$$\hat{h}_i^u = [h_i^u; e_i^{v\to u}] \tag{7}$$

$$\alpha_i^u = W_u^T \hat{h}_i^u \tag{8}$$

$$\beta_i^u = \frac{exp(\alpha_i^u)}{\sum_{i=1}^{M} exp(\alpha_i^u)} \tag{9}$$

$$u = \sum_{i=1}^{M} \beta_i^u \cdot \hat{h}_i^u \tag{10}$$

where W_u is a trainable parameter and M is the max number of words in the joint sentence of the chat history. $[;]$ is the concatenation operation.

Response Feature Refinement As the visual information of the video fragment could be the supplement information for the chat history, the visual information of the video fragment and the semantic information of the chat history both could provide the supplement information for the following response. Therefore, we utilize two cross-attention mechanisms to capture relevant visual information of the video fragment and semantic information of the chat history, respectively. Besides, similar to word selection for the chat history, we also apply a self-attention mechanism to select important words to product the global response feature.

Firstly, we calculate the relevance score $\alpha_{i,j}^{v\to r}$ for the pair of i-th word in the following response and the j-th frame in the video fragment, and normalize the relevance score over all video frames to produce the relevant weight $\beta_{i,j}^{v\to r}$. Then, the weighted summation of all video frame features are regarded as the visual supplement feature for the i-th word in the following response.

Secondly, we calculate the relevance score $\alpha_{i,j}^{u\to r}$ for the pair of i-th word in the following response and the j-th word in the chat history, and normalize the relevance score over all words of the chat history to produce the relevant weight $\beta_{i,j}^{u\to r}$. Then, the weighted summation of all word features of the chat history are regarded as the semantic supplement feature for the i-th word in the following response.

Thirdly, we fuse the visual supplement feature $e_i^{v\to r}$, the semantic supplement feature $e_i^{u\to r}$, and the original feature h_i^r by a concatenation operation. The result is regarded as the new feature for the i-th word of the following response. After that, we calculate the importance score α_i^r for the i-th word in the following response, and normalize the importance score over all words to produce the important weight β_i^r. Then, the global response feature r is formulated as the dynamic weighted summation of all word features.

Multi-Task Prediction Module

We design a multi-task prediction module for MSVD-TL with multiple subtasks. Specifically, the response prediction module judge the correctness of the input response, and the frame-level temporal localization module localize the relevant video sub-segment in frame level.

Response Prediction Module Our model targets to predict the correctness of the following response. The following response is based on both the visual information of the video fragment and the semantic information of the chat history. Given the global features of the video fragment, the chat history, and the following response, we calculate the relevance score x among them and normalize the relevance score to the relevant probability $p(v, u, r)$ by the sigmoid function.

$$x = W_{pv} v \odot r + W_{pu} u \odot r \tag{11}$$

$$p(v, u, r) = \sigma(W_p^T x + b) \tag{12}$$

where W_{pv}, W_{pu}, and W_p are trainable parameters. $\odot$ is the element-wise product operation. σ is the sigmoid function.

Frame-Level Temporal Localization Module Our model aims to localize the relevant video sub-segment which is corresponding to the chat history and the following response in frame level. Given frame features of the video fragment, the global feature of the chat history, and the global feature of the following response, we firstly use the global features of the chat history and the following response as the semantic guidance to calculate the relevance score z_i of each frame feature. Then, the relevance score is normalized to the relevant probability $p(f_i|v,u,r)$ using the sigmoid function.

$$z_i = W_{lv}h_i^v \odot W_{lu}u + W_{lv}h_i^v \odot W_{lr}r \quad (13)$$

$$p(f_i|v,u,r) = \sigma(W_l^T z_i + b) \quad (14)$$

where W_{lv}, W_{lu}, W_{lr}, and W_l are trainable parameters. $\odot$ is the element-wise product operation. σ is the sigmoid function.

