End-to-end Recurrent Cross-Modality Attention for Video Dialogue

Yun-Wei Chu[1] Kuan-Yen Lin[2] Chao-Chun Hsu[3] Lun-Wei Ku[4]
[1]Purdue University, [2]Cornell Tech, [3]University of Chicago, [4]Academia Sinica,
chu198@purdue.edu, kl924@cornell.edu, chaochunh@uchicago.edu, lwku@iis.sinica.edu.tw

***Abstract*—Visual dialogue systems need to understand dynamic visual scenes and comprehend semantics in order to converse with users. Constructing video dialogue systems is more challenging than traditional image dialogue systems because the large feature space of videos makes it difficult to capture semantic information. Furthermore, the dialogue system also needs to precisely answer users' question based on comprehensive understanding of the videos and the previous dialogue. In order to improve the performance of video dialogue system, we proposed an end-to-end recurrent cross-modality attention (ReCMA) model to answer a series of questions about a video from both visual and textual modality. The answer representation of the question is updated based on both visual representation and textual representation in each step of the reasoning process to have a better understanding of both modalities' information. We evaluate our method on the challenging DSTC7 video scene-aware dialog dataset and the proposed ReCMA achieves a relative 20.8% improvement over the baseline on CIDEr.**

I. Introduction

DEEP neural networks have been successfully understand both visual information and natural language, with applications ranging from image captioning [9], [20], [23], [39], [43], [50], [68], [69] to image-based visual question answering (VQA) [4], [14], [24], [41], [46], [72], [75], [81]. Different from image-based VQA, which the model can generate an answer of a single question about a static image, image-based visual dialogue [11], [17], [28], [38], [47], [58], [64] was introduced to hold a meaningful dialogue with humans about an image using conversational language. However, intelligent systems are difficult to well interact with human users when only accessing a single image without dynamic scenes. Therefore, the ability to reason on a video is important and deserves to be discussed.

Moving from a single image to video is challenging for vision-to-language systems because systems need to understand dynamic visual scenes, natural language, and multiple modalities interaction. To grasp the semantics of dynamic scenes, recent research has focused on video captioning [19], [27], [36], [52], [54] and video question answering [32], [65], [77], [82], [83]. Instead of answering single question of a video, video dialogue system [1], [16], [56] is designed to understand dialogue context and answer series of question of a given video. Developing video dialogue systems is more challenging than constructing traditional image-language systems because feature space of video is more complicated than image-based features. To be more specific, videos contain diverse objects, flow of actions, and dynamic light source, yielding video processing more difficult than image processing. Moreover, instead of answering one question or selecting an answer from multiple choices, generating an open-end answer from a series of question-answer pairs causes video dialogue system more complex than video question answering. Figure 1 shows an example of dialogue corresponds to a video from the Dialog System Technology Challenge 7 (DSTC7) dataset [16].

Video Caption: a person is sitting on the sofa. the person gets up and walks over to the table.

Video Summary: a man is sitting on a bench and wiping his eyes and face with his hands. he places his glasses on and then gets up and goes into another room where he rumbles through a drawer.

Question-Answer Pairs:
Q_1: is he wiping his eyes in the beginning ?
A_1: yes , it looks like he is wiping his eyes and rubbing his face
Q_2: are his glasses off as the video starts ?
A_2: yes , he is holding his glasses in his hand
⋮
Q_{10}: is there anything else noteworthy about this video ? does he take anything out of the drawer ?
A_{10}: not really. he doesn't take anything out of the drawer. he kind of looks bored and like he may be looking for something

Fig. 1. A sample of caption, summary, and question-answer pairs for a given video from the DSTC7 dataset. i-th turn question and answer are denoted as Q_i and A_i respectively.

End-to-end vision-to-language systems have been growing awareness because they can be trained by paired input and output texts, without pre-designed data processing modules. The development cost of systems can also be reduced by end-to-end training procedure, and this approach shows improvement when utilizing large conversational datasets [37], [67]. Recently, attention mechanisms [5], [74] have shown benefits on several end-to-end vision-to-language systems. Attention mechanisms can capture relevant region on both visual feature and textual feature that correspond to the query. However,

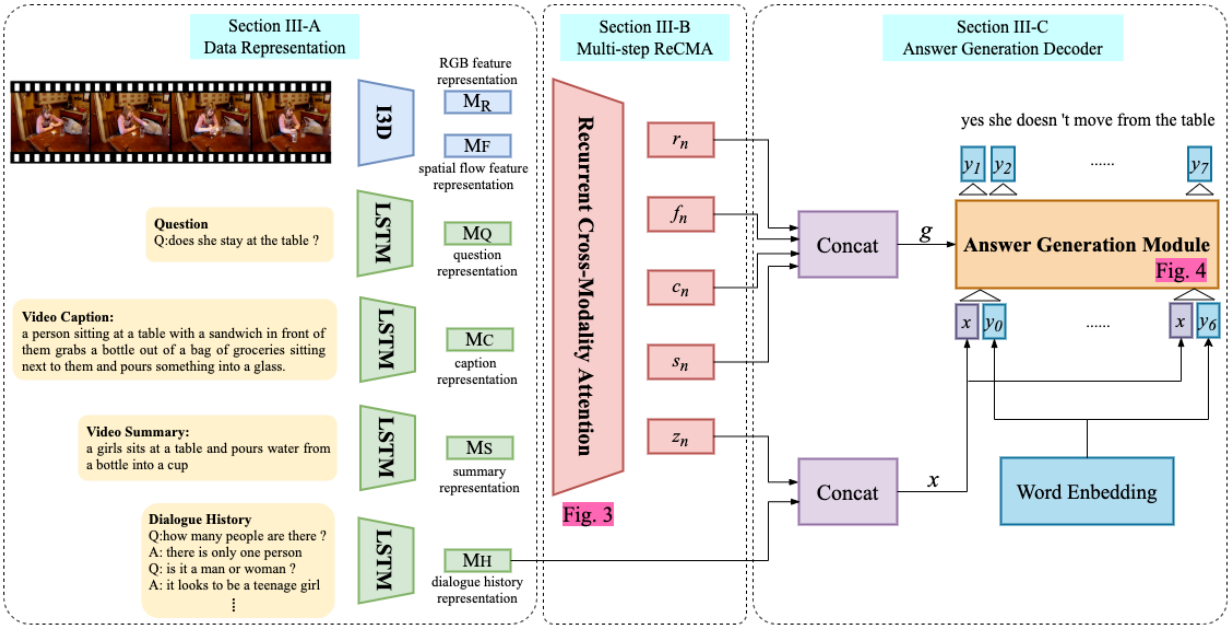


Fig. 2. Overview of our end-to-end video dialogue system. The system first produces representations of each data. The proposed recurrent cross-modality attention then learns the important regions from both visual and textual domain. Finally, the system outputs the answer corresponds to the input question by answer generation module. The detailed descriptions are introduced in Section III.

