DMRM: A Dual-Channel Multi-Hop Reasoning Model for Visual Dialog

Feilong Chen,[1,2,3,4*] **Fandong Meng,**[2] **Jiaming Xu,**[1,3†] **Peng Li,**[2] **Bo Xu,**[1,3,4,5] **Jie Zhou**[2]
[1]Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China.
[2]Pattern Recognition Center, WeChat AI, Tencent Inc., China
[3]Research Center for Brain-inspired Intelligence, CASIA
[4]University of Chinese Academy of Sciences
[5]Center for Excellence in Brain Science and Intelligence Technology, CAS. China
{chenfeilong2018, jiaming.xu, xubo}@ia.ac.cn
{fandongmeng, patrickpli, withtomzhou}@tencent.com

Abstract

Visual Dialog is a vision-language task that requires an AI agent to engage in a conversation with humans grounded in an image. It remains a challenging task since it requires the agent to fully understand a given question before making an appropriate response not only from the textual dialog history, but also from the visually-grounded information. While previous models typically leverage single-hop reasoning or single-channel reasoning to deal with this complex multimodal reasoning task, which is intuitively insufficient. In this paper, we thus propose a novel and more powerful Dual-channel Multi-hop Reasoning Model for Visual Dialog, named DMRM. DMRM synchronously captures information from the dialog history and the image to enrich the semantic representation of the question by exploiting dual-channel reasoning. Specifically, DMRM maintains a dual channel to obtain the question- and history-aware image features and the question- and image-aware dialog history features by a mulit-hop reasoning process in each channel. Additionally, we also design an effective multimodal attention to further enhance the decoder to generate more accurate responses. Experimental results on the VisDial v0.9 and v1.0 datasets demonstrate that the proposed model is effective and outperforms compared models by a significant margin.

Introduction

With the rapid development of both computer vision and natural language processing, visual-language tasks such as image caption (Xu et al. 2015; Anderson et al. 2016; 2018) and visual question answering (Ren, Kiros, and Zemel 2015; Gao et al. 2015; Lu et al. 2016; Anderson et al. 2018) have attracted increasing attention in recent years. Although these tasks have inspired tremendous efforts on integrating vision and language to develop smarter AI, they are mostly *single-round* while human conversations are generally multi-round. Therefore, the Visual Dialog task is proposed to encourage research on multi-round visually-grounded dialog by Das et al. (2017).

In Visual Dialog, an agent is required to answer a question given the dialog history and the visual context. In order to make an appropriate response, it is necessary for the agent to gain a proper understanding of the question, which requires it to exploit the textual dialog history and the visual context. To this end, some studies (Das et al. 2017; Lu et al. 2017a) design models to obtain features from both modalities. Das et al. (2017) propose Late Fusion (LF), which directly concatenates individual representations of the question, the dialog history, and the image, and then generates a new joint representation by a linear transformation on them. Lu et al. (2017a) design a history-conditioned attention image encoder to generate the representation of the question, the question-aware dialog history and the history-conditioned image features, and then concatenate them to be joint representations.

Nevertheless, the approaches mentioned above are of single-hop approaches, which show the limited ability of reasoning and neglect latent information of the interactions among the question, the dialog history and the image. For better solutions, researchers (Das et al. 2017; Wu et al. 2018; Niu et al. 2019; Kang, Lim, and Zhang 2019) investigate multi-hop reasoning approaches (Hudson and Manning 2018; Hu et al. 2018) to conduct interactions among modalities. For example, Wu et al. (2018) provide a sequential co-attention encoder, which firstly obtains question-aware image features, secondly extracts history features by co-attention mechanism with the question features and the extracted image features, thirdly gets the attended question features by the extracted history features and image features, and finally joints the three attended features and send them to the decoder. Niu et al. (2019) propose a recursive visual attention model, which recursively reviews the dialog history to find the reference of the question, and then extracts the image features by the attention model with the extracted history context and the question. These approaches are of single-channel approaches, which firstly use the question to find reference from the dialog history and then extract the image context from both the question and the history context. However, humans usually deal with a visually-grounded, multi-turn dialog by simultaneously comprehending the two aspects of information, namely both the textual

*This work was done when Feilong Chen was interning at Pattern Recognition Center, WeChat AI, Tencent Inc., China
†Corresponding author.

dialog history and the visual context. That is to say, the question can find reference first from the image and then form the dialog history to enrich the question representation, and vice versa.

