Dual Visual Attention Network for Visual Dialog

Dan Guo, **Hui Wang** and **Meng Wang**
School of Computer Science and Information Engineering, Hefei University of Technology
guodan@hfut.edu.cn, wanghui.hfut@gmail.com, eric.mengwang@gmail.com

Abstract

Visual dialog is a challenging task, which involves multi-round semantic transformations between vision and language. This paper aims to address cross-modal semantic correlation for visual dialog. Motivated by that V_g (global vision), V_l (local vision), Q (question) and H (history) have inseparable relevances, the paper proposes a novel Dual Visual Attention Network (DVAN) to realize $(V_g, V_l, Q, H) \Rightarrow A$. DVAN is a three-stage query-adaptive attention model. In order to acquire accurate A (answer), it first explores the textual attention, which imposes the question on history to pick out related context H'. Then, based on Q and H', it implements respective visual attentions to discover related global image visual hints V_g' and local object-based visual hints V_l'. Next, a dual crossing visual attention is proposed. V_g' and V_l' are mutually embedded to learn the complementary of visual semantics. Finally, the attended textual and visual features are combined to infer the answer. Experimental results on the VisDial v0.9 and v1.0 datasets validate the effectiveness of the proposed approach.

1 Introduction

In recent years, the cross-modal semantic understanding between vision and language has gained more and more interest and attention in the computer vision and natural language processing fields. Great progresses have been achieved in a variety of multi-modal applications including image captioning [Karpathy and Fei-Fei, 2015; Xu *et al.*, 2015; Lu *et al.*, 2017b], referring expressions [Hu *et al.*, 2016; Zhang *et al.*, 2018], visual question answering (VQA) [Antol *et al.*, 2015; Patro and Namboodiri, 2018; Anderson *et al.*, 2018], and visual dialog [Das *et al.*, 2017]. In this paper, we focus on the visual dialog, which can be regarded as originating from VQA. Based on a single question, VQA requires the agent to identify the interest area in the image and infer an answer. As an extension of VQA, visual dialog [Das *et al.*, 2017] is in the style of multi-round question-answer (QA) pairs in a GuessWhat game. Semantic co-reference among question, history and visual cues is a crucial problem in the visual dialog task. Based on textual and visual features,

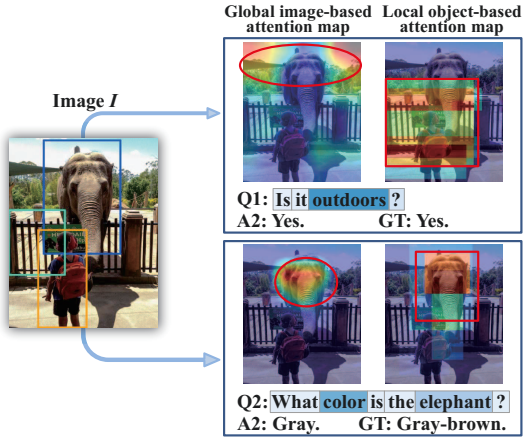

Figure 1: An example in the VisDial v0.9 dataset. Given an image I, there are two types of visual attention maps obtained by the proposed model. If these two maps both focus on the same regions, the model has high confidence with the visual reasoning; otherwise, these visual cues are complementary. Utilizing both global and local visions is beneficial to enhance accurate semantic inference.

early studies focused on semantic fusion [Das *et al.*, 2017]. In addition, to acquire the multi-modal semantic correlation, state-of-the-art visual dialog approaches [Lu *et al.*, 2017a; Wu *et al.*, 2018] applied various attention mechanisms on feature representations of vision and language, which yield a lot of promising results. However, these approaches only refer to one type of visual feature, *i.e.*, global image-based feature [Lu *et al.*, 2017a; Kottur *et al.*, 2018; Wu *et al.*, 2018] or local object-based feature [Niu *et al.*, 2018].

As illustrated in Figure 1, visual dialog requires the agent to understand the image content comprehensively. Given an image I, question 1 (Q1: "Is it outdoors?") requires the agent to understand the global visual context, while question 2 (Q2: "What color is the elephant?") focuses on specific objects in the image to infer the answer. Accurate visual grounding regions perform crucial impacts on semantic reasoning. For different questions, local and global attention maps reflect different visual responses related to the rich textual semantics (question and history). Therefore, effectively utilizing the visual complementarity is beneficial to address the cross-modal semantic correlation for visual dialog.

