Learning Dual Encoding Model for Adaptive Visual Understanding in Visual Dialogue

Jing Yu, Xiaoze Jiang, Zengchang Qin, Weifeng Zhang, Yue Hu, and Qi Wu

***Abstract*— Different from Visual Question Answering task that requires to answer only one question about an image, Visual Dialogue task involves multiple rounds of dialogues which cover a broad range of visual content that could be related to any objects, relationships or high-level semantics. Thus one of the key challenges in Visual Dialogue task is to learn a more comprehensive and semantic-rich image representation that can adaptively attend to the visual content referred by variant questions. In this paper, we first propose a novel scheme to depict an image from both visual and semantic views. Specifically, the visual view aims to capture the appearance-level information in an image, including objects and their visual relationships, while the semantic view enables the agent to understand high-level visual semantics from the whole image to the local regions. Furthermore, on top of such dual-view image representations, we propose a Dual Encoding Visual Dialogue (DualVD) module, which is able to adaptively select question-relevant information from the visual and semantic views in a hierarchical mode. To demonstrate the effectiveness of DualVD, we propose two novel visual dialogue models by applying it to the Late Fusion framework and Memory Network framework. The proposed models achieve state-of-the-art results on three benchmark datasets. A critical advantage of the DualVD module lies in its interpretability. We can analyze which modality (visual or semantic) has more contribution in answering the current question by explicitly visualizing the gate values. It gives us insights in understanding of information selection mode in the Visual Dialogue task. The code is available at https://github.com/JXZe/Learning_DualVD.**

***Index Terms*— Dual encoding, visual module, semantic module, visual relationship, dense caption, visual dialogue.**

I. Introduction

TO UNDERSTAND the real world by analyzing vision and language together is a priority for AI to achieve human-like abilities, which enables the development of diverse applications, such as Visual Question Answering (VQA) [1], Referring Expressions [2], Image Captioning [3], [4], Cross-modal Information Retrieval [5], [6] *etc*. To move a step further, this work focuses on the Visual Dialogue [7] problem, which requires the agent to answer a series of questions in natural language regarding an image. It is more challenging because it demands the agent to adaptively focus on diverse visual content with respect to the current question in light of previous dialogue, while other visual-semantic problems mostly attend to some specific objects or regions in an image. Considering the dialogue in Figure 1: Given " *Q1: Is the man on the skateboard?*", the agent should be aware of the foreground visual content, i.e. *the man, the skateboard*, while " *Q5: Is there sky in the picture?*" changes the attention of the agent to the background of *sky*. Besides appearance-level questions like *Q1* and *Q5*, " *Q4: Is he young or older?*" requires the agent to reason about the visual content for higher-level semantics. How to adaptively capture the question-relevant visual evidence through dialogue becomes one of the most critical challenges in visual dialogue.

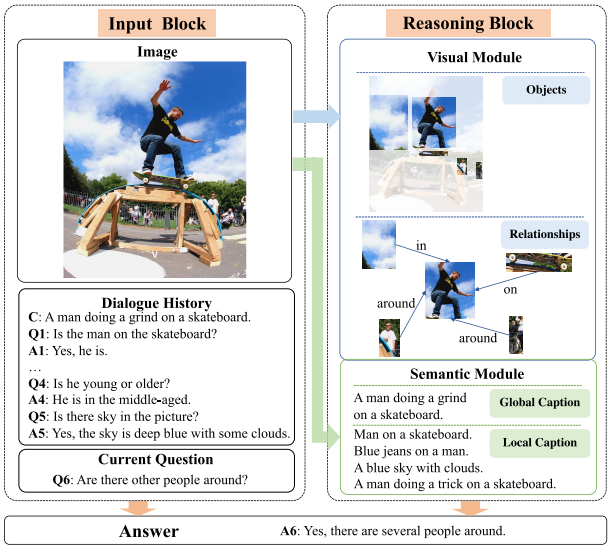


Fig. 1. Illustration of DualVD. Left: the input of the dialogue system. Right: visual and semantic modules designed to understand the visual content like humans. The answer is inferred depending on multi-modal evidence.

The typical solution for visual dialogue is to firstly fuse visual (*i.e.* image) features learned from Convolutional Neural Networks (CNN) and textual (*i.e.* dialogue history, current

Manuscript received December 20, 2019; revised August 31, 2020; accepted October 8, 2020. Date of publication November 3, 2020; date of current version November 19, 2020. This work was supported by the National Natural Science Foundation of China under Grant 62006222. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Julian Fierrez. *(Corresponding authors: Zengchang Qin; Weifeng Zhang.)*

Jing Yu and Yue Hu are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yujing02@iie.ac.cn; huyue@iie.ac.cn).