the performance of attention-based vision-to-language systems often declines when the answer lies in a specific region of image or video with plenty of objects or dynamic background. Some may hypothesize the limited performance of the attention-based vision-to-language systems for the deficient ability of single-step reasoning [13]. Moreover, attention-based vision-to-language systems rarely consider information from different modalities (e.g., attending visual modality with query without the information of textual modality), yielding weak understanding when the number of feature type increases. Motivated by the weakness of current attention-based vision-to-language systems, we propose Recurrent Cross-Modality Attention (ReCMA) that utilizes multi-step reasoning and jointly learns attention from multiple modalities in order to have a better understanding of video dialogue.

Figure 2 shows an overview of the proposed ReCMA framework. First, the system encodes both the visual features and the textual features. The visual features are extracted by Two-Stream Inflated 3D ConvNets (I3D) [7], which capture RGB pixel information and spatial flow information from the input video. The video summary, video caption, question, and dialogue history are fed into corresponding LSTM-based encoder to build up textual features. Recurrent Cross-Modality Attention (ReCMA) then attend question with both visual features and textual features to gather question-aware visual representation and question-aware textual representation. Moreover, with the increasing reasoning steps of ReCMA, the model learns the important visual regions and salient textual parts that correspond to the query. When attending one domain feature, the proposed model also attends the heterogeneous domain feature to have full comprehension of the video. Jointly considering both visual and textual representation, a LSTM-based answer-generation decoder then generate an open-end answer that most relevant to the given question, video, and context.

In summary, this work includes two contributions: (1) We introduced an end-to-end model to understand dynamic scenes and conversational dialogue, instead of answering a single question about a static image. (2) We proposed a ReCMA framework that performing multiple reasoning steps to get focused features' representation and developing cross-modality attention to comprehend different modalities' information. Our proposed model enhance 20.8% of CIDEr on the DSTC7 dataset.

II. Related Work

A significant amount of research has been developed to enhance the performance of vision-to-language systems, and these systems can be classified as visual captioning, visual question answering, and visual dialogue. In the following section, we briefly review those related work.

A. Visual Captioning

Image captioning intends to describe the content of an image, and most of the work [3], [8], [9], [12], [20], [22], [43], [69], [79] majorly adopt recurrent neural networks (RNNs) as the core architecture for generating captions and learning long-term visual concepts. RNNs show exceptional results on learning the matching between image patches and single text. Xu et al. [74] proposed a soft attention mechanism [10] on the input image to capture salient regions and improved the performance of image captioning. Implementing attention mechanism also shows promising outcome on much research work [2], [23], [39], [50], [68].

Generating natural language description from images to videos is more challenging because of videos' larger feature

space. Some research [27], [54] comprehends human activities in the video in order to benefit vision-language systems and some work [19], [49], [52], [59], [60] also utilize RNNs to increase long-term memory of the model. Instead of focusing on salient regions on frames in the video, some work [36], [48], [62], [76], [80] also utilizes temporal attention to select the most relevant temporal segments. To conclude, visual captioning introduces visual understanding by generating semantics that well fit the visual features.

B. Visual Question Answering

Instead of generating a description of visual scenes, visual question answering (VQA) uplifts the interaction of vision-to-language systems and humans. Giving a natural language question that targets on visual features, the task is to provide an accurate answer relevant to the question. Visual question answering requires a more precise understanding of visual features and question semantics than producing visual captions. Because systems need to identify the most relevant region in the visual features based on question semantics, attention mechanisms show powerful capability of focusing on the salient regions. A large amount of image-based VQA research [2], [15], [24], [40], [41], [55], [57], [70], [73], [75], [81] conduct systems based on attention mechanism and perform significant result on plenty of image-based VQA datasets [4], [14], [42], [53], [84].

In order to enhance the performance of answering a question for a video, systems need to analyze relevant objects in the frames and memorize temporal events. Such challenging task makes much work [25], [32], [45], [77], [83] design more complicated attention mechanisms to concentrate on the most important part of videos. Furthermore, some research [18], [31], [65] provide video datasets from movie or TV series for systems to output an accurate answer from possible multiple choice. In sum, visual question answering majorly learns to focus on the regions of the visual features and takes the question semantics as guidance.

C. Visual Dialogue

Instead of answering a single question, visual dialogue increases the ability of human-machine interaction for vision-to-language systems. Taking historical question-answer pairs into consideration, models learn both visual representation and conversational semantics in order to answer a question. Das et al. [11] first proposed visual dialogue dataset (VisDial) which contains images from COCO dataset [35] and 1 dialog with 10 question-answer pairs. Most visual dialogue tasks follow the encoder-decoder framework proposed by Sutskever et al. [64]. Several deep neural network architectures have been developed from different aspects, including fusing multi-features [17], generating more human-like responses [71], utilizing ranking discriminator [38], and using conditional probabilistic auto-encoders [44]. Attention mechanisms also play a role in image-based visual dialogue model to capture important regions of visual feature, including reasoning multiple steps on image and dialogue [13], performing dynamic attentions combination [58], recursively increasing visual co-reference resolution [47], and employing a multi-head attention mechanism [21].

Understanding dynamic scenes and conversational semantics for video-based dialogue systems is more challenging than image-based dialogue systems, and one basic reason is the limited availability of such data. Recently, Hori et al. [16] proposed a visual scene-aware dialog dataset on Dialog System Technology Challenge 7 (DSTC7). DSTC7 dataset contains videos from Charades dataset [61] and 1 dialog with 10 question-answer pairs. Different from existing video question answering datasets that selects an answer from multiple choice, DSTC7 dataset provides a free-form answer that yields this task more difficult. Alamri et al. [1] and Schwartz et al. [56] introduce end-to-end models to propose a simple baseline for DSTC7 dataset. The spanning feature space across video frames and historical dialogue make this task important at dealing with multi-modality. In sum, visual dialogue provides promising future for vision-to-language systems and contributes more application in real-world scenarios.