Dual-channel reasoning, i.e., acquiring information from the dialog history and the image simultaneously, is beneficial for gaining an original understanding of the question from the dialog history and the image. Meanwhile, multi-hop reasoning, i.e., reasoning among the question, the dialog history and the image, is conducive to utilizing abundant latent information among the three inputs. Therefore, in this paper, we propose a Dual-channel Multi-hop Reasoning Model for Visual Dialog, named DMRM. DMRM synchronously captures information from the dialog history and the image to enrich the semantic representation of the question by exploiting dual-channel reasoning, which is composed of a Track Module and a Locate Module. Track Module aims to enrich the representation of the question from the visual information while Locate Module aims to reach the same goal from the textual dialog history. Specifically, DMRM maintains dual channels to obtain the question- and history-aware image features and the question- and image-aware dialog history features by a mulit-hop reasoning process in each channel. In addition, we design an effective multimodal attention to further enhance the decoder to generate more accurate responses.

We validate the DMRM model on large-scale datasets: VisDial v0.9 and v1.0 (Das et al. 2017). DMRM achieves the state-of-the-art results on some metrics compared to other methods. We also conduct ablation studies to demonstrate the effectiveness of our proposed components. Furthermore, we conduct the human evaluation to indicate the effectiveness of our model in inferring answers.

Our main contributions are threefold:

- We propose a dual-channel multi-hop reasoning model to deal with this complex multimodal reasoning task which enriches the semantic representation of the question, and thus the agent can make an appropriate response.
- We are the first to apply multimodal attention to the decoder for visual dialog and demonstrate the necessity and effectiveness of this attention mechanism for the decoding of visual dialog.
- We evaluate our method on two large-scale datasets and conduct ablation studies, human evaluation. Experimental results on VisDial v0.9 and v1.0 demonstrate that the proposed model achieves the state-of-the-art results on some metrics[1].

Our Approach

In this section, we formally describe the visual dialog task and our proposed method, Dual-channel Multi-hop Reasoning Model (DMRM). According to Das et al.(2017), inputs of a visual dialog agent consist of an image I, a caption C describing the image, a dialog history (question-answer pairs) till round $t-1$: $H = (\underbrace{C}_{H_0}, \underbrace{(Q_1, A_1)}_{H_1}, \ldots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})$ and the current question Q_t at round t. The goal of the visual dialog agent is to generate a response A_t to the question Q_t.

Given the problem setup, DMRM for visual dialog consists of four components: (1) Input Representation, where the representations of the image and the textual information are generated for reasoning; (2) Dual-channel Multi-hop Reasoning, where our reasoning is applied to encode input representations; (3) Multimal Fusion, where we fuse the multimodal information; (4) Generative Decoder, where we use our multimodal attention decoder to generate the response. Specifically, we use Track Module and Locate Module to implement our dual-channel multi-hop reasoning. As shown in Figure 1, Track Module aims to enrich the representation of the question from the visual information by exploiting the question and the dialog history. Locate Module aims to enrich the representation of the question from the textual dialog history by exploiting the question and the image. Answer decoder takes the outputs of Track Module and Locate module as inputs, and generates an appropriate response.

We first introduce the representations of inputs (both image features and the language features). Then we describe the detailed architectures of dual-channel multi-hop reasoning and multimodal fusion operation. Finally, we present the multimodal attention answer decoder.

Input Representation

Image Features We use a pre-trained Faster R-CNN (Ren et al. 2015) to extract object-level image features. Specifically, the image features v for the image I are represented by:

$$v = \text{Faster R} - \text{CNN}(I) \in \mathbb{R}^{K \times V}, \tag{1}$$

where K denotes the total number of the object detection features per image and V denotes the dimension of each features, respectively. We extract the object features by using a fixed number K.

Language Features We first embed each word in the current question Q_t to $\{w_{t,1}, \ldots, w_{t,L}\}$ by using pre-trained Glove embeddings (Pennington, Socher, and Manning 2014), where L denotes the number of tokens in Q_t. We use a one-layer BiLSTM to generate a sequence of hidden states $\{x_{t,1}, \ldots, x_{t,L}\}$. We use the last hidden state of the BiLSTM as question features $q_t \in \mathbb{R}^L$ as follows:

$$\overrightarrow{x_{t,j}} = \text{LSTM}_f(w_{t,j}, x_{t,j-1}),\ j \in \{0, \ldots, L-1\}, \tag{2}$$

$$\overleftarrow{x_{t,j}} = \text{LSTM}_b(w_{t,j}, x_{t,j+1}),\ j \in \{L-1, \ldots, 0\}, \tag{3}$$

$$q_t = [\overrightarrow{x_{t,L-1}}, \overleftarrow{x_{t,0}}], \tag{4}$$

Also, each question-answer pair in the dialog history $H = \{H_0, H_1, \ldots, H_{t-1}\}$ and the answer A_t are embedded in the same way as the current question, yielding the dialog history features $u = \{u_0, u_1, \ldots, u_{t-1}\}$ and the answer features a_t. Q_t, H and A_t are embedded with the same word embedding vectors but three different BiLSTMs.