Xiaoze Jiang and Zengchang Qin are with the Intelligent Computing and Machine Learning Lab, School of ASEE, Beihang University, Beijing 100191, China (e-mail: xzjiang@buaa.edu.cn; zcqin@buaa.edu.cn).

Weifeng Zhang is with the College of Mathematics, Physics and Information Engineering, Jiaxing University, Jiaxing 314001, China (e-mail: zhangweifeng@zjxu.edu.cn).

Qi Wu is with the Australian Centre for Robotic Vision, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: qi.wu01@adelaide.edu.au).

Digital Object Identifier 10.1109/TIP.2020.3034494

question) features learned from Recurrent Neural Networks (RNN) together and then to infer the correct answer. Most approaches focus on analyzing the dialogue history by recovering the dialogue relational structure [8], imperfect dialogue history [9], and dialogue consistency [10]. However, the role of visual information was less studied. Existing models simply use CNN [11] or R-CNN [12] to extract visual features and focus on the question-relevant content via attention mechanism. Such visual features have limited expressive ability due to the monolithic representations [2]. On one hand, questions in visual dialogue tasks refer to a wide range of visual content, including objects, visual relationships and high-level semantics, which can not be covered by monolithic features. Though the monolithic features involve appearance-level clues, they have limited relational semantics and lower level of semantic abstraction. On the other hand, the referred visual content may change remarkably from visual appearance to high-level semantics through the dialogue. In order to be involved more in the conversation, the agent needs to adaptively capture visual evidence relevant to the current question and continue to infer correct answers, which is difficult for monolithic features to achieve.

Our work is inspired by the Dual-coding theory [13] of human cognition process. *Dual-coding theory* postulates that our brain encodes information in two ways: *visual imagery* and *textual associations*. For example, a person has stored the concept "cat" as both the word "cat" and the image of a cat. When asked to act upon a concept, our brain retrieves either an image or words, or both simultaneously. The ability to encode a concept in two different ways strengthens the capacity of memory and understanding compared to encode a concept in only one way.

In this work, we first propose a novel scheme to comprehensively depict an image from both visual and semantic views, which are coordinated to support subtle and abstract cognition simultaneously. Specifically, the visual view keeps the major objects and their visual relationships in an image and organizes them as a scene graph, which supports visual reasoning regarding to the question. The semantic view explicitly converts the image into high-level global and local textural captions. These captions describe the key concepts and their relationships in the language form, which is unified with the question and dialogue history in text domain. On top of the visual representation scheme, we propose a module, called *Dual Encoding Visual Dialogue* (*DualVD*), to adaptively select question-relevant information from the image in a hierarchical mode: intra-modal selection first captures the visual and semantic information individually from the object-relational visual features and global-local semantic features; then inter-modal selection obtains the joint visual-semantic knowledge by correlating vision and semantics. This hierarchical framework imitates human cognition theory to capture targeted visual clues from multiple perceptual views and semantic levels. To prove the effectiveness of DualVD, we integrate it into two typical visual dialogue frameworks, where Late Fusion framework focuses on multi-modal feature fusion and Memory Network framework focuses on dialogue history reasoning.

The main contributions are summarized as follows:

(1) We exploit the possibility of cognition theory in visual dialogue by depicting an image from both visual and semantic views, which covers a broad range of visual content referred by most of questions in the visual dialogue task;

(2) We propose a hierarchical visual information selection module DualVD, which can select question-adaptive clues for answering diverse questions. It supports explicit visualization in visual-semantic knowledge selection and reveals which modality has more contribution to answer the question;

(3) We propose two novel models for the visual dialogue task by integrating our proposed DualVD module with two typical frameworks: *DualVD-LF* based on Late Fusion framework and *DualVD-MN* based on Memory Network framework. The proposed models outperform state-of-the-art approaches on three visual dialogue datasets, including VisDial v0.9, VisDial v1.0 and Visual-Q, which demonstrates the feasibility and effectiveness of the proposed models.

A previous version of our dual encoding model was published in AAAI 2020 [14]. In this extension version, we extend our DualVD on the memory network for the visual dialogue so that we can pay more attention on "dialogue history reasoning" as well as the "visual reasoning". This also suggests that our DualVD module is complementary with the improvements in dialogue modeling and can be plugged into existing visual dialogue models. We also conduct more in-depth experiments and improve the performance on the visual dialogue task. The proposed dual encoding module shows great generalization ability and can be applied to existing visual dialogue models for complementary benefits.