III. Recurrent Cross-Modality Attention

In this section, we propose the detailed interpretation of our end-to-end video dialogue system and introduce how the proposed ReCMA algorithm learns attention from both visual and textual modality.

A. Data Representation

The inputs of proposed video dialogue system are a video V, a video caption C, a video summary S, a dialogue history H, and a question Q. From a raw video V, we extract its RGB feature F_R and spatial flow feature F_F as visual features. The 2048-dimensional F_R and F_F are extracted from the "Mixed_5c" layer of Two-Stream Inflated 3D ConvNets (I3D) [7].

$$F_R, F_F = \text{I3D}(V) \in \mathbb{R}^{n_v \times d_v}, \tag{1}$$

where n_v is the temporal length of a video and d_v is the dimension of feature vector. Both visual features are transformed into new vectors that have the same dimension as the query vector by the single layer perceptron.

$$M_R = \tanh(W_R F_R) \in \mathbb{R}^{n_h \times d_v}, \tag{2}$$

$$M_F = \tanh(W_F F_F) \in \mathbb{R}^{n_h \times d_v}, \tag{3}$$

where W_R and $W_F \in \mathbb{R}^{n_h \times n_v}$ are the matrices of neural weights.

For textual feature of caption C, summary S, dialogue history H, and question Q, corresponding LSTM-based encoders are used to obtain the textual representation. To generate question representation, we find the longest sentence and zero-pad shorter ones for each batch. Every words are embedded by using a linear-embedding layer, followed by a single layer LSTM with dropout. The last hidden state of the LSTM is the question representation $M_Q \in \mathbb{R}^{n_q \times d_q}$, where n_q is the maximal query sentence length for the given batch and d_q is the question embedding dimension. With the same concept as generating question representation, caption and summary separately pass through their own LSTM-net to obtain caption

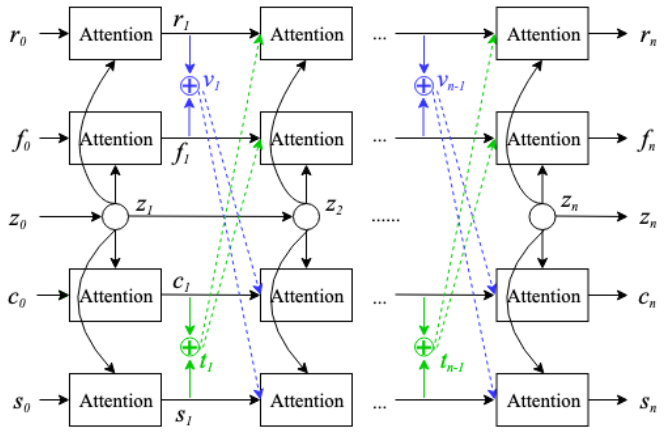


Fig. 3. The proposed recurrent cross-modality attention for multi-step video reasoning. The blue arrows indicate passing joint attended visual feature v_n to textual modality. Likewise, the green arrows represent passing joint attended textual feature t_n to visual modality.

representation $M_C \in \mathbb{R}^{n_c \times d_c}$ and summary representation $M_S \in \mathbb{R}^{n_s \times d_s}$. For the given batch, n_c and n_s represent the maximal sentence length of caption and summary, and d_c and d_s symbolize the dimension of caption and summary. Dialogue history H consists of l-turns question-answer pairs (Q_i, A_i) (i = 1,2,...l), and a LSTM-net also performed to obtain dialogue history representation $M_H \in \mathbb{R}^{n_H \times d_H}$, where n_H and d_H indicate the maximum length and the dimension of the dialogue history snippet.

B. Multi-step ReCMA

The overview of meulti-step recurrent cross-modality attention mechanism are describe in Figure 3. The framework is based on recurrent neural network (RNN), where the hidden state z_n indicates the current question representation and the lower index n is the number of reasoning step. r_n and f_n denote as attended RGB feature and attended spatial flow feature of the video. Likewise, s_n and c_n represent the attended summary and the attended caption of the video. Specifically, we add up r_n and f_n as joint attended visual feature v_n, c_n and s_n also aggregate as joint attended textual feature t_n. Both joint attended features v_n and t_n start after first reasoning step ($n = 1$). Different from attending single-domain modality with query, we found that attending heterogeneous domain modality enhances the performance of video understanding. For instance, we take joint attended textual feature t_n into account when attending RGB feature r_n with the query. When the number of reasoning step increases, ReCMA focuses on the salient region of both visual and textual features by taking the knowledge from cross-modality.

1) Question Self-Attention: The hidden state of ReCMA is the current question representation $z_n \in \mathbb{R}^{1 \times d_q}$, and self-attention is applied to the question representation.

$$a_z = \text{softmax}(p_z \cdot \tanh(W_z z_{n-1}^T)), \tag{4}$$

$$z_n = a_z \cdot z_{n-1}, \tag{5}$$

where the initial hidden state z_0 of RNN is M_Q^T. The attention score of question is $a_z \in \mathbb{R}^{1 \times d_q}$. The matrices of parameter weight are $p_z \in \mathbb{R}^{1 \times d_q}$ and $W_z \in \mathbb{R}^{d_q \times d_q}$.

2) Attending Question and Previous Joint Attended Textual Feature to Visual Features: To find the salient regions of frames in the video, the attended RGB feature $r_n \in \mathbb{R}^{1 \times d_v}$ and attended spatial flow feature $f_n \in \mathbb{R}^{1 \times d_v}$ are updated by their previous state (r_{n-1} and f_{n-1}) and z_n. Furthermore, we pass the previous joint attended textual feature $t_{n-1} \in \mathbb{R}^{1 \times d_v}$ after the first reasoning step in order to use important textual information to find valuable visual information.

$$a_\alpha = \text{softmax}(p_\alpha \cdot \tanh(W_\alpha \alpha_{n-1} + \widetilde{W}_z z_n^T + W_t t_{n-1}^T)), \tag{6}$$

$$\alpha_n = a_\alpha \cdot \alpha_{n-1}, \tag{7}$$

where $\alpha \in \{r, f\}$ is the index of visual components (RGB and spatial flow), and the parameter weight matrices are $p_\alpha \in \mathbb{R}^{1 \times d_v}$, and W_α, $\widetilde{W}_z$, $W_t \in \mathbb{R}^{d_v \times d_v}$. The attention score of visual agent is $a_\alpha \in \mathbb{R}^{1 \times d_v}$. We let the initial visual agent r_0 and f_0 be M_R and M_F. After reasoning step $n = 1$, the system starts to aggregate r_n and f_n as joint attended visual feature v_n, and v_n is delivering to the heterogeneous domain to attend with textual modality.