Multi-Task Loss

To optimize our model, we use a multi-task loss L, including L_{pred} for response prediction and L_{loc} for temporal localization. We minimize the objective function over all the training set.

$$L = L_{pred} + \lambda L_{loc} \quad (15)$$

where λ is a hyper-parameter to balance two losses and it is set to 1.0 in our experiments.

We use the max-margin loss function (Mao et al. 2016; Yu et al. 2017) as L_{pred} for response prediction. Given a positive training triple (v,u,r), we use the corresponding negative training triples of $(v^{'},u,r)$, $(v,u^{'},r)$, and $(v,u,r^{'})$ which replace a correct input with a wrong input at a time for each of v, u, and r. And $p(v,u,r)$ represents the probability that the triple (v,u,r) is positive.

$$L_{pred} = \sum max(0, \rho + log\, p(v,u,r) - log\, p(v^{'},u,r)) \quad (16)$$

$$+max(0, \rho + log\, p(v,u,r) - log\, p(v,u^{'},r)) \quad (17)$$

$$+max(0, \rho + log\, p(v,u,r) - log\, p(v,u,r^{'})) \quad (18)$$

where the summation is over all the training triples. ρ is a hyper-parameter to tune the margin between positive and negative training triples. We set ρ to 0.1 in our experiments.

We use the cross-entropy loss function as L_{loc} for temporal localization. $p(f_i|v,u,r)$ is the probability that i-th frame is belong to the relevant video sub-segment, and $gt(f_i|v,u,r) \in \{0,1\}$ is the ground truth label to present i-th frame is or not belong to the relevant video sub-segment.

$$L_{loc} = -\sum \sum_{i=1}^{T} gt(f_i|v,u,r) log\, p(f_i|v,u,r) \quad (19)$$

$$+(1 - gt(f_i|v,u,r)) log\,(1 - p(f_i|v,u,r)) \quad (20)$$

where the summation is over all the training triples, and T is the total frame number of the video fragment.

Experiments

To solve the MSVD-TL task, we build a new dataset on top of Twitch-FIFA dataset (Pasunuru and Bansal 2018) which provides collected soccer game videos along with users' live chat conversations about the game. We firstly divide the whole video into several 50-second video fragments. Then we randomly select users' live chat conversations in a 20-second video sub-segment as the chat history. And we regard the most relevant utterance in the following 10-second video clip as the ground truth of the following response. There are 49 game videos totally, which are divided into 33 videos for training, 8 videos for validation, and 8 videos for testing. And each video is several hours long, which provides a great amount of data. After processing, there are 10510 samples in the training set, 2153 samples in the validation set, and 2780 in the test set, respectively. For each sample, there are a set of candidate responses and only one is positive while the others are negative. The number of candidate responses in the training set, validation set, and test set are 2, 10, and 10, respectively.

Evaluation Metrics

Since the original Twitch-FIFA dataset (Pasunuru and Bansal 2018) has provided a set of candidate responses for each chat history, we use the retrieval metric R@k (the percentage of the response in top-k ranked responses) to evaluate the performance of response prediction instead of BLEU (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and METEOR (Denkowski and Lavie 2014) in other language generation tasks. And higher R@k is better for response prediction.

Experimental Settings

Before training the model, we randomly select three frames for each second in the video fragment to extract static frame features. The total frame number of each video fragment T is 150. And the maximum word number of the chat history M is 70 while the maximum word number of the following response N is 10. During the training phase, the embedding size of words is 100. All RNNs in our model are bidirectional single-layer Long short-term memory networks (LSTM) (Schuster and Paliwal 1997; Hochreiter and Schmidhuber 1997). The size of hidden states in RNNs is 256. Therefore, the dimension of frame features for the video fragment, word features for the chat history, and word features for the following response are all 512. We rely on the Adam (Kingma and Ba 2014) algorithm to update all parameters in our model with the learning rate of 10^{-5}. The experimental hardware environment is 1080ti GPU. During the training process, the batch size is set to 16 and the model is trained for 30,000 iterations.