In this paper, we propose a novel Dual Visual Attention Network (DVAN), which explores visual cues from different views related to the current question. It utilizes the merits of both global image-based and local object-based visual features. As shown in Figure 2, DVAN first imposes the current question on history to acquire the attended history feature. Here it is sentence-level textual attention. Next, the attended history and question features are used to refer the related image regions and detected bounding boxes. This process consists of two visual reasoning steps. In the first visual reasoning step, the proposed DVAN model imposes textual semantic features on global and local visual features respectively. Essentially, this step is a intra-visual reasoning. For the second step, the model explores mutual correlation through a dual crossing attention between global and local visions, *i.e.*, inter-visual reasoning. Up to now, the sentence-level textual semantics have already instructed the respective independent and mutual crossing correlation learning between local and global visions. We further consider a fine-grained representation of question Q at word-level, which enhances the keywords' semantics. Finally, an answer is inferred by a multi-model semantic fusion scheme.

The main contributions are summarized as follows:

- The paper proposes a Dual Visual Attention Network (DVAN) to enhance the question-related cues on history, global and local visions progressively. Experimental results on VisDial v0.9 and v1.0 show that the proposed approach achieves state-of-the-art performances.
- It tackles visual co-reference using both independent and mutual visual attention mechanisms. Image-based global feaure and object-based local visual feature are introduced to the DVAN model.
- DVAN is a question-adaptive attention model, which considers both sentence-level and word-level textual attentions. Both of them perform well for visual dialog.

2 Related Work

2.1 Visual Dialog

As an extension of the vision-language task, visual dialog was introduced by [de Vries *et al.*, 2017; Das *et al.*, 2017]. Specifically, De Vries et al. [de Vries *et al.*, 2017] collected a GuessWhat dataset by a two-player guessing game, where one agent asks questions guessing which object has been selected in the image, and the second agent answers in yes/no/NA. Das et al. [Das *et al.*, 2017] proposed a large visual dialog dataset VisDial, which pairs two annotators on Amazon Mechanical Turk to collect free-form questions and answers about an image. The questioner asked questions to help himself better imagine the unseen image.

Das et al. [Das *et al.*, 2017] introduced three baseline models, *i.e.*, late fusion (LF), hierarchical recurrent network (HRE), and memory network (MN). LF encoded the question, dialog history, and image separately and then concatenated them to a joint representation for answer inferring. HRE used a hierarchical recurrent encoder for history encoding. MN stored each question-answer pair as a 'fact' and answered the current question based on the facts. In addition, to improve performance, attention mechanisms have been widely used in VisDial task. In [Lu *et al.*, 2017a], a history-conditioned image attention model (HCIAE) was proposed, which discriminatively attended on image features spatially according to the dialogue context. In order to capture more accurate visual regions, Wu et al. [Wu *et al.*, 2018] introduced a sequential co-attention model (CoAtt), which applied multi-step co-attentions over question, history, and VGG visual feature maps. Seo et al. [Seo *et al.*, 2017] proposed an attention memory network (AMEM), which addressed the co-reference of the question and history. Niu et al. [Niu *et al.*, 2018] proposed a recursive visual attention model which selectively reviewed the dialog history and recursively refined the visual attention for visual grounding. Kottur et al. [Kottur *et al.*, 2018] applied neural module network [Andreas *et al.*, 2016] to solve this visual reference problem at word-level.

2.2 Dual Visual Representation

For visual correlation learning, the point of dual visual representation has been explored in VQA [Lu *et al.*, 2018; Farazi and Khan, 2018]. Lu et al. [Lu *et al.*, 2018] employed a multiplicative embedding scheme to realize co-attention among question, global and local visions. Similarly, in [Farazi and Khan, 2018], a hierarchical co-attention scheme using global and local visions was proposed, which focused on multi-modal semantic fusion. The method most related to this paper is [Lu *et al.*, 2018]. Lu et al. applied a crossing visual attention performing on original global and local visual features at once time for VQA. In contrast, we adopt a progressive attention mechanism. Based on global and local visual features, the proposed DVAN model first explores respective independent visual attention to learn intra-visual correlation and then implements mutual crossing visual attention to model the inter-visual correlation. Moreover, in the fusion stage, Lu et al. directly fused the question and visual representations, while we use a self-attended question at word-level to further enhance related visual semantics, and then jointly fuse the multi-modal features to infer the answer.

3 Proposed Method

A visual dialog is defined as that at current round t, given an image I and previous history H including the image caption c and t-1 question-answer pairs (*i.e.*, $H = (c, (q_1, a_1), ..., (q_{t-1}, a_{t-1}))$), the dialog agent has to answer the follow-up question Q. The agent is divided into two types, *i.e.*, discriminative and generative models. Discriminative models select the answer with the maximum score from a list of 100 candidate answers $A_t = \{a_t^{(1)}, ..., a_t^{(100)}\}$, while generative models decode an answer by sequential learning models. Generative models are optimized by maximizing the log-likelihood of the ground truth answer $a_t^{gt} \in A_t$.