[1]Code is available at https://github.com/phellonchen/DMRM.

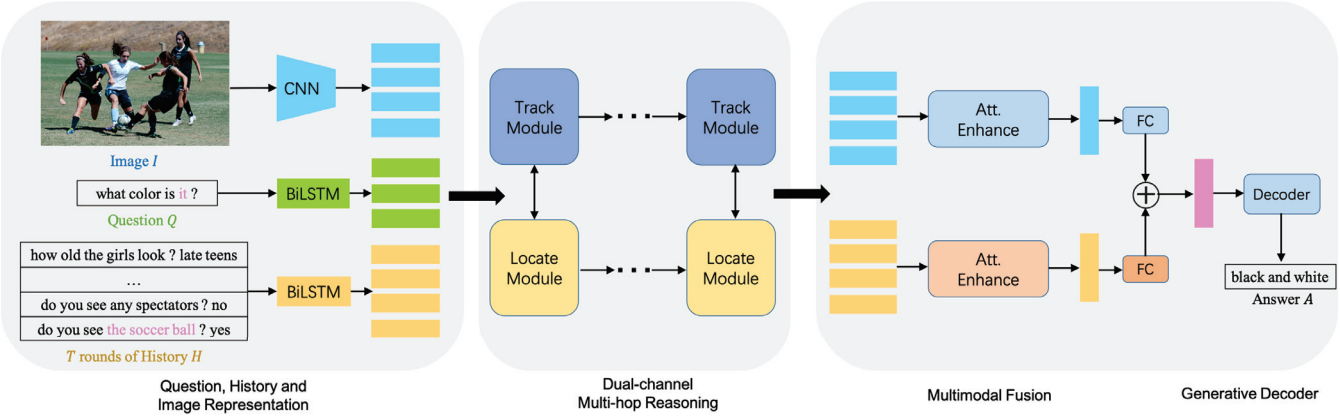

Figure 1: The framework of the DMRM model. DMRM synchronously captures information from the dialog history and the image to enrich the semantic representation of the question by exploiting dual-channel reasoning, which is composed of Track Module and Locate Module. Track Module aims to make a fully understanding of the question from the aspect of the image. Locate Module aims to make a fully understanding of the question from the aspect of the dialog history. Finally, the outputs of the dual-channel reasoning are sent to the decoder after att-enhanece and multimodal fusion operation.

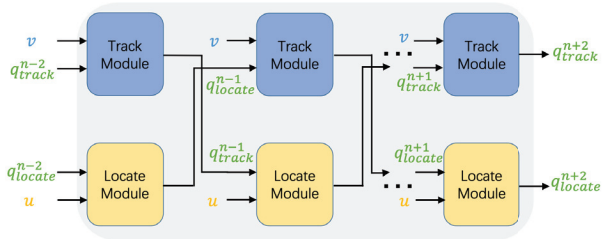

Figure 2: Schematic representation of multi-hop reasoning. Please see Section Dual-channel Multi-hop Reasoning for details. All q at different hops denote different query features, v denotes the image features and u denotes the dialog history features.

Dual-channel Multi-hop Reasoning

The dual-channel multi-hop reasoning framework is implemented via two modules, i.e., Track Module and Locate Module. Track Module aims to make a fully understanding of the question from the aspect of the image. Locate Module aims to make a fully understanding of the question from the aspect of the dialog history. The multi-hop reasoning pathway of Track Module is illustrated as $I_1 \rightarrow H_2 \rightarrow I_3 \cdots \rightarrow I_n$ and the multi-hop reasoning pathway of Locate Module is illustrated as $H_1 \rightarrow I_2 \rightarrow H_3 \cdots \rightarrow H_n$. Next, we formally describe the single-hop Track Module and Locate Module, and then extend them to multi-hop ones. We use the 3-hop reasoning in this paper.

Track Module Track Module is designed to help enrich the semantic presentation of the question from the image. In order to obtain the question- and history-aware representation of the image, we implement Track Module by taking the inspiration from bottom-up attention mechanism (Anderson et al. 2018). Track Module takes the query features q_{track} (for instance, the question feature q at reasoning hop 1) and image features v (Eq.1) as inputs, and then outputs query-aware representation of the image. We first project these two vectors to d_{track} dimension and compute soft attention over all the object detection features as follows:

$$S = f^q_{track}(q_{track}) \circ f^v_{track}(v), \quad (5)$$

$$\alpha = \text{softmax}(W^S S + b^S), \quad (6)$$

where $f^q_{track}(\cdot)$ and $f^v_{track}(\cdot)$ denote the 2-layer perceptrons with ReLU activation which transform the dimension of input features to d_{track}, W^S is the project matrix for the softmax activation and $\circ$ denotes Hadamard product. From these equations, we get the query-aware attention weights $\alpha \in \mathbb{R}^{K\times 1}$. Next we apply the query-aware attention weights to image features v to compute the query-aware representation of the image as follows:

$$q^{out}_{track} = \sum_{j=1}^{K} \alpha_j v_j. \quad (7)$$

We use $\text{Track}(\cdot,\cdot)$ to represent the operations of Track Module, namely Eq.5 - Eq.7, here and after.