3) Attending Question and Previous Joint Attended Visual Feature to Textual Features: To generate the important part of context, the attended caption $c_n \in \mathbb{R}^{1 \times d_c}$ and the attended summary $s_n \in \mathbb{R}^{1 \times d_s}$ are updated by attending their previous form (c_{n-1} and s_{n-1}) to z_n. The previous joint attended visual feature v_{n-1} then transfers into textual modality in order to utilize salient visual information to discover important textual information.

$$a_\beta = \text{softmax}(p_\beta \cdot \tanh(W_\beta \beta_{n-1} + \widehat{W}_z z_n^T + W_v v_{n-1}^T)), \tag{8}$$

$$\beta_n = a_\beta \cdot \beta_{n-1}, \tag{9}$$

where $\beta \in \{c, s\}$ is the index of textual components (caption and summary), and the matrices of parameter weight are $p_\beta \in \mathbb{R}^{1 \times d_\beta}$, and W_β, $\widehat{W}_z$, $W_v \in \mathbb{R}^{d_\beta \times d_\beta}$. The attention score of textual agent is $a_\beta \in \mathbb{R}^{1 \times d_\beta}$. The initial textual agents c_0 and s_0 are set to be M_C and M_S. Moreover, the system starts to add up c_n and s_n as joint attended textual feature $t_n \in \mathbb{R}^{1 \times d_\beta}$ when reasoning step greater than or equal to 1, and t_n then passes to visual modality as an additional information.

C. Answer Generation Decoder

After performing proposed ReCMA, the system concatenates all attended features r_n, f_n, c_n, and s_n as the context vector g. In Figure 4, a generative LSTM-based decoder is used to decode the context vector into an answer $y = (y_1, y_2, ..., y_L)$, where L is the number of word, and $y_\ell \in \Upsilon_\ell = \{1, 2, ..., |\Upsilon_\ell|\}$ represents the a vocabulary of possible words Υ_ℓ. An FC-layer with dropout and softmax performed after the answer generation LSTM-decoder to compute the conditional probability $p(y_\ell | x, y_{\ell-1}, h_{\ell-1})$ for possible word y_ℓ, where $h_{\ell-1}$ denotes the previous hidden state of the decoder, and context vector g serves as the initial hidden state of the RNN, i.e. h_0 = g.

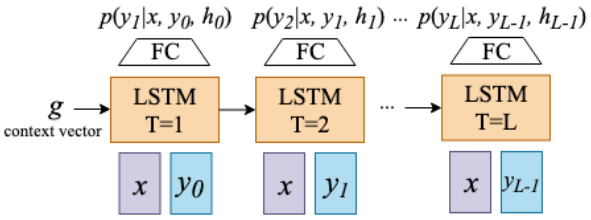


Fig. 4. The answer generation decoder of proposed end-to-end video dialogue system. The context vector acts as initial state of the LSTM-net. The input is the concatenation of x and y_{T-1}, where concatenating question representation with dialogue history representation indicates x, and y_{T-1} represents the previous answer word.

	Training	Validation	Test
# of Dialogs	7,659	1,787	1,710
# of Turns	153,180	35,740	13,490
# of Words	1,450,754	339,006	110,252

TABLE I
THE DATA DISTRIBUTION OF DSTC7 DATASET

IV. EXPERIMENTS AND RESULTS

To test the performance of proposed ReCMA, we conduct several experiments and describe detailed analysis on DSTC7 dataset [16]. From Charades video dataset [61], DSTC7 dataset proposes video caption, video summary, and 1 dialog with 10 question-answer pairs. The dialogue was generated by two Amazon Mechanical Turk workers who had a discussion about events in the video. Table I summarizes the data distribution of DSTC7 dataset.

A. Baselines for comparison

We first compare our model with AVSD-baseline [1], which is a naive baseline provided by the organizers. We also compare our performance with other participants that compete in this AVSD challenge. Kumar et al. [29] implemented topics of the dialog into the architecture and performed multimodal attention. Yeh et al. [78] introduced a fusion technique that well integrates multimodal features. Le et al. [30] proposed a hierarchical attention approach and applied a nonlinear feature fusion technique to combine the visual and audio features. Lin et al. [34] proposed an entropy-enhanced dynamic memory network to effectively model video modalities.

B. Experimental setup and evaluation metrics

In the training process, the dimension of textual and visual feature are set to 128 and 2048. Each text is transferred into a vector by GloVe [51]. We use the Adam optimizer [26] with a learning rate 0.001, a batch size of 32, and a drop out rate [63] 0.2. The hyper-parameters were determined by optimizing the cross-entropy loss between prediction and target. We evaluate the performance of our model by using 4 automatic evaluation metrics: BLEU score, METEOR [6], ROUGE-L [33], and CIDEr [66].

	B-1	B-2	B-3	B-4	M	R	C
reasoning step n = 1							
Q+H	0.623	0.478	0.374	0.295	0.220	0.492	0.775
Q+H+C	0.632	0.486	0.381	0.303	0.238	0.518	0.896
Q+H+S	0.628	0.483	0.378	0.302	0.238	0.518	0.889
Q+H+rgb	0.636	0.484	0.390	0.312	0.234	0.517	0.882
Q+H+flow	0.641	0.493	0.392	0.306	0.233	0.520	0.895
Q+H+C+S	0.644	0.488	0.383	0.302	0.238	0.518	0.891
Q+H+rgb+flow	0.648	0.499	0.390	0.309	**0.240**	0.520	0.890
Q+H+C+S+rgb+flow	**0.657**	**0.510**	**0.400**	**0.318**	0.238	**0.527**	**0.911**

TABLE II
RESULTS FOR EVALUATING THE ROBUSTNESS OF EACH FEATURE USING OBJECTIVE EVALUATION METRICS. B-i DENOTES AS THE i-GRAM PRECISION SCORE OF BLEU METRIC. METEOR, ROUGE-L, AND CIDER STAND FOR M, R, AND C IN THE TABLE

C. Robustness of modalities

To fully analyze the influence of the multi-modal features to video dialogue task, we start from inputting mono-type feature then adding other features. We first consider current question Q and conversational dialogue history H, so the simplest input representation is concatenating M_H with M_Q as the context vector x, without the information of videos. Both textual feature are encoded using a word embedding and a single layer LSTM-net. Only taking question and dialogue history, the model trained by this simplest input is denoted as Q+H in Table II.