As illustrated in Figure 2, our model consists of three modules. (1) Feature Embedding Learning (Section 3.1). To better learn the correlation among multi-modal features, we embed each feature to the same feature dimension. (2) Question-adaptive Dual Visual Attention (Section 3.2). A multi-stage attention mechanism is proposed to progressively enhance the question-adaptive cues on history, global and local visions. The whole process is question-driven. It first tackles the

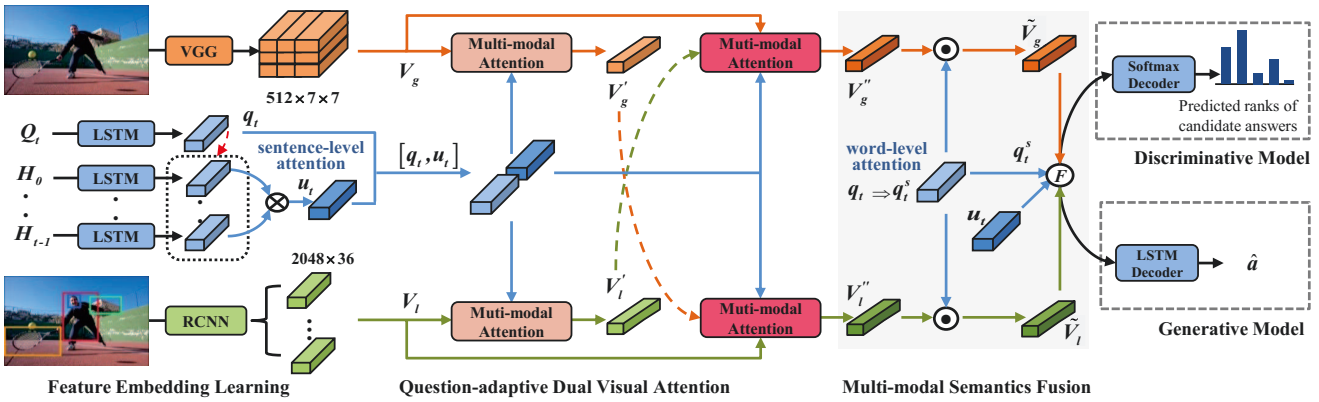

Figure 2: The overall framework of Dual Visual Attention Network (DVAN) for visual dialog.

textual inferring, and then attends the textual cues on global and local visual features individually. Finally, this part implements co-reference by mutual global and local visual correlation. (3) Multi-modal Semantics Fusion (Section 3.3). In this module, the attended textual and visual features are fused together to infer the answer. In the following, we discuss each module in detail.

3.1 Feature Embedding Learning

For visual embedding learning, we first adopt a pre-trained VGG19 [Simonyan and Zisserman, 2014] as feature extractor. The output of the last pooling layer of VGG19 is taken as global visual feature $V_g^{(0)} \in \mathbb{R}^{d_g \times M}$, where $M = 7 \times 7$ is spatial size and $d_g = 512$ is the channel number of the feature maps. Then, we use the Faster R-CNN [Ren *et al.*, 2015] pre-trained on Visual Genome dataset [Krishna *et al.*, 2017] to obtain local visual feature $V_l^{(0)} \in \mathbb{R}^{d_l \times K}$, where $K = 36$ is the number of detected objects for per image and $d_l = 2048$ is the feature dimension. In order to better calculate the correlation between these two types of visual features, we employ the fully connected (FC) layer to map original $V_g^{(0)}$ and $V_l^{(0)}$ into the same dimension d respectively:

$$\begin{cases} V_g = tanh(W_g V_g^{(0)}) \in \mathbb{R}^{d \times M} \\ V_l = tanh(W_l V_l^{(0)}) \in \mathbb{R}^{d \times K} \end{cases} \quad (1)$$

where $W_g \in \mathbb{R}^{d \times d_g}$ and $W_l \in \mathbb{R}^{d \times d_l}$ are learnable parameters of respective FC layer. All bias terms of FC layers in this paper are omitted for simplicity.

For textual embedding learning, at the t-th dialogue round, we first assign each word x_i (in the current question Q) into a one-hot vector. Next, word x_i is embedded into a vector w_i through a learnable word embedding matrix W_e. We obtain the textual features of question Q as follows:

$$w_i = W_e x_i, \quad Q = [w_1, w_2, ..., w_N] \quad (2)$$

where N is the word number of question Q.