Furthermore, we use Track Module in the multi-hop reasoning process to enrich the semantic presentation of the question from the image. Details are to be formalized in Section Multi-hop Reasoning.

Locate Module Locate Module is designed to get a rich representation of the question from the dialog history. Similar with Track Module, Locate Module takes the query features q_{locate} (for instance, the question feature q at reasoning hop 1) and dialog history features u (Eq.4) as inputs, and then outputs query-aware representation of the dialog history features as follows:

$$Z = f^q_{locate}(q_{locate}) \circ f^u_{locate}(u), \quad (8)$$

$$\eta = \text{softmax}(W^Z Z + b^Z), \quad (9)$$

where $f^q_{locate}(\cdot)$ and $f^v_{locate}(\cdot)$ denote the two layer multi-layer perceptrons with ReLU activation which transform the dimension of input features to d_{locate}, W^Z is the project matrix for the softmax activation and $\circ$ denotes Hadamard product. From these equations, we get the query-aware attention weights $\eta \in \mathbb{R}^{T\times 1}$. Next we apply the query-aware attention weights to the dialog history features u to compute the query-aware representation of the dialog history as follows:

$$\hat{u} = \sum_{j=1}^{T} \eta_j u_j. \tag{10}$$

Next we apply $\hat{u}$ to two layer multi-layer perceptrons with ReLU activation in between, then add it with the representation of the caption u_0. Layer normalization (Kang, Lim, and Zhang 2019) is also applied in this step.

$$\begin{aligned} g &= W^2_u \mathrm{ReLU}(W^1_u \hat{u} + b^1_u) + b^2_u, & (11)\\ q^{out}_{locate} &= \mathrm{LayerNorm}(g + u_0). & (12) \end{aligned}$$

We use $\mathrm{Locate}(\cdot,\cdot)$ to represent the operations of locate module, namely Eq.8 - Eq.12, here and after.

Furthermore, we use Locate Module in the multi-hop reasoning process to enrich the semantic presentation of the question from the dialog history. Details are to be formalized in Section Multi-hop Reasoning.

Multi-hop Reasoning Dual-channel multi-hop reasoning contains two types of multi-hop reasoning. One is multi-hop reasoning, starting from and ending with the image, illustrated as $I_1 \rightarrow H_2 \rightarrow I_3 \cdots \rightarrow I_n$. The other one is multi-hop reasoning, starting from and ending with the dialog history, illustrated as $H_1 \rightarrow I_2 \rightarrow H_3 \cdots \rightarrow H_n$. We implement each reasoning pathway via Track Module and Locate Module. The reasoning pathway $I_1 \rightarrow H_2 \rightarrow I_3 \cdots \rightarrow I_n$ includes the following steps:

$$\begin{aligned} \text{step } 1: &\quad \mathrm{Track}(q, v) \rightarrow q^1_{track};\\ \text{step } 2: &\quad \mathrm{Locate}(q^1_{track}, u) \rightarrow q^2_{track};\\ \text{step } 3: &\quad \mathrm{Track}(q^2_{track}, v) \rightarrow q^3_{track};\\ &\cdots\\ \text{step } n: &\quad \mathrm{Track}(q^{n-1}_{track}, v) \rightarrow q^n_{track}. \end{aligned}$$

The reasoning pathway $H_1 \rightarrow I_2 \rightarrow H_3 \cdots \rightarrow H_n$ includes the following steps:

$$\begin{aligned} \text{step } 1: &\quad \mathrm{Locate}(q, u) \rightarrow q^1_{locate};\\ \text{step } 2: &\quad \mathrm{Track}(q^1_{locate}, v) \rightarrow q^2_{locate};\\ \text{step } 3: &\quad \mathrm{Locate}(q^2_{locate}, u) \rightarrow q^3_{locate};\\ &\cdots\\ \text{step } n: &\quad \mathrm{Locate}(q^{n-1}_{locate}, u) \rightarrow q^n_{locate}. \end{aligned}$$

Parameters of modules at each reasoning hop are not shared with the other. Note that the reasoning process is valid only if n is an odd number. In this paper, we use 3-hop reasoning for Visual Dialog.