In order to improve video understanding, we then add the video-related features which are RGB, spatial flow, caption, and summary of videos as the third input. The models that individually add video caption, video summary, RGB, and spatial flow are represented as Q+H+C, Q+H+S, Q+H+rgb, and Q+H+flow respectively in Table II. To test the performance of each feature, the third feature then pass through ReCMA with question representation M_Q. Attending individual feature to question representation one time, the reasoning step n of ReCMA is set to be 1. Therefore, the context vector x of Q+H+C, Q+H+S, Q+H+rgb, and Q+H+flow is c_1, s_1, r_1, and f_1 respectively.

Because most of the current question Q is asking what happened in the video and is hard to generate answer from previous dialogue, all models with video-related features as the third input reasonably outperform simple model Q+H. Moreover, we observe that the models with visual features (Q+H+rgb and Q+H+flow) have better performance than the models with textual features (Q+H+C and Q+H+S). The caption and summary for each video in DSTC7 dataset only have approximate 2 sentences, so visual features are more informative than textual features when answering a given question.

After analyzing the models with additional mono-type feature, we then evaluate the performance of the model combining different features. With one reasoning step, Q+H+C+S in the third part of Table II take textual features (caption and summary) into account. To be more specific, the context vector x of Q+H+C+S is the concatenation of c_1 and s_1. Likewise, Q+H+rgb+flow considers visual features (RGB and optical flow) in first reasoning step, and the context vector x of this model is the concatenation of r_1 and f_1. The results show that the models combining two features

(Q+H+C+S and Q+H+rgb+flow) have a better performance than the models with additional mono-type feature. Examining textual domain, Q+H+C+S slightly outperforms both Q+H+C and Q+H+S. Moreover, Q+H+rgb+flow surpasses both Q+H+rgb and Q+H+flow for visual domain. We observe that the model combining visual features (Q+H+rgb+flow) exhibit better performance than the model combining textual features (Q+H+C+S). Similar to the results of models with additional mono-type feature, we think that visual features will help our system to generate better responses.

D. Recurrent reasoning step

In order to fully unitize the information provided by DSTC7 dataset, we select all features to build our proposed end-to-end model. Furthermore, we aim at measuring the video understanding performance of proposed ReCMA in different reasoning step. Including current question Q and dialogue history H, all video-related features, which are RGB, spatial flow, caption, and summary of videos, are added. We first consider attending each feature to question one time, where reasoning step equals to 1, and the context vector (x is the concatenation of c_1, s_1, r_1, and f_1. The model that utilize every feature representations (M_R, M_F, M_C, M_S, M_Q, and M_H) is denoted as Q+H+C+S+rgb+flow in Table III. Though without multiple reasoning steps, the model Q+H+C+S+rgb+flow also all models in Table II, and this outcome reveals the usefulness of all features proposed by DSTC7 dataset.

After the first reasoning step (n = 1), proposed ReCMA targets on more specific regions of textual representation and visual representation corresponding to the input question. In order to learn the important regions from heterogeneous domain, we proposed joint attended features (v_n and t_n). Both joint attended features are designed by aggregating homogeneous attended features after first reasoning step, $i.e.$, joint attended visual feature v_n equals to attended RGB feature r_n plus attended spatial flow feature f_n. However, in order to test the effectiveness of proposed joint attended features, we first evaluate ReCMA without them. By eliminating the term of t^T_{n-1} and v^T_{n-1} in Equation (6) and (8), the model re-attend attended features to question representation in different reasoning steps. The result in Table III shows that even increasing 1 reasoning step (n = 2), ReCMA without joint attended features improves the performance compared with Q+H+C+S+rgb+flow model. Furthermore, the performance further improves along with the reasoning step n increases, and the outcomes show the success of multiple reasoning steps. Nevertheless, for reasoning step n greater than 3 , the model did not show significant increase on every metrics.

Inspired by finding important information of current domain from the salient regions of a heterogeneous modality, we add joint attended features (v_n and t_n) into ReCMA when reasoning step increases. Take Figure 3 as an example, when generating r_2 and f_2, the attention mechanism also takes joint attended textual feature t_1 (combined by c_1 and s_1) into account. Likewise, even increasing 1 reasoning step (n = 2), the performance of ReCMA surpasses Q+H+C+S+rgb+flow model. Moreover, comparing to ReCMA without joint attended features, ReCMA preforms better when the reasoning step n is the same. The result shows that adding important information from cross-domain helps the model fully understand videos and generate accurate answers. Moreover, ReCMA also shows steady and relative improvement over the baseline when reasoning step n increases. Unlike ReCMA without joint attended features, the accuracy of ReCMA still consistently increases in higher reasoning step (e.g., n exceeds 3). However, after the fifth reasoning step, the model only achieves slight improvement.

	B-1	B-2	B-3	B-4	M	R	C
ReCMA w/ reasoning step n = 1							
Q+H+C+S+rgb+flow	0.657	0.510	0.400	0.318	0.238	0.527	0.911
ReCMA w/o joint attended features (v_n and t_n)							
n = 2	0.657	0.511	0.402	0.323	0.238	0.524	0.917
n = 3	0.660	0.515	0.406	**0.324**	0.241	**0.528**	**0.93**
n = 4	0.656	0.508	0.400	0.318	**0.242**	0.527	0.922
n = 5	**0.662**	**0.516**	**0.407**	**0.324**	0.241	0.527	0.911
ReCMA w/ joint attended features (v_n and t_n)							
n = 2	0.658	0.513	0.406	0.325	0.239	0.523	0.917
n = 3	0.663	0.517	0.408	0.327	0.239	0.527	0.917
n = 4	0.662	0.517	0.412	0.333	0.242	0.532	0.935
n = 5	**0.667**	**0.521**	**0.413**	**0.334**	**0.242**	**0.533**	**0.941**

TABLE III
RESULTS FOR PROPOSED RECMA WITH INCREASING REASONING STEP. THE BEST RESULT FOR RECMA WITH AND WITHOUT JOINT ATTENDED FEATURE IN EACH METRIC IS HIGHLIGHTED IN BOLD.