To learn the temporal correlation of words in question Q, LSTM [Hochreiter and Schmidhuber, 1997], a basic RNN unit, is used to get the embedding representation of Q. We implement $Q' = LSTM\,(Q)$, where $Q' \in \mathbb{R}^{d \times N}$, and d is the same dimension as the embedded visual features. We take the last hidden state $LSTM(w_N)$ as the embedding feature of Q and denote it as q_t. Similarly, we adopt another LSTM to obtain the history embedding features H. Each previous round of history is encoded. Here is $H = [h_0, h_1, ..., h_{t-1}] \in \mathbb{R}^{d \times t}$, where h_0 denotes the textual embedding feature of image caption c (*i.e.*, $h_0 = LSTM(c)$).

3.2 Question-adaptive Dual Visual Attention

In order to achieve effective cross-modality semantic understanding on V_g (global vision), V_l (local vision), Q and H for inferring answer A, a novel Dual Visual Attention Network (DVAN) is proposed. It is a three-stage query-adaptive attention model. The details are presented as follows.

The 1st-stage Attention: Question Attending to History

Visual dialog has a multi-round conversation about an image. The questions in the dialog usually contain at least one pronoun (*e.g.*, "he", "she", "it", "this"). This textual co-reference in dialog context has a great impact on the final answer inference. In other words, the agent has to find the targeted subjects in the previous history snippets. In the proposed DVAN framework, we propose a question-conditioned attention mechanism, which focuses on identifying the related history snippet semantics. It realizes a sentence-level textual feature representation by imposing question q_t on history H. The first stage attention is formulated as follows:

$$\begin{cases} z_t^h = tanh((W_q q_t)\mathbb{1}^T + W_h H) \\ \alpha_t^h = softmax(P^T z_t^h) \\ u_t = \alpha_t^h H \end{cases} \quad (3)$$

where $W_q \in \mathbb{R}^{d \times d}, W_h \in \mathbb{R}^{d \times d}$, and $P \in \mathbb{R}^{d \times 1}$ are to-be-learned attention parameters, $\mathbb{1} \in \mathbb{R}^t$ is a vector with all elements set to 1, $\alpha_t^h \in \mathbb{R}^t$ is the attention weight vector for history H, and $u_t \in \mathbb{R}^d$ is the outputted attended history representation. Then, q_t and u_t are jointly used to attend different visual features in the following sections.

The 2nd-stage Attention: Independent Visual Attention

Here we focus on addressing the visual reference problem. The question and attended history are jointly used to localize

the related visual regions in the image. In order to capture the correlation between different data modalities (vision and language), a multi-modal attention mechanism is designed to acquire attended visual features including both global image-based and local object-based representations. To effectively capture the context-aware visual semantics, here we explore the intra-visual relation under the guidance of Q and H, while the inter-visual relation is discussed in next subsection.

In this second-stage attention step, the proposed model implements the inference Q+H+$V \rightarrow V'$. It jointly combines question q_t, attended history u_t, and original visual features V to rethink V itself and obtain more discriminative representation V'. Global and local visual features are independently updated by formulae 4 and 5.

$$\begin{cases} z_t^{g1} = tanh((W_{q1}q_t + W_{h1}u_t)\mathbb{1}_g^T + W_{g1}V_g) \\ \alpha_t^{g1} = softmax(P_{g1}^T z_t^{g1}) \\ V_g^{'} = \alpha_t^{g1} V_g^T \end{cases} \tag{4}$$

$$\begin{cases} z_t^{l1} = tanh((W_{q1}^{'}q_t + W_{h1}^{'}u_t)\mathbb{1}_l^T + W_{l1}V_l) \\ \alpha_t^{l1} = softmax(P_{l1}^T z_t^{l1}) \\ V_l^{'} = \alpha_t^{l1} V_l^T \end{cases} \tag{5}$$

where $\alpha_t^{g1} \in \mathbb{R}^M$ and $\alpha_t^{l1} \in \mathbb{R}^K$ are respective attention weights over global and local visual features. $V_g^{'}$ and $V_l^{'}$ just consider respective intra-visual relation in each own feature space.

The 3rd-stage Attention: Dual Crossing Visual Attention Through the second-stage attention, the proposed model has obtained attended intra-visual embeddings V_g' and V_l' under the guidance of textual features. However, the visual complementarity between global and local visions are not considered. As shown in Figure 1, if both global and local visual attention maps focus on the same regions in image I, the model has confidence in both sides of visual reasoning. Otherwise, with different visual response regions, V_g' and V_l' can complement each other. This can help repair neglected or uncorrected visual cues. Therefore, we design the third-stage attention to capture the correlation between global and local visions (*i.e.*, inter-visual reasoning).