Experimental Results for MSVD w/o TL

Firstly, we directly utilize the 20-second relevant video sub-segment to predict the following response without frame-level temporal localization. Therefore, inputs of our model are a 20-second relevant video sub-segment, the chat history, and the candidate response. "C" represents that we just explore the semantic information of the chat history to predict the following response. "V" represents that we just use the visual information of the relevant video sub-segment to predict the following response. "C+V" represents that we utilize both the semantic information of the chat history and the visual information of the relevant video sub-segment to predict the following response. We compare our approach with the state-of-the-art methods for the MSVD w/o TL task on Twitch-FIFA dataset (Pasunuru and Bansal 2018). "Dual Encoder" and "Triple Encoder" (Pasunuru and Bansal 2018) directly fuse features of inputs and use the joint representation to predict the correctness of the input triple. "TriDAF+self Attn" (Pasunuru and Bansal 2018) apply several bidirectional attentions and self attentions to enable attention flows across all three modalities of the relevant video sub-segment, the chat history, and the candidate response.

Table 1: Experimental results compared with the state-of-the-art methods for MSVD w/o TL (20s).

Models	R@1	R@2	R@5
Dual Encoder (C) (Pasunuru and Bansal 2018)	17.1	30.3	61.9
Dual Encoder (V) (Pasunuru and Bansal 2018)	16.3	30.5	61.1
Triple Encoder (C+V) (Pasunuru and Bansal 2018)	18.1	33.6	68.5
TriDAF+self Attn (C+V) (Pasunuru and Bansal 2018)	20.7	35.3	69.4
SelRef (C)	17.9	32.3	67.0
SelRef (V)	19.5	33.5	68.7
SelRef (C+V)	21.4	36.0	69.7

The experimental results are shown in Table 1. "Dual Encoder" and "Triple Encoder" (Pasunuru and Bansal 2018) do not consider the relation among inputs of the relevant video-segment, the chat history, and the following response. Although "TriDAF+self Attn" (Pasunuru and Bansal 2018) enables attention flows across all three modalities directly and ignores the relation among inputs during the video dialog generation process. Our approach "SelRef (C+V)" outperforms "Triple Encoder (C+V)" with the improvement of 3.3% in R@1 score. It demonstrates that the relation among inputs really enhances representations of inputs. Besides, Our approach "SelRef(C+V)" outperforms "TriDAF+self Attn (C+V)" with the improvement of 0.70% in R@1 score. It illustrates that considering the characteristic of video dialog generation process really helps to enhance representations of inputs.

Table 2: Experimental results compared with the state-of-the-art methods for MSVD w/o TL (50s).

Models	R@1	R@2	R@5
TriDAF+self Attn (C+V) (Pasunuru and Bansal 2018)	20.3	34.6	68.2
SelRef (C)	17.4	32.9	67.0
SelRef (V)	20.2	34.2	67.6
SelRef (C+V)	20.9	35.9	69.1

Additionally, in order to show the necessity of the temporal localization phase in video dialog, we utilize the 50-second video fragment to predict the following response without the temporal localization phase. Therefore, the inputs of our model are a 50-second video fragment, the chat history, and the candidate response. And we compare our approach with the state-of-the-art method (Pasunuru and Bansal 2018) for MSVD w/o TL on Twitch-FIFA dataset. The experimental results are shown in Table 2. Compared with experimental results in Table 1, the performance of "TriDAF+self Attn (C+V)" is reduced by 1.2% in R@5 score. The performance of "SelRef (C+V)" is reduced by 0.6% in R@5 score. It further demonstrates the necessary of temporal localization in video dialog.