	B-1	B-2	B-3	B-4	M	R	C
AVSD-baseline [1]	0.621	0.480	0.379	0.305	0.217	0.481	0.733
Kumar et al. [29]	0.274	0.175	0.121	0.087	0.117	0.294	0.789
Yeh et al. [78]	0.640	0.513	0.416	0.342	0.223	0.504	0.837
Le et al. [30]	0.633	0.490	0.386	0.310	0.242	0.515	0.856
Lin et al. [34]	0.410	0.493	0.388	0.310	0.241	0.527	0.912
ReCMA($n = 5$)	**0.667**	**0.521**	0.413	0.334	**0.242**	**0.533**	**0.941**

TABLE IV
THE COMPARISON OF PROPOSED RECMA WITH OTHER STATE-OF-THE-ART BASELINES.

E. Quantitative results with state-of-the-art baselines

Table IV shows the automatic evaluation result of our proposed method and other state-of-the-art baselines, and our model improves on most of metrics compared with others. The result shows the usefulness of proposed ReCMA and we believe performing attention mechanism on heterogeneous domain with higher reasoning steps can also benefit other models.

To show the effectiveness of joint attended features (v_n and t_n) in multiple reasoning process, we conduct ablation studies and the results are shown in Table V. By considering the fifth reasoning step, the result of taking only one modality feature (v_n or t_n) in each reasoning step lie between the result with and without both features. The results justify our assumption that using both visual and textual modality in attention mechanism would benefit dialogue systems' performance. In sum,

	B-1	B-2	B-3	B-4	M	R	C
ReCMA w/ reasoning step $n = 5$							
w/o v_n and t_n	0.662	0.516	0.407	0.324	0.241	0.527	0.911
w/o t_n	0.662	0.517	0.410	0.327	0.230	0.525	0.923
w/o v_n	0.664	0.519	0.409	0.330	0.233	0.527	0.930
w/ v_n and t_n	**0.667**	**0.521**	**0.413**	**0.334**	**0.242**	**0.533**	**0.941**

TABLE V
ABLATION STUDIES OF JOINT ATTENDED FEATURES

	B-1	B-2	B-3	B-4	M	R	C	Human
AVSD-baseline [16]	0.614	0.467	0.365	0.289	0.210	0.480	0.651	2.885
ReCMA($n = 5$)	**0.645**	**0.504**	**0.402**	**0.324**	**0.232**	**0.521**	**0.875**	**3.123**

TABLE VI
RELEASED BY THE AVSD ORGANIZERS, THE FINAL RESULT OF OBJECTIVE EVALUATION AND HUMAN RATING ON DSTC8-AVSD TEST SET.

our best performance (ReCMA with joint attended features at reasoning step n= 5) improve relative 20.8% improvement over baseline on CIDEr metric.

F. Qualitative analysis

Table VI shows the evaluation result on DSTC8-AVSD test set. The result was released by AVSD organizers. Both automatic evaluation metrics and human rating of proposed ReCMA outperforms baseline. Figure 5 shows the ground truth proposed by DSTC7 dataset and the answers generated by AVSD-baseline model and proposed ReCMA model. The answers generated by proposed ReCMA model illustrate that multiple reasoning steps benefit the inference process and hence lead to accurate answers of questions. For example, the proposed model can focus on the people in the frame and correctly answer the number and gender of people in the dynamic scenes video. Compared with "*the same position*" generated by the baseline model, the question "*does she goes out of washroom ?*" is provided with a more precise answer "*in the bathroom*" by the proposed model. Moreover, baseline model cannot capture instant event in the video, but our model can focus on instant emotion of people. When the question ask temporal-related issue, like " *was he present at the beginning ?*" or " *the video ends with him sitting in a chair watching tb right ?*", proposed ReCMA model can also answer precisely compared with baseline model. Figure 6 shows and example of proposed ReCMA with different reasoning step. Though the answer generated by ReCMA with lower reasoning step (e.g, $n = 2$ or 3) correctly answer the question, the answer generated by ReCMA with higher reasoning step (e.g, $n = 4$ and 5) specifically answer "*sitting on the chair*".

In order to fully comprehend video dialogue task, we did some analysis on DSTC dataset. Though DSTC7 dataset is promising and challenging, the quality of the dataset need to be improved. Some questions are challenging to answer and the ground truth answer provides an ambiguous answer. For example, a question "*is he at work ?*" is answered with "*hard to say, he is sitting in the hallway by himself*". Moreover, lot of to-be-answered questions in the training data ask additional information, such as "*anything else that i need to know ?*" is answered with "*no, that is all that happens*". Furthermore, the reference sometimes gives the answer outside the question. For instance, the question is "*is the man sitting on a chair ?*", and proposed ReCMA answer "*yes, the man is sitting on a chair*". However, the ground truth answer "*yes, the man sits on the couch staring at the tv for a long duration*", it not only subjects to the word "*sitting*" of the question but also provides more irrelevant information. If the dataset precise questions and answers, the model can have clear a understanding of the video. Therefore, we believe the quality of the dataset requires enhancement, but the effort of DSTC7 dataset can not be discounted.

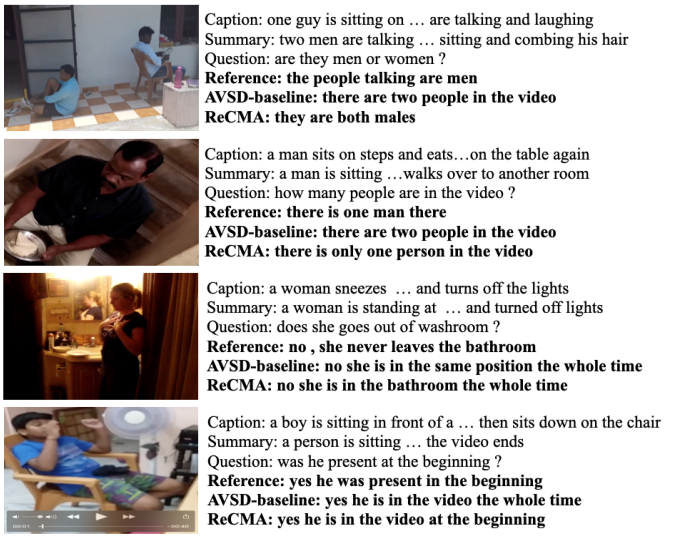


Fig. 5. Some examples of the answers generated by proposed ReCMA (n = 5) and the reference answers given by DSTC7 dataset. For simplicity, only parts of video caption and video summary are shown. The results show that our ReCMA not only can comprehend in-frame events but also answer the temporal questions.