As shown in Figure 2, this attention step implements the dual visual crossing attention for visual reference, which also belongs to the multi-modal attention under the guidance of the question and the attended history. We implement a mutual visual correlation calculation to obtain a new global visual semantic, *i.e.*, Q+H+V_g+$V_l' \rightarrow V_g''$. Similarly, Q+H+V_l+$V_g' \rightarrow V_l''$ is conducted. The visual semantics V_l'' and V_g'' are mutually updated by the formulae 6 and 7.

$$\begin{cases} z_t^{g2} = tanh((W_{q2}q_t + W_{h2}u_t + W_{l2}V_l^{'})\mathbb{1}_g^T + W_{g2}V_g) \\ \alpha_t^{g2} = softmax(P_{g2}^T z_t^{g2}) \\ V_g^{''} = \alpha_t^{g2} V_g^T \end{cases} \tag{6}$$

$$\begin{cases} z_t^{l2} = tanh((W_{q2}^{'}q_t + W_{h2}^{'}u_t + W_{g2}^{'}V_g^{'})\mathbb{1}_l^T + W_{l2}^{'}V_l) \\ \alpha_t^{l2} = softmax(P_{l2}^T z_t^{l2}) \\ V_l^{''} = \alpha_t^{l2} V_l^T \end{cases} \tag{7}$$

where $\alpha_t^{g2} \in \mathbb{R}^M$ and $\alpha_t^{l2} \in \mathbb{R}^K$ are respective attention weights over global and local visual features.

3.3 Multi-modal Semantics Fusion

Up to now, the sentence-level textual semantics (q_t and u_t) have already instructed the respective independent and mutual crossing correlation learning between local and global visions. Here, we consider a fine-grained representation of question Q at word-level. As illustrated in Figure 1, words "outdoors", "color", and "elephant" are most related to the answer. Enhancing these keywords' semantics in Q can help infer the final answer. To this end, we apply a self-attention mechanism on question Q to focus on the keywords:

$$\begin{cases} \alpha_t^q = softmax(P_q^T tanh(W_{sq}Q')) \\ q_t^s = \alpha_t^q Q \end{cases} \tag{8}$$

where $\alpha_t^q \in \mathbb{R}^N$, $q_t^s \in \mathbb{R}^{d_m}$, d_m is the dimension of word embedding, and q_t^s is the self-attended question representation.

Then, we use the self-attended q_t^s to further impose its impact on both local and global visual contents. The refined visual features are obtained by Hadamard (element-wise) product (denoted as symbol '$\odot$'), which is similar to the gate operation within LSTM and GRU. This process is expressed as follows:

$$\begin{cases} \widetilde{V}_g = V_g^{''} \odot tanh(W_s q_t^s) \\ \widetilde{V}_l = V_l^{''} \odot tanh(W_s q_t^s) \end{cases} \tag{9}$$

where $W_s \in \mathbb{R}^{d \times d_m}$ is a to-be-learned attention parameter. $\widetilde{V}_g$ and $\widetilde{V}_l$ are the final visual embedding representations.

Finally, with self-attended question feature q_t^s, question-conditioned history feature u_t, and refined visual features $(\widetilde{V}_g, \widetilde{V}_l)$, we fuse them to obtain the final embedding e_t.

$$e_t = tanh\left(W_e[q_t^s, u_t, \widetilde{V}_g, \widetilde{V}_l]\right) \tag{10}$$

where $[,]$ is the concatenation operation. In the generative model, the feature e_t is fed into a single LSTM decoder to infer the answer $\hat{a}$. For the discriminative model, e_t is fed into a softmax decoder to sort the candidate answers in A_t. The details of the training setting of generative and discriminative models are explained in Section 4.2.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate the proposed model on the VisDial v0.9 and v1.0 [Das *et al.*, 2017] datasets. VisDial v0.9 contains 83k dialogs on COCO-train images and 40k dialogs on COCO-val images (totally 1.2M QA pairs). The dialog of each image has 10-round question-answer pairs, which were collected by a two-player image guessing chat game. Specifically, the "questioner" attempted to imagine the unseen image better by sequentially asking questions, while the "answerer" answered questions based on the observed picture. VisDial v1.0 is an updated version of the VisDial v0.9, in which VisDial v0.9 is set to be the train split. And the new val and test splits of VisDial v1.0 contains 2k and 8k dialogs collected on COCO-like Flickr images, respectively. It is worth noting that