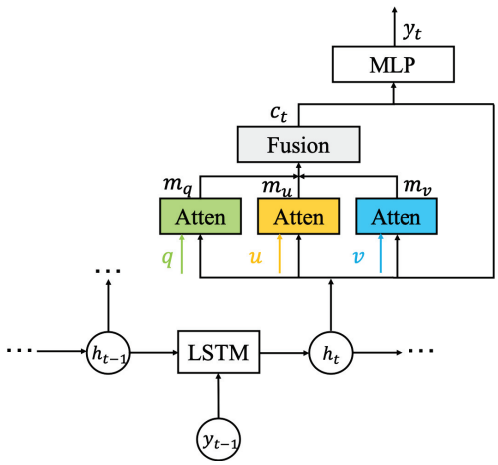

Figure 3: Multi-modal Attention Decoder. We use the multi-modal context vector e (Eq.16) to initial the deocder LSTM, utilize hidden h_t to attend to the question features q, history features u, image features v and combine the attended representations m_q, m_u, m_v to predict the next word united with hidden h_t.

Multimodal Fusion

In this section, we introduce multimoal fusion. As shown in Figure 1, before we fuse the multimodal representations q^n_{track} and q^n_{locate} generated by Track Module and Locate Module, we use question features q to enhance the representations q^n_{track} and q^n_{locate} as follows:

$$\begin{aligned} \hat{q}^n_{track} &= f^q_{att}(q) \circ f^h_{att}(q^n_{track}), & (13)\\ \hat{q}^n_{locate} &= f^q_{att}(q) \circ f^v_{att}(q^n_{locate}), & (14) \end{aligned}$$

where $f^q_{att}(\cdot)$, $f^h_{att}(\cdot)$ and $f^v_{att}(\cdot)$ denote 2-layer perceptrons with ReLU activation. Both Eq.13 and Eq.14 are named as the Att-Enhance module. We also use Att-Enhance modules between 2-hop and 3-hop. Then we fuse the representations of two channels as follows:

$$\begin{aligned} e &= [W^1_f \hat{q}^n_{track} + b^1_f, W^2_f \hat{q}^n_{locate} + b^2_f,], & (15)\\ \hat{e} &= \tanh(W^3_f e + b^3_f), & (16) \end{aligned}$$

where $[\cdot]$ denotes the concatenation operation, W^1_f, W^2_f, W^3_f and b^1_f, b^2_f, b^2_f are learned parameters.

Generative Decoder

As illustrated in Figure 3, our generative decoder is adapted from spatial attention based decoders (Lu et al. 2017b). In the encoder-decoder framework, with recurrent neural network (RNN), we model the conditional probability as:

$$p(y_t | y_1, \ldots, y_{t-1}, q, v, u) = f(h_t, c_t), \tag{17}$$

where f is 2-layer perceptrons with ReLU activation, c_t is the mulitmodal context vector at time t and h_t is the hidden state of the RNN at time t. In this paper, we use LSTM and h_t is modeled as:

$$h_t = \mathrm{LSTM}(y_{t-1}, h_{t-1}), \tag{18}$$

where y_{t-1} is the representation of the generative answer at time step $t-1$.

Given the question features q, dialog history features u, image features v, and hidden state h_t, we feed them through a 1-layer perceptron with a softmax function to generate the three attention distribution over the question, T rounds of the history and K object detection features per image, respectively. First, the attended question vector m_q is as defined:

$$z_t^q = W_h^q \tanh(W_q q + (W_g^q h_t)\mathbb{1}^T), \quad (19)$$

$$\alpha_t^q = \text{softmax}(z_t^q), \quad (20)$$

where $\mathbb{1}$ is a vector with all elements set to 1, W_q, W_g^q, W_h^q are learned parameters. All the bias terms in description of Eq.19 and Eq.22 are omitted for simplicity. Then we obtain the attended question vector m_q as follows:

$$m_q = \sum_{i=1}^{l} \alpha_{t,i}^q q_i. \quad (21)$$

Similar with the computation of attended question, we obtain the attended history vector m_u and attended image vector m_v. Then we fuse these three context vectors to obtain the context vector c_t by:

$$c_t = \tanh(W_c[m_q, m_h, m_v]), \quad (22)$$

where $[\cdot]$ denotes concatenation and w_c is learned parameters. c_t and h_t are combined to predict next word y_{t+1}.

In addition, we use the encoder output $\hat{e}$ as embedding input to initialize our decoder LSTM. Formally,

$$h_0 = \text{LSTM}(\hat{e}, s_q), \quad (23)$$

where s_q is the last state of the question LSTM in the encoder and h_0 is used as the initial state of the decoder LSTM.