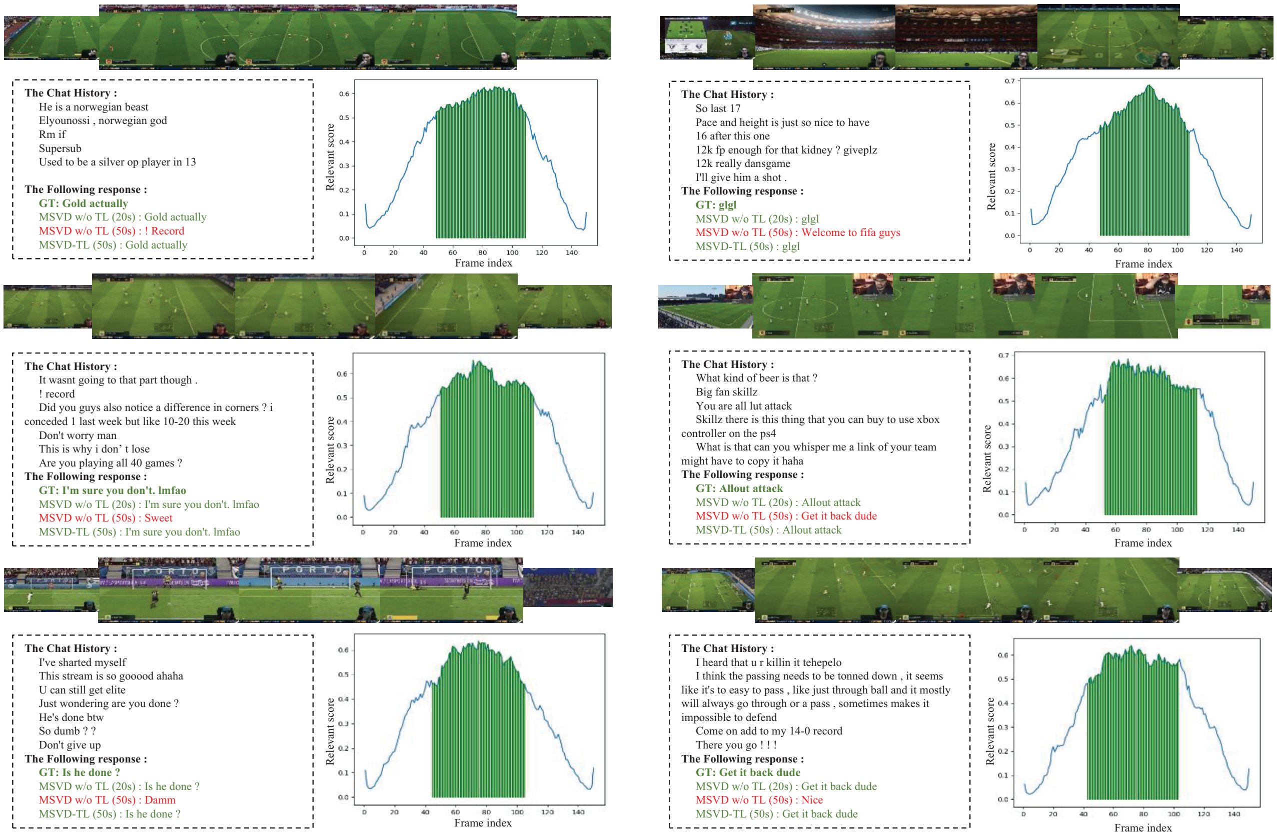
Experimental Results for MSVD-TL

To evaluate our approach for the MSVD-TL task, we construct several experiments, which utilize the 50-second video fragment to predict the following response with the temporal localization phase. Therefore, the inputs of our model are a 50-second video fragment, the chat history, and the candidate response. "C" represents that we just explore the semantic information of the chat history to predict the following response, while the semantic of both the chat history and the following response are used to localize the relevant video sub-segment. "V" represents that we just explore the visual information of the video fragment to predict the following response and localize the relevant video sub-segment based on the following response. "C+V" represents that both the semantic information of the chat history and the visual information of the video fragment are used for response prediction, while the semantic information of both the chat history and the following response are used to localize the relevant video sub-segment.