Model	MRR	R@1	R@5	R@10	Mean
LF	0.5199	41.83	61.78	67.59	17.07
HRE	0.5237	42.29	62.18	67.92	17.07
HREA	0.5242	42.28	62.33	68.17	16.79
MN	0.5259	42.29	62.85	68.88	17.06
HCIAE	0.5386	44.06	63.55	69.24	16.01
CoAtt	0.5411	44.32	63.82	69.75	16.47
DVAN w/o OF	0.5443	44.58	64.30	70.11	15.27
DVAN w/o GF	0.5538	46.01	65.06	70.68	15.11
DVAN w/o Att3	0.5573	46.32	65.28	70.92	14.91
DVAN w/o SQ	0.5579	46.40	65.33	71.02	14.95
DVAN	**0.5594**	**46.58**	**65.50**	**71.25**	**14.79**

Table 1: Generative model comparison on VisDial v0.9 val.

in the new test split, each dialog has flexible n rounds of QA pairs, where n is in the range of 1 to 10.

Following the evaluation protocol in [Das *et al.*, 2017], the experimental performance is evaluated by retrieving the ground-truth answer from a list of 100 optional answers. The retrieval metrics include (1) **Mean**: mean rank of the ground truth answer option, (2) **Recall@K**: existence of the ground truth answer option in top-k ranked responses, and (3) **MRR**: mean reciprocal rank of the ground truth answer option. For VisDial v1.0, **NDCG** (normalized discounted cumulative gain) was introduced, which penalizes the lower rank of answers with high relevance.

4.2 Implementation Details

We tokenize all text inputs by Python NLTK toolkit and construct a vocabulary of words that appear at least 4 times in the training split. The captions, questions, and answers are truncated to 24/16/8 words for generative models, and 40/20/20 words for discriminative models, respectively. Next, each word is embedded into a 300-dim vector initialized by the GloVe embedding [Pennington *et al.*, 2014]. All the LSTMs in our model are 1-layered with 512 hidden states. The Adam optimizer [Kingma and Ba, 2014] is adopted with initialized learning rate 4×10^{-4}, multiplied by 0.5 after each 20 epochs. We also apply Dropout [Srivastava *et al.*, 2014] with radio 0.5 for LSTM, attention modules, and the output of encoder. Finally, generative models are trained with a MLE loss (maximum likelihood estimation), while discriminative models are trained with a multi-class N-pair loss [Lu *et al.*, 2017a].

4.3 Experiment Results

The compared baseline models (**LF**, **HRE**, and **MN**) are proposed by [Das *et al.*, 2017]. Specifically, **LF** directly extracted multi-modal features and fused them in the later stage. **HRE** used a hierarchical recurrent encoder [Serban *et al.*, 2017] to encode the dialog history, and **HREA** applied an attention mechanism to attend the relevant history. **MN** designed a memory bank to store previous dialog history. In recent works, **HCIAE** [Lu *et al.*, 2017a] applied a history-conditioned attention mechanism to attend on image and dialog features, and **CoAtt** [Wu *et al.*, 2018] proposed a sequential co-attention mechanism over multi-modal inputs.

Ablation Study

In order to verify each attention component, we test and verify the following ablative models:

Model	MRR	R@1	R@5	R@10	Mean
HCIAE-DIS*	0.5467	44.35	65.28	71.55	**14.23**
CoAtt-RL*	0.5578	46.10	**65.69**	**71.74**	14.43
DVAN	**0.5594**	**46.58**	65.50	71.25	14.79

Table 2: Generative model comparison on VisDial v0.9 val. * indicates that the models are trained with additional loss functions.

- **DVAN w/o OF:** DVAN with only global image-based visual feature $V_g^{(0)}$.
- **DVAN w/o GF:** DVAN with only local object-based visual feature $V_l^{(0)}$.
- **DVAN w/o Att3:** DVAN without the visual cross attention scheme (the third-stage attention). It demonstrates that there is no crossing visual attention between global and local visual features.
- **DVAN w/o SQ:** DVAN without self-attention on question Q. It means that DVAN adopts only sentence-level attention (q_t) without word-level attention (q_t^s) in the final multi-modal fusion scheme.