Experiments

Datasets

We evaluate our proposed approach on the VisDial v0.9 and v1.0 datasets (Das et al. 2017). VisDial v0.9 contains 83k dialog on COCO-train (Lu et al. 2017a) and 40k dialog on COCO-val (Lu et al. 2017a) images, for a total of 1.23M dialog question-answer pairs. VisDial v1.0 dateset is an extension of VisDial v0.9 dateset with an additional 10k COCO-like images from Flickr. Overall, VisDial v1.0 dateset contains 123k (all images from v0.9), 2k and 8k images as train, validation and test splits, respectively.

Evaluation Metrics

We follow Das et al. (2017) to use a retrieval setting to evaluate the individual responses at each round of a dialog. Specifically, at test time, apart from the image, ground truth dialog history and the question, a list of 100 candidates answers are also given. The model is evaluated on retrieval metrics: (1) rank of human response, (2) existence of the human response in $top-k$ ranked responses, i.e., recall@k and (3) mean reciprocal rank (MRR) of the human response. Since we focus on evaluating the generalization ability of our generator, the sum of the log-likelihood of each option is used for ranking.

Model	MRR	R@1	R@5	R@10	Mean
AP (Das et al. 2017)	37.35	23.55	48.52	53.23	26.50
NN (Das et al. 2017)	42.74	33.13	50.83	58.69	19.62
LF (Das et al. 2017)	51.99	41.83	61.78	67.59	17.07
HREA (Das et al. 2017)	52.42	42.28	62.33	68.71	16.79
MN (Das et al. 2017)	52.59	42.29	62.85	68.88	17.06
HCIAE (Lu et al. 2017a)	53.86	44.06	63.55	69.24	16.01
CoAtt (Wu et al. 2018)	54.11	44.32	63.82	69.75	16.47
CoAtt (Wu et al. 2018)†	55.78	46.10	65.69	71.74	14.43
RvA (Niu et al. 2019)	55.43	45.37	65.27	**72.97**	**10.71**
DMRM	**55.96**	**46.20**	**66.02**	72.43	13.15

Table 1: Performance on VisDial val v0.9 (Das et al. 2017). Higher the better for mean reciprocal rank (MRR) and recall@k (R@1, R@5, R@10), while lower the better for mean rank. Our proposed model outperforms all other models on MRR, R@5, and mean rank. † indicates that the model is trained by using reinforcement learning.

Model	MRR	R@1	R@5	R@10	Mean
MN (Das et al. 2017)‡	47.99	38.18	57.54	64.32	18.60
HCIAE (Lu et al. 2017a)‡	49.10	39.35	58.49	64.70	18.46
CoAtt (Wu et al. 2018)‡	49.25	39.66	58.83	65.38	18.15
ReDAN (Gan et al. 2019)	49.69	**40.19**	59.35	66.06	17.92
DMRM	**50.16**	40.15	**60.02**	**67.21**	**15.19**

Table 2: Performance on VisDial val v1.0 (Das et al. 2017). ‡ All the models are re-implemented by Gan et al. (2019).

Implementation Details

To process the data, we first lowercase all the texts, convert digits to words, then remove contractions before tokenizing. The captions, questions and answers are further truncated to ensure that they are no longer than 24, 16 or 8 tokens, respectively. We then construct the vocabulary of tokens that appear at least 5 times in the training split, giving us a vocabulary of 8,958 words on VisDial v0.9 and 10,366 words on VisDial v1.0. All the BiLSTMs in our model are 1-layered with 512 hidden states. The Adam optimizer (Kingma and Ba 2014) is used with the base learning rate of 1e-3, further decreasing to 1e-5 with a warm-up process.

Results and Analysis

We compare our proposed models to the state-of-the-art generative models developed in previous works. As shown in Table 1 and Table 2, our proposed model achieves the state-of-the-art results on some metrics on the VisDial v0.9 and v1.0 datasets. The key observations are as follows:

- By comparing with singe-hop approaches (LF (Das et al. 2017) and HCAIE (Lu et al. 2017a)), we demonstrate the validity of multi-hop reasoning, because it utilizes the abundant latent information between modalities.
- By comparing with singe-channel approaches (CoAtt (Wu et al. 2018) and RvA (Niu et al. 2019)), we come to the conclusion that dual-channel reasoning is beneficial for gaining an original understanding of the question from the dialog history and the image.

Model	MRR	R@1	R@5	R@10	Mean
DMRM w/ 1-hop	55.04	45.55	64.46	70.49	14.68
DMRM w/ 2-hop	54.87	44.85	65.05	71.75	13.66
DMRM w/ 3-hop	55.57	45.80	65.54	72.09	13.51
DMRM w/o Locate	54.77	45.35	64.04	70.01	14.81
DMRM w/o Track	53.28	43.06	63.47	70.06	14.54
DMRM w/o AttD	55.57	45.80	65.54	72.09	13.51
DMRM	**55.96**	**46.20**	**66.02**	**72.43**	**13.15**

Table 3: Ablation study of our proposed model on VisDial val v0.9 (Das et al. 2017). "DMRM w/ n-hop" means the model use n-hop reasoning. "DMRM w/o AttD" means the model is not use multimodal attention decoder. Note that "DMRM w/ 2-hop" is an incomplete reasoning process under our designed architecture and the ablation study of n-hop reasoning is based on the model "DMRM w/o AttD".