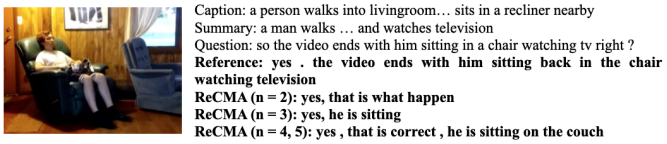


Fig. 6. An example of the answers generated by proposed ReCMA with different reasoning step. The results show that ReCMA with higher reasoning step can provide more specific answer.

V. CONCLUSION

We presented an end-to-end video dialogue system to converse about videos and understand dynamic scenes. We also showed that proposed ReCMA, a recurrent cross-modality attention, can take cross-modality information to enhance attention mechanism. Through multiple reasoning steps of ReCMA, the model achieved a better comprehension of multimodal context, thus boosting video question answering performance over state-of-the-art baseline. We evaluated proposed ReCMA on DSTC7 dataset, where ReCMA achieved a relative 20.8% improvement over the baseline on CIDEr metric. In this paper, we also demonstrated the effectiveness of attention mechanism for video dialogue system and the usefulness of each modalities on DSTC7 dataset. Video question answering is a new and promising research area, and a possible improvement to our work is adding pre-trained word embedding such as BERT to improve the semantic understanding of the model.

VI. ACKNOWLEDGEMENT

This research is supported by Ministry of Science and Technology, Taiwan under the project contract 108-2221-E-001-012-MY3 and 109-2221-E-001-015-.

REFERENCES

[1] Huda AlAmri, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Jue Wang, Irfan Essa, Dhruv Batra, Devi Parikh, Anoop Cherian, Tim K. Marks, and Chiori Hori. Audio visual scene-aware dialog (AVSD) challenge at DSTC7. *CoRR*, abs/1806.00525, 2018.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998, 2017.

[3] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. *CoRR*, abs/1711.09151, 2017.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[7] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017.

[8] Moitreya Chatterjee and Alexander G. Schwing. Diverse and coherent paragraph generation from images. *CoRR*, abs/1809.00681, 2018.

[9] Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *CoRR*, abs/1411.5654, 2014.

[10] KyungHyun Cho, Aaron C. Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *CoRR*, abs/1507.01053, 2015.

[11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. *CoRR*, abs/1611.08669, 2016.

[12] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389, 2014.

[13] Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. *CoRR*, abs/1902.00579, 2019.

[14] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. *CoRR*, abs/1505.05612, 2015.

[15] Peng Gao, Hongsheng Li, Haoxuan You, Zhengkai Jiang, Pan Lu, Steven Hoi, and Xiaogang Wang. Dynamic fusion with intra- and inter- modality attention flow for visual question answering. *CoRR*, abs/1812.05252, 2018.

[16] Chiori Hori, Huda AlAmri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Irfan Essa, Dhruv Batra, and Devi Parikh. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. *CoRR*, abs/1806.08409, 2018.

[17] Unnat Jain, Svetlana Lazebnik, and Alexander G. Schwing. Two can play this game: Visual dialog with discriminative question generation and answering. *CoRR*, abs/1803.11186, 2018.

[18] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. *CoRR*, abs/1704.04497, 2017.

[19] Qin Jin, Jia Chen, Shizhe Chen, Yifan Xiong, and Alexander Hauptmann. Describing videos using multi-modal fusion. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 1087–1091, New York, NY, USA, 2016. ACM.

[20] Justin Johnson, Andrej Karpathy, and Fei-Fei Li. Densecap: Fully convolutional localization networks for dense captioning. *CoRR*, abs/1511.07571, 2015.

[21] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. *CoRR*, abs/1902.09368, 2019.

[22] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

[23] Mahmoud Khademi and Oliver Schulte. Image caption generation with hierarchical contextual visual spatial attention. pages 2024–20248, 06 2018.

[24] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *CoRR*, abs/1805.07932, 2018.

[25] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story QA by deep embedded memory networks. *CoRR*, abs/1707.00836, 2017.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *Int. J. Comput. Vision*, 50(2):171–184, November 2002.

[28] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. *CoRR*, abs/1809.01816, 2018.

[29] Shachi H. Kumar, Eda Okur, Saurav Sahay, Juan Jose Alvarado Leanos, Jonathan Huang, and Lama Nachman. Context, attention and audio feature explorations for audio visual scene-aware dialog. *CoRR*, abs/1812.08407, 2018.

[30] Hung Le, Doyen Sahoo, Nancy F. Chen, and Steven C.H. Hoi. Hierarchical multimodal attention for end-to-end audio-visual scene-aware dialogue response generation. *Computer Speech Language*, 63:101095, 2020.

[31] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: localized, compositional video question answering. *CoRR*, abs/1809.01696, 2018.

[32] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander G. Hauptmann. Focal visual-text attention for visual question answering. *CoRR*, abs/1806.01873, 2018.

[33] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[34] Kuan-Yen Lin, Chao-Chun Hsu, Yun-Nung Chen, and Lun-Wei Ku. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. *CoRR*, abs/1908.08191, 2019.

[35] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[36] Xiang Long, Chuang Gan, and Gerard de Melo. Video captioning with multi-faceted attention. *CoRR*, abs/1612.00234, 2016.

[37] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, abs/1506.08909, 2015.

[38] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. *CoRR*, abs/1706.01554, 2017.

[39] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. *CoRR*, abs/1612.01887, 2016.

[40] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061, 2016.

[41] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. *CoRR*, abs/1711.06794, 2017.

[42] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *CoRR*, abs/1410.0210, 2014.

[43] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Explain images with multimodal recurrent neural networks. *CoRR*, abs/1410.1090, 2014.