As shown in Table 1, compared to **DVAN w/o OF**, the full model **DVAN** improves R@1 from 44.58 to 46.58. It shows that introducing local visual feature can capture richer visual cues for the agent to infer the answer. As for **DVAN w/o GF** which considers only local visual feature, R@1 drops to 46.01. It indicates that the global and local visions are complementary. The combination usage of them performs much better than only one vision type. After removing the crossing attention module (the three-stage attention), **DVAN w/o Att3** drops R@1 from 46.58 to 46.32. This shows the crossing attention correlation between global and local visions can effectively enhance the visual semantics. A self-attended question is further proposed to consider the word-level textual semantics. **DVAN** can get better performance than **DVAN w/o SQ** by average 0.2% improvement.

Evaluation on Generative Models

For a fair comparison, all models in Table 1 use the MLE loss for generative training. **DVAN w/o OF** already outperforms other compared methods. This is interpretable, as other methods only consider the question at sentence-level representation. In contrast, we explore a word-level representation. In addition, by integrating global and local visions, the **DVAN** model obtains significant performances on all evaluation metrics. Comparing to the state-of-the-art model **CoAtt**, the **DVAN** model achieves nearly 2 points improvement on R@K and 2.1% on MRR.

Table 2 shows another experimental comparison with other approaches. **HCIAE-DIS** [Lu *et al.*, 2017a] trained a generative model with the MLE loss, and then used the knowledge from a pre-trained discriminative model to fine-tune the generative model. This strategy requires additional candidate answers. Compared to **HCIAE-DIS**, the **DVAN** model performs better on MRR, R@1 and R@5. Notably, no additional candidate answers are needed in the end-to-end **DVAN** training. **CoAtt-RL** [Wu *et al.*, 2018], which was pre-trained with the MLE loss, adopted both reinforcement learning and adversarial learning for further offline training. Compared to

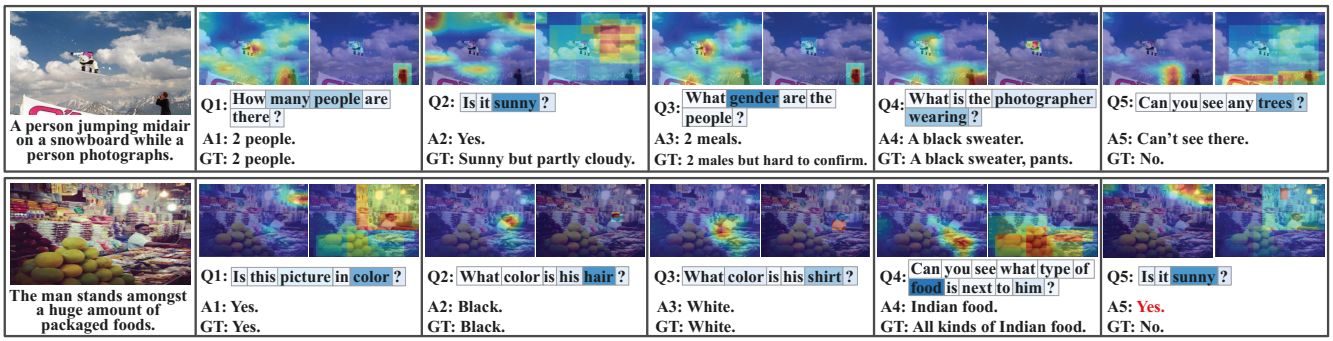

Figure 3: Two examples in the VisDial v0.9 dataset. In the box at each round, the first column shows the global visual attention map, and the second column depicts the object-based visual attention map.

Model	MRR	R@1	R@5	R@10	Mean
LF	0.5807	43.82	74.68	84.07	5.78
HRE	0.5846	44.67	74.50	84.22	5.72
HREA	0.5868	44.82	74.81	84.36	5.66
MN	0.5965	45.55	76.22	85.37	5.46
HCIAE	0.6222	48.48	78.75	87.59	4.81
AMEM	0.6227	48.53	78.66	87.43	4.86
CoAtt	0.6398	50.29	80.71	88.81	4.47
CorefNMN‡	0.6410	50.92	80.18	88.81	4.45
DVAN w/o OF	0.6381	50.09	80.58	89.03	4.38
DVAN w/o GF	0.6522	51.86	81.64	89.96	4.22
DVAN w/o Att3	0.6601	52.78	82.22	90.21	4.09
DVAN w/o SQ	0.6604	52.83	82.41	90.37	4.03
DVAN	**0.6667**	**53.62**	**82.85**	**90.72**	**3.93**

Table 3: Discriminative model comparison on VisDial v0.9 val. ‡ indicates that the model uses ResNet-152 features.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF	0.5542	40.95	72.45	82.83	5.95	45.31
HRE	0.5416	39.93	70.45	81.50	6.41	45.46
MN	0.5549	40.98	72.30	83.30	5.92	47.50
CorefNMN‡	0.6150	47.55	78.10	88.80	4.40	**54.70**
DVAN	**0.6258**	**48.90**	**79.35**	**89.03**	**4.36**	**54.70**

Table 4: Discriminative model comparison on VisDial v1.0 test-std. ‡ indicates that the model uses ResNet-152 features.