	HCIAE	Ours
Human evaluation method 1 (M1):	0.60	**0.65**
Human evaluation method 2 (M2):	0.53	**0.62**

Table 4: Human evaluation on 100 sampled responses on VisDial val v0.9. M1: percentage of responses pass the Turing Test. M2: percentage of responses evaluated as better or equal to human responses.

- By comparing with other methods and the state-of-the-art approaches (HREA (Das et al. 2017), MN (Das et al. 2017) and ReDAN (Gan et al. 2019)), our approach achieves the state-of-the-art results on some metrics that demonstrate the superiority of our model.

Ablation Study As shown in Table 3, different settings explain the importance of each part of our model. By comparing "DMRM w/ n-hop", we see the effectiveness of multi-hop reasoning. By comparing "DMRM w/o Locate" and "DMRM w/o Track" with DMRM, we see the effectiveness of our dual-channel models. By comparing our final model DMRM with "DMRM w/o AttD", we illustrates the improvement due to multimodal attention in the decoder.

Significance Test We use t-test and analysis of variance (ANOVA) to analyze results of sentences generated by our model and the HCIAE model (Lu et al. 2017a). The p-values of these two analytical methods are all less than 0.01, indicating that the results are significantly different.

Human Evaluation We randomly extract 100 samples for human evaluation. The evaluation results are as shown in Table 4, which show the effectiveness of our model.

Quantitative Results Analysis As shown in Figure 4, our model generate responses of a high degree of consistency with human answers, which shows the effectiveness of it. Compared with "DMRM w/o AttD", DMRM generate more correct and meaningful responses. Figure 5 is the visualization of our reasoning process. For the question "*what color is his bike ?*", the model infers step by step, finally pays attention to the bike to answer the current question.

Related Work

Vision-language Task Vision-language tasks, such as image caption (Ren, Kiros, and Zemel 2015; Gao et al. 2015; Kinghorn, Zhang, and Shao 2018; Tan and Chan 2019; Ding et al. 2019) and visual question answering (VQA) (Yang et al. 2016; Anderson et al. 2018; Alberti et al. 2019; Cadene et al. 2019; Vedantam et al. 2019), have aroused great interest in recent years. Image caption is a task of describing the visual content of an image by using one or more sentences while visual question answering focuses on providing a natural language answer given an image and free-form, open question. Visual dialog (Wu et al. 2018; Lu et al. 2017a; Seo et al. 2017; Guo, Xu, and Tao 2019) can be seen as an extension of image caption and VQA tasks. Visual dialog enables an AI agent not only to interact with the visual environment but also to have a continuous conversation with human.

Visual Dialog Visual dialog has attracted widespread attention. Some previous works are similar to our work, but fundamentally different from ours. Das et al. (2017) propose a dialog-RNN, which takes the question, the image and the last round history as inputs, and then produces both an encoding representation for this round and a dialog context for the next round. Das et al. (2017) exploit a dialog-RNN to deal with the mutil-turn dialog only by using the information of last round history while we leverage a multi-hop reasoning for visual dialog at each turn and exploit the whole dialog history. Besides, Gan et al. (2019) provide a multi-step reasoning model via a RNN, which firstly leverage the query and history to attend to the image, secondly use the query and the image to attend to the history, and finally utilize the image and the history to update RNN State. Nevertheless, we propose the dual-channel multi-hop reasoning via two modules where Track Module only deals with the image and Locate Module only utilizes the information of the dialog history. Moreover, we conduct the representation of the question to the representation generated by Track Module and Locate module between reasoning hops.

Conclusion

We introduce our Dual-channel Multi-hop Reasoning Model (DMRM) for visual dialog, a new framework to simultaneously capture information from the dialog history and the image to enrich the semantic representation of the question by exploiting dual-channel reasoning. This dual-channel mulit-hop reasoning process provides a more fine-grained understanding of the question by utilizing the textual information and the visual context simultaneously via multi-hop reasoning, thus boosting the performance of answer generation. Experiments conducted on the VisDial v0.9 and v1.0 certify the effectiveness of our proposed method.