The remaining of this paper is organized as follows. We briefly review the related works in Section II. We detail the proposed DualVD module two novel models integrated with DualVD in Section III. Experimental settings and results are presented in Section IV. We conclude our work in Section V.

II. Related Work

Visual Question Answering (VQA) focuses on answering arbitrary natural language questions conditioned on an image. The typical solutions in VQA build multi-modal representations upon CNN-RNN architecture [1], [15], [16]. They adopt deep Convolutional Neural Networks (CNNs) to represent images and Recurrent Neural Networks (RNNs) to represent questions. The extracted visual and textual feature vectors are then jointly embedded to infer the answer. One of the key challenges in VQA is to effectively understand and extract visual features that better adapt to the question. Existing approaches incorporate context-aware visual features and the trend for modeling the visual context is progressively from global level to fine-grained level. For example, [15] applies CNN features of the whole image as global context, [17] and [18] adopt patches and salient objects learned by attention mechanism as the region context, and [19] exploits inter-object relationships via graph attention networks to model the relational context. However, how to leverage the external visual-semantic knowledge to learn more informative relational representations and combine them with higher-level visual features for better semantic understanding has not been well exploited yet.

Another emerging line of work represents visual content explicitly by natural language and solves VQA as a reading comprehension problem. In [20], the image is wholly converted into descriptive captions, which preserves information at semantic-level in textual domain. However, this kind of approaches use the generated captions, which could not be correct as we desired, and that they fully abandon the informative and subtle visual features. Besides the specific tasks, our model has notable progress compared to the above approaches. We adopt dual encoding mechanism to provide both appearance-level and semantic-level visual information, so that it incorporates the strong points of the above two kinds of approaches.

Visual Dialogue aims to answer a current question conditioned on an image and dialogue history. Compared with Visual Question Answering task, Visual Dialogue involves multi-round dialogue history as context besides the image and the question. Most existing works are based on late fusion framework and focus on modeling the dialogue history. Sequential co-attention mechanism [10] enables the model to identify question-relevant image regions and dialogue history to keep the dialogue consistency. Reference [9] introduces false response in dialogue history for an adverse critic on the historic error. To investigate semantic dependencies between entities underlying dialogue, [8] introduces an Expectation Maximization (EM) algorithm to infer the dialogue structure and the answers via graph neural networks. Reference [21] proposed a novel image-question-answer synergistic network to value the role of the answer for precise visual dialogue. The most recent works [22]–[24] proposed graph inference or causal intervention to reason about the answer on the image and dialogues.

By contrast to extensive study on modeling dialogue history, the image content has been less studied. Although some works [25], [26] devise attention mechanism to focus on the essential visual features most relevant to the question and dialogue history, the monolithic visual representations used in these models still have limited expressive abilities. Since the referred visual content may change remarkably from visual appearance to high-level semantics through the dialogue, monolithic visual representations are hard to convey such diverse visual clues. Different from existing works merely modeling the appearance, our model is able to adaptively capture visual and semantic information in a hierarchical mode inspired by the Dual-coding theory to provide adequate visual clues for diverse questions in visual dialogue.

Visual Relationship Understanding aims to represent and infer the relationships between two objects in an image, which is critical to improve AI's capacity of combinatorial generalization by learning towards relational visual representations. In the early works [27], shallow geometric relationships (e.g. below, above, and inside) based on spatial information have been explored to improve visual segmentation. Later on, visual relationships have been extended to richer definition [28], including geometric, comparative, composition, interaction, *etc*. One limitation of the above approaches is that they merely represent the visual relationships by rigid-categorized labels, which are difficult to accurately model the subtle relationships conditioned on the contexts. For example, an image of <woman, ride, horse> and another image of <man, ride, motorcycle> have quite different visual appearance and semantics even they belong to the same relationship. Even for the same visual appearance, it may contain different relationships. For example, an image where a boy is kicking a football conveys two different relationships, <boy, kick, football> and <boy, look at, football>. The rigid labels can hardly depict such ambiguity relationships. To solve these problems, [29] proposed to infer relations between all the implicit object-like patch pairs via a plug-and-play MLP module for visual question answering. Reference [30] proposed the MotifNet which mines information from motifs — regularly appearing substructures in scene graphs. Recently, [31] proposed a visual and a semantic module that maps features from the two modalities into a shared space and use language high-level "context" to constrain the visual relationship embeddings, which achieves superior performance on large and imbalanced benchmarks. Our model builds upon the visual relationship model [31] and uses continuous embeddings for relationships instead of discrete labels.