CoAtt-RL, the **DAVN** model performs better on MRR and R@1, lifting MRR from 0.5578 to 0.5594 and R@1 from 46.10 to 46.58. With richer textual and visual semantics, the **DVAN** model achieves promising results even in a simple end-to-end training mode.

Evaluation on Discriminative Models

In discriminative task, the compared methods are **AMEM** [Seo *et al.*, 2017] and **CorefNMN** [Kottur *et al.*, 2018]. These two works also focused on the visual co-reference, which involves word referring (typically noun phrases and pronouns) to the same entities in an image. In contrast, the **DAVN** model emphasizes the semantic enhancement by the sentence-level and the word-level textual attentions, and the respective independent and mutual crossing visual attentions. Table 3 shows that using the VisDial v0.9 dataset, the proposed model outperforms the state-of-the-art models on all metrics by average 2%. We also evaluate on VisDial v1.0. As shown in Table 4, the **DAVN** model still achieves the best performance.

4.4 Qualitative Results

As shown in Figure 3, we give two examples from VisDial v0.9. There are some conclusions. On one side, the textual attention distribution at word-level relative to each round answering is interpretable, which captures the keywords of each question. On the other side, we discuss the two types of visual attention maps, *i.e.*, global and local visual attention maps. We denote the first row as Example 1, and the second row as Example 2. (1) For questions $Q1$ and $Q3$ in Example 1, and $Q2$ and $Q3$ in Example 2, the proposed model attends the consistent visual regions on both global and local visions. (2) However, for questions $Q2$ and $Q5$ in Example 1, and $Q1$ and $Q4$ in Example 2, global and local visions form the visual complementarily to infer the correct answer. (3) For $Q4$ in Example 1, the global map can effectively focus on the target object–"photographer", while the local map attends on the "skier" incorrectly. In this case, global map focuses on a part of the targets, while local map attends irrelevant regions. Only one visual attention map attending the related regions can also infer the correct answer. (4) There are also some challenging cases, *e.g.*, $Q5$ in Example 2. Both of the two attention maps focus on the same regions where the light is bright as "sunny", but the answer is wrong.

5 Conclusion

In this paper, we propose a Dual Visual Attention Network (DVAN) for visual dialog. DAVN aims at learning effective visual correlation under textual cues. DVAN first applies a textual attention mechanism on question and history to get related textual semantics, then imposes them on global and local visual features to acquire accurate visual semantics. The visual attention mechanism in DVAN tackles both respective independent and mutual crossing visual reasoning. Experiments conducted on the VisDial 0.9 and 1.0 datasets validate the effectiveness of the proposed model.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under grants 61725203, 61732008, and 61876058.

References

[Anderson *et al.*, 2018] 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 visual question answering. In *CVPR*, pages 6077–6086, 2018.

[Andreas *et al.*, 2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, pages 39–48, 2016.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, pages 326–335, 2017.

[de Vries *et al.*, 2017] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, pages 5503–5512, 2017.

[Farazi and Khan, 2018] Moshiur R. Farazi and Salman Khan. Reciprocal attention fusion for visual question answering. In *BMVC*, 2018.

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

[Hu *et al.*, 2016] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[Kottur *et al.*, 2018] Satwik Kottur, Jose M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, pages 153–169, 2018.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[Lu *et al.*, 2017a] 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. In *NIPS*, pages 314–324. 2017.

[Lu *et al.*, 2017b] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, pages 375–383, 2017.

[Lu *et al.*, 2018] 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. In *AAAI*, pages 7218–7225, 2018.

[Niu *et al.*, 2018] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. *arXiv preprint arXiv:1812.02664*, 2018.

[Patro and Namboodiri, 2018] Badri N. Patro and Vinay P. Namboodiri. Differential attention for visual question answering. In *CVPR*, pages 7680–7688, 2018.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99. 2015.

[Seo *et al.*, 2017] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. In *NIPS*, pages 3719–3729. 2017.

[Serban *et al.*, 2017] Iulian V. Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[Wu *et al.*, 2018] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*, pages 6106–6115, 2018.

[Xu *et al.*, 2015] 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. In *ICML*, pages 2048–2057, 2015.

[Zhang *et al.*, 2018] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *CVPR*, pages 4158–4166, 2018.