Table 3: Experimental results compared with the state-of-the-art methods for MSVD-TL (50s).

Models	R@1	R@2	R@5
TriDAF+self Attn (C+V) (Pasunuru and Bansal 2018)	20.40	35.0	69.17
SelRef (C)	17.3	32.1	66.7
SelRef (V)	19.5	35.3	68.4
SelRef (C+V)	21.5	36.1	70.1

The experimental results are shown in Table 3. Compared with the experimental results in Table 1, the performance of our approach for MSVD-TL (50s) is similar to the performance of our approach for MSVD w/o TL (20s). Compared with the experimental results in Table 2, the performance of our approach for MSVD-TL (50s) is better than the performance of our approach for MSVD w/o TL (50s). It further demonstrates that the temporal localization module in our model could accurately localize the relevant video sub-segment, which could improve the applicability of video dialog in practice.



The Chat History :
He is a norwegian beast
Elyounossi , norwegian god
Rm if
Supersub
Used to be a silver op player in 13

The Following response :
GT: Gold actually
MSVD w/o TL (20s) : Gold actually
MSVD w/o TL (50s) : ! Record
MSVD-TL (50s) : Gold actually

The Chat History :
So last 17
Pace and height is just so nice to have
16 after this one
12k fp enough for that kidney ? giveplz
12k really dansgame
I'll give him a shot .
The Following response :
GT: glgl
MSVD w/o TL (20s) : glgl
MSVD w/o TL (50s) : Welcome to fifa guys
MSVD-TL (50s) : glgl

The Chat History :
It wasnt going to that part though .
! record
Did you guys also notice a difference in corners ? i conceded 1 last week but like 10-20 this week
Don't worry man
This is why i don’ t lose
Are you playing all 40 games ?
The Following response :
GT: I'm sure you don't. lmfao
MSVD w/o TL (20s) : I'm sure you don't. lmfao
MSVD w/o TL (50s) : Sweet
MSVD-TL (50s) : I'm sure you don't. lmfao

The Chat History :
What kind of beer is that ?
Big fan skillz
You are all lut attack
Skillz there is this thing that you can buy to use xbox controller on the ps4
What is that can you whisper me a link of your team might have to copy it haha
The Following response :
GT: Allout attack
MSVD w/o TL (20s) : Allout attack
MSVD w/o TL (50s) : Get it back dude
MSVD-TL (50s) : Allout attack

The Chat History :
I've sharted myself
This stream is so gooood ahaha
U can still get elite
Just wondering are you done ?
He's done btw
So dumb ? ?
Don't give up
The Following response :
GT: Is he done ?
MSVD w/o TL (20s) : Is he done ?
MSVD w/o TL (50s) : Damm
MSVD-TL (50s) : Is he done ?

The Chat History :
I heard that u r killin it tehepelo
I think the passing needs to be tonned down , it seems like it's to easy to pass , like just through ball and it mostly will always go through or a pass , sometimes makes it impossible to defend
Come on add to my 14-0 record
There you go ! ! !
The Following response :
GT: Get it back dude
MSVD w/o TL (20s) : Get it back dude
MSVD w/o TL (50s) : Nice
MSVD-TL (50s) : Get it back dude

Figure 3: Examples in the MSVD w/o TL (20s) task, the MSVD w/o TL (50s) task, and the MSVD-TL (50s) task. For each sample, the top shows some frames of the video including both the relevant frames (larger) and the irrelevant frames (smaller). The left-bottom shows the chat history, the ground truth response, and the predicted response for different tasks. And the right-bottom shows the result of temporal localization for the relevant video sub-segment in the MSVD-TL (50s) task.