Image+Caption	Question	Human-Answer	DMRM w/o AttD	DMRM
 2 baskets of oranges on a table near a written note	Q1: is this photo in color ?	A1: yes	A1: yes	A1: yes , it is a color photo
	Q2: what color is the basket ?	A2: dark brown	A2: white	A2: brown
	Q3: can you tell how many oranges in the basket ?	A3: probably about 15 in each	A3: yes	A3: more than 20
	Q4: is the basket wicker ?	A4: i ca n't really tell	A4: yes	A4: no , it looks like a plastic basket
	Q5: is there a tablecloth on the table ?	A5: no	A5: no	A5: i do n't think so
	Q6: do you see any people ?	A6: no	A6: no	A6: no
 a crowd of people standing on top of a field flying colorful kites	Q1: are there a lot of kites ?	A1: yes	A1: yes	A1: yes , there are a lot of kites
	Q2: is it sunny ?	A2: yes	A2: yes	A2: yes , it 's a nice sunny day
	Q3: is the grass green ?	A3: yes	A3: yes	A3: yes , it 's a nice green field
	Q4: can you see trees ?	A4: yes	A4: yes	A4: yes in the far background in the distance
	Q5: do you see children ?	A5: yes	A5: no	A6: yes , there are a lot of children
 a group of 3 men standing next to each other	Q1: is this picture in color ?	A1: yes this picture is in color	A1: yes	A1: yes , it is a color photo
	Q2: are they wearing hats ?	A2: no they are not wearing hats	A2: no they are not	A2: no they are not wearing hats
	Q3: does any 1 have glasses on ?	A3: no	A3: yes 1 woman is wearing glasses	A3: no they do not have any glasses on
	Q4: is it day time ?	A4: i can not tell	A4: it s indoors	A4: i ca n't tell if it is day
	Q5: is it raining ?	A5: this photo is indoors so there is not	A5: no	A5: it does not look like it is raining

Figure 4: Qualitative results of our final model (DMRM) on VisDial v0.9 comparing to human ground-truth answers and our baseline model ("DMRM w/o AttD"). Compared with "DMRM w/o AttD", DMRM utilizes the muliti-modal attention in the decdoer. The improvement of correctness (marked in green and red) and interpretability (marked in blue) of generated answers due to our multi-modal attention in the decoder are partially colored.

(a) (b) (c)

Figure 5: Visualization of our reasoning process. (a) The attended image at 1-hop via Track Module. (b) The attended image at 2-hop via Track Module. (c) The attended image at 3-hop via Track Module. With the question "what color is his bike ?", our model finally attends to the bike to get the answer.

Acknowledgments

This work was supported by the Major Project for New Generation of AI (Grant No. 2018AAA0100400), the National Natural Science Foundation of China (Grant No. 61602479), and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB32070000).

References

Alberti, C.; Ling, J.; Collins, M.; and Reitter, D. 2019. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2131–2140.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic propositional image caption evaluation. *Adaptive Behavior* 11(4):382–398.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Cadene, R.; Ben-Younes, H.; Cord, M.; and Thome, N. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1989–1998.

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*, 326–335.

Ding, S.; Qu, S.; Xi, Y.; Sangaiah, A. K.; and Wan, S. 2019. Image caption generation with high-level image features. *Pattern Recognition Letters* 123:89–95.

Gan, Z.; Cheng, Y.; Kholy, A. E.; Li, L.; Liu, J.; and Gao, J. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6463–6474.

Gao, H.; Mao, J.; Zhou, J.; Huang, Z.; Wang, L.; and Xu, W. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, 2296–2304.

Guo, D.; Xu, C.; and Tao, D. 2019. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10434–10443.

Hu, R.; Andreas, J.; Darrell, T.; and Saenko, K. 2018. Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision*, 53–69.

Hudson, D. A., and Manning, C. D. 2018. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*.

Kang, G.-C.; Lim, J.; and Zhang, B.-T. 2019. Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2024–2033.

Kinghorn, P.; Zhang, L.; and Shao, L. 2018. A region-based image caption generator with refined descriptions. *Neurocomputing* 272:416–424.

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

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, 289–297.

Lu, J.; Kannan, A.; Yang, J.; Parikh, D.; and Batra, D. 2017a. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, 314–324.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017b. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 375–383.

Niu, Y.; Zhang, H.; Zhang, M.; Zhang, J.; Lu, Z.; and Wen, J.-R. 2019. Recursive visual attention in visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6679–6688.

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.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 91–99.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, 2953–2961.

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.

Tan, Y. H., and Chan, C. S. 2019. Phrase-based image caption generator with hierarchical lstm network. *Neurocomputing* 333:86–100.

Vedantam, R.; Desai, K.; Lee, S.; Rohrbach, M.; Batra, D.; and Parikh, D. 2019. Probabilistic neural symbolic models for interpretable visual question answering. In *Proceedings of International Conference on Machine Learning*, 6428–6437.

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*, 6106–6115.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of International Conference on Machine Learning*, 2048–2057.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21–29.