[44] Daniela Massiceti, N. Siddharth, Puneet Kumar Dokania, and Philip H. S. Torr. Flipdial: A generative model for two-way visual dialogue. *CoRR*, abs/1802.03803, 2018.

[45] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. *CoRR*, abs/1709.09345, 2017.

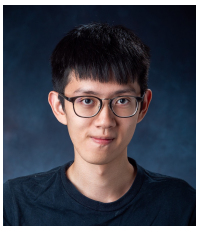
[46] Medhini Narasimhan and Alexander G. Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. *CoRR*, abs/1809.01124, 2018.

[47] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. *CoRR*, abs/1812.02664, 2018.

[48] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. *CoRR*, abs/1511.03476, 2015.
[49] Ramakanth Pasunuru and Mohit Bansal. Multi-task video captioning with video and entailment generation. *CoRR*, abs/1704.07489, 2017.
[50] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. *CoRR*, abs/1612.01033, 2016.
[51] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
[52] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. Multimodal video description. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 1092–1096, New York, NY, USA, 2016. ACM.
[53] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *CoRR*, abs/1505.02074, 2015.
[54] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. *2013 IEEE International Conference on Computer Vision*, pages 433–440, 2013.
[55] Idan Schwartz, Alexander Schwing, and Tamir Hazan. High-order attention models for visual question answering. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3664–3674. Curran Associates, Inc., 2017.
[56] Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. *CoRR*, abs/1904.05876, 2019.
[57] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G. Schwing. Factor graph attention. *CoRR*, abs/1904.05880, 2019.
[58] Paul Hongsuck Seo, Andreas M. Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. *CoRR*, abs/1709.07992, 2017.
[59] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly supervised dense video captioning. *CoRR*, abs/1704.01502, 2017.
[60] Rakshith Shetty and Jorma Laaksonen. Frame- and segment-level features and candidate pool evaluation for video caption generation. *CoRR*, abs/1608.04959, 2016.
[61] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016.
[62] Jingkuan Song, Xiangpeng Li, Lianli Gao, and Heng Tao Shen. Hierarchical lstms with adaptive attention for visual captioning. *CoRR*, abs/1812.11004, 2018.
[63] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.
[64] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.
[65] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. *CoRR*, abs/1512.02902, 2015.
[66] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014.
[67] Oriol Vinyals and Quoc V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015.
[68] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.
[69] Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. *CoRR*, abs/1711.07068, 2017.
[70] Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. Object-difference attention: A simple relational attention for visual question answering. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 519–527, New York, NY, USA, 2018. ACM.
[71] Qi Wu, Peng Wang, Chunhua Shen, Ian D. Reid, and Anton van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. *CoRR*, abs/1711.07613, 2017.
[72] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *CoRR*, abs/1603.01417, 2016.
[73] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *CoRR*, abs/1511.05234, 2015.
[74] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.
[75] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015.
[76] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4507–4515, Washington, DC, USA, 2015. IEEE Computer Society.
[77] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video question answering via attribute-augmented attention network learning. *CoRR*, abs/1707.06355, 2017.
[78] Yi Ting Yeh, Tzu-Chuan Lin, Hsiao-Hua Cheng, Yu-Hsuan Deng, Shang-Yu Su, and Yun-Nung Chen. Reactive multi-stage feature fusion for multimodal dialogue modeling. *CoRR*, abs/1908.05067, 2019.
[79] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *CoRR*, abs/1603.03925, 2016.
[80] Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. *CoRR*, abs/1707.06029, 2017.
[81] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *CoRR*, abs/1708.01471, 2017.
[82] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3518–3524, 2017.
[83] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3683–3689. International Joint Conferences on Artificial Intelligence Organization, 7 2018.
[84] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. *CoRR*, abs/1511.03416, 2015.

Yun-Wei Chu Yun-Wei Chu received the M.S. degree in electrical control engineering from National Chiao Tung University in 2017 and the B.S. degree in electrical engineering from National Chi Nan University in 2015. He is currently a Ph.D. student in Electrical and Computer Engineering at Purdue University. Prior to this, he was a research assistant working on natural language processing and vision-to-language processing at Academia Sinica, Taiwan. His research interests are social learning networks and natural language processing.

Kuan-Yen Lin Kuan-Yen Lin received the B.E. in Bio-Industrial Mechatronics Engineering from National Taiwan University in 2019 and received the M.Eng. in Computer Science from Cornell University in 2020. Her research interests include machine learning and natural language processing.



Chao-Chun Hsu Chao-Chun Hsu received the B.S. degree in Computer Science and Information Engineering from National Taiwan University in 2017. From 2017 to 2019, he was the research assistant working on natural language processing at Academia Sinica, Taiwan. He is currently a Ph.D. student in Computer Science at the University of Chicago. His research interests are human-centered machine learning and natural language processing.

Lun-Wei Ku Lun-Wei Ku is now an associate research fellow in Institute of Information Science, Academia Sinica, adjunct associate professor of National Yang Ming Chiao Tung university (NYCU), and the secretary-general of Association for Computational Linguistics and Chinese Language Processing (ACLCLP). She received her M.S. and Ph.D. degrees from Department of Computer Science and Information Engineering, National Taiwan University. Her research interests include natural language processing, information retrieval, and computational linguistics. She has been working on sentiment analysis since year 2005 and was the co-organizer of NTCIR MOAT Task (Multilingual Opinion Analysis Task, traditional Chinese side) from year 2006 to 2010. Her international recognition includes Good Design Award Selected (2012), CyberLink Technical Elite Fellowship (2007), IBM Ph.D. Fellowship (2008), and ROCLING Doctorial Dissertation Distinction Award (2009). Other professional international activities she involved include: General Chair, StarSem 2021, Program Chair, StarSem 2019 and ARIS 2019, Best Paper Committee, ACL 2019; Student Workshop Chair, AACL-IJCNLP; Area Chair, ACL 2021, NAACL 2021, ACL 2020, COLING 2020, EMNLP 2019, ACL 2017, CCL 2016, NLPCC 2016, ACL-IJCNLP 2015 and EMNLP 2015; Financial Chair, IJCNLP 2017; Publication Co-Chair, IJCNLP 2013; Publicity Chair, AIRS 2010. She is also active in industrial collaborations and currently working with banks and medical companies to improve their NLP technology.