Dense Captioning aims to jointly localize and describe image regions in natural language. It provides fine-grained visual descriptions for each local image region compared with object detection and image captioning. The task generalizes object detection when the descriptions are simple labels and image captioning when one predicted region covers the full image. It simultaneously takes the object detection and description task into account. Previous work on holistic description of visual element [32]–[34] are either limited to salient objects of images, or tend to broadly depict the entire visual scene. These descriptions are far from complete visual understanding. Reference [3] proposed to use dense captions for better interpretation of image content. The model consists of a Fully Convolutional Localization Network and an LSTM based Language Model that produces both bounding boxes for interest regions and associated captions in a single forward pass. In this work, we leverage dense captions to describe the semantic-level visual content for the local relationships between objects. In this way, visual information is represented in linguistic form that is closer to human's cognition.

III. Methodology

The visual dialogue task can be described as follows: given an image I and its caption C, a dialogue history till round t-1, $H_t = \{C, (Q_1, A_1), \ldots, (Q_{t-1}, A_{t-1})\}$, and the current question Q_t, the task is to rank a list of N candidate answers $\mathbb{A} = \{A_1, A_2, \ldots, A_N\}$ and return the best answer A_t to Q_t. In this section, we first introduce the scheme of depicting an image from both visual and semantic perspectives in Section III-A. It covers a broad range of visual content like objects, relationships, global semantics and local semantics. In Section III-B, we introduce the hierarchical feature selection module DualVD, which can adaptively capture question-relevant visual-semantic information. In order to prove the effectiveness of the proposed visual representation, two novel models integrated with DualVD are proposed on

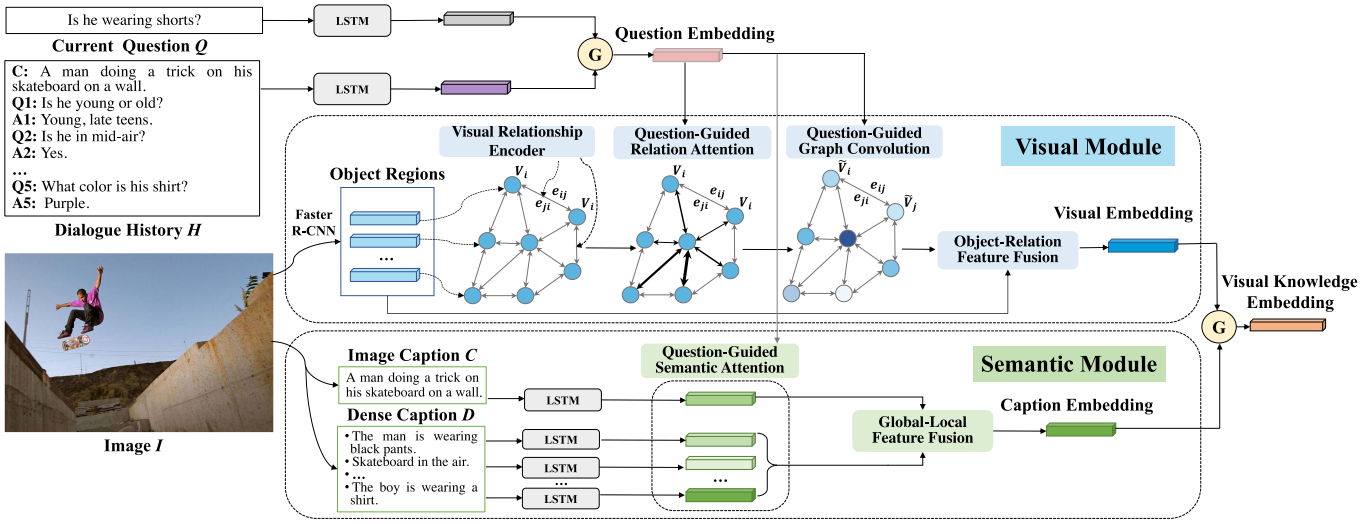


Fig. 2. Overview structure of the DualVD module. It mainly contains two parts: Visual Module and Semantic Module, where "G" represents gate operation.

top of Late Fusion framework in Section III-C and Memory Network framework in Section III-D.

A. Visual-Semantic Dual Encoding

In visual dialogue, two types of information play the primary role to depict an image and answer the diverse questions: visual information and semantic information (Figure 2). For visual information, the major objects and relationships should be kept. In semantic information, higher-level abstraction of the image content should be provided, which involves prior knowledge and complex cognition. Considering the above information together, the agent is capable to answer most questions from various perspectives. In this section, we introduce a dual encoding scheme to generate both visual and semantic representations to depict an image. A scene graph is proposed to represent the visual information while multi-level captions in natural language are leveraged to represent the semantic information. These representations are served as the input of our DualVD module.

1) Scene Graph Construction: Each image is represented as a featured scene graph. Let $V = \{v_i\}^N$ denotes its nodes, which represents object set detected by a pre-trained object detector and let $E = \{e_{ij}\}^{N\times N}$ denotes its edge set, which represents the semantic visual relationships embedded by our visual relationship encoder. We use a pre-trained Faster-RCNN [12] to detect N objects in an image and represent the object v_i as a $2,048$-dimensional vector, denoted by h_i. The visual relationship encoder [31], which is pre-trained on a visual relationship benchmark, i.e. *GQA* [35], encodes relationships between the subject v_i and object v_j as a 512-dimensional relation embedding, denoted as r_{ij}. We assume that certain relationship exists between any pair of objects by considering "unknown-relationship" as a special kind of relationship. Therefore, the constructed scene graph is fully-connected.

The detailed structure of the visual relationship encoder [31] is illustrated in Figure 3. It embeds the relationships between

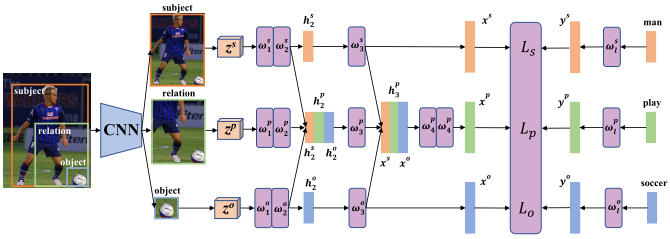


Fig. 3. The framework of visual relationship encoder.

the subject and the object into a semantic space which is aligned with their corresponding descriptions in natural language. Such continuous representations instead of discrete labels can preserve the discriminative capability and contextual awareness. The encoder consists of two parts: visual module and semantic module. In visual module, the input is the region features of subject z^s, relation z^p and object z^o and the outputs are the representations of subject x^s, relation x^p and object x^o. In semantic module, the input is the label of subject, relation and object encoded by a shared GRU, and the outputs are the corresponding embeddings y^s, y^p and y^o. Then the visual and semantic embeddings are fed into a series of fully connected layers separately in order to align their representations respectively in their semantic space, learned by the supervision of three triplet losses, denoted as subject loss L_s, relation loss L_p and object loss L_o. More detail of these losses can be found in [31]. The learnt x^p is denoted as the relation embedding r_{ij} between subject v_i and object v_j, which is used as the representation of edge e_{ij} in the constructed scene graph.

2) Multi-Level Image Captions: We propose to represent the semantics of an image by its captions from both global and local levels. The advantage of captions compared to visual features mainly lies in that captions are represented by natural language with high-level semantics, which can provide straightforward clues for the questions without "heterogenous gap". Global image caption C (provided by the

dataset) is beneficial to response to questions exploring the whole scene. Meanwhile, dense captions [3], denoted as $Z = \{z_1, z_2, \ldots, z_k\}$ (k is the number of dense captions), provide a set of local-level semantics, including the object properties (position, color, shape, *etc.*), the prior knowledge related to the objects (weather, species, emotion, *etc.*), and the relationships between objects (interactions, spatial positions, comparison, *etc.*). Dense captions can support answering questions referring to the local parts or properties. The words in both C and Z are represented by concatenated GloVe [36] and ELMo [37] word embeddings. Then C and Z are separately encoded with two different LSTMs [38], denoted as $\widetilde{C}$ and $\widetilde{Z} = \{\widetilde{z}_1, \widetilde{z}_2, \ldots, \widetilde{z}_k\}$, respectively.

B. Adaptive Visual-Semantic Knowledge Selection

On top of the visual and semantic image representations, we propose a novel feature selection framework to adaptively select question-relevant information from an image. Under the guidance of the current question, the feature selection process is devised in a hierarchical mode as shown in Figure 2: intra-modal selection first captures the visual and semantic information respectively from the *visual module* and *semantic module*; then inter-modal selection obtains the desired visual knowledge from both the visual module and semantic module via *selective visual-semantic fusion