Examples of Experimental Results

In Figure 3, we show some examples in the MSVD w/o TL (20s) task, the MSVD w/o TL (50s) task, and the MSVD-TL (50s) task. For each sample, the top shows some frames of the video including both the relevant frames (larger) and the irrelevant frames (smaller). The left-bottom shows the chat history, the ground truth response, and the predicted response for different tasks. And the right-bottom shows the result of temporal localization for the relevant video sub-segment in the MSVD-TL (50s) task. In the result of temporal localization, the green area corresponds to the ground truth of the relevant video sub-segment, while each point in the blue line represents the probability that the corresponding frame is belong to the relevant video sub-segment. We can see that the predicted response is correct when we directly use the relevant video sub-segment in the MSVD w/o TL (20s) task. The predicted response is wrong when we expand the video length and directly use the video fragment in the MSVD w/o TL (50s) task. And the predicted response is correct again when we expand the video length and apply the frame-level temporal localization simultaneously. Besides, the results of temporal localization shows that frames with more than 0.5 relevance score are consistent with the corresponding ground truth. It illustrates that the frame-level temporal localization really learn the ability to search the relevant video sub-segment automatically. And it will be more applicable in practice without the requirement of the given relevant video sub-segment in advance.

Conclusions

In this paper, we firstly introduce a novel task of Multi-Speaker Video Dialog with frame-level Temporal Localization (MSVD-TL). Given a long video fragment and the chat history, MSVD-TL targets to predict the following response and localize the relevant video sub-segment in frame level, simultaneously. We develop a new multi-task model to solve the two sub-tasks of response prediction and temporal localization. In additional, we design a selective refinement module to exploit the relation among the video fragment, the chat history, and the following response to refine their representations. Finally, we construct a series of experiments to evaluate our approach on MSVD-TL. The experimental results demonstrate that the added temporal localization module could help to select relevant video sub-segment within a longer video, which is more applicable in the real life.

References

Alamri, H.; Cartillier, V.; Lopes, R. G.; Das, A.; Wang, J.; Essa, I.; Batra, D.; Parikh, D.; Cherian, A.; Marks, T. K.; et al. 2018. Audio visual scene-aware dialog (avsd) challenge at dstc7. *arXiv preprint arXiv:1806.00525*.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

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

Hori, C.; Alamri, H.; Wang, J.; Winchern, G.; Hori, T.; Cherian, A.; Marks, T. K.; Cartillier, V.; Lopes, R. G.; Das, A.; et al. 2018. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. *arXiv preprint arXiv:1806.08409*.

Jain, U.; Lazebnik, S.; and Schwing, A. G. 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*, volume 1.

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

Kottur, S.; Moura, J. M.; Parikh, D.; Batra, D.; and Rohrbach, M. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 153–169.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Lu, J.; Kannan, A.; Yang, J.; Parikh, D.; and Batra, D. 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.

Luan, Y.; Ji, Y.; and Ostendorf, M. 2016. Lstm based conversation models. *arXiv preprint arXiv:1603.09457*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Nguyen, D. T.; Sharma, S.; Schulz, H.; and Asri, L. E. 2018. From film to video: Multi-turn question answering with multi-modal context. *arXiv preprint arXiv:1812.07023*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318.

Pasunuru, R., and Bansal, M. 2018. Game-based video-context dialogue. *arXiv preprint arXiv:1809.04560*.

Pasunuru, R., and Bansal, M. 2019. Dstc7-avsd: Scene-aware video-dialogue systems with dual attention. In *DSTC7 at AAAI2019 workshop*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, 1532–1543.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Seo, P. H.; Lehrmann, A.; Han, B.; and Sigal, L. 2017. Visual reference resolution using attention memory for visual dialog. In *Advances in Neural Information Processing Systems*, 3719–3729.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shaikh, S.; Strzalkowski, T.; Broadwell, G. A.; Stromer-Galley, J.; Taylor, S. M.; and Webb, N. 2010. Mpc: A multi-party chat corpus for modeling social phenomena in discourse. In *proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Vinyals, O., and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wu, Q.; Wang, P.; Shen, C.; Reid, I.; and van den Hengel, A. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yu, L.; Tan, H.; Bansal, M.; and Berg, T. L. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7282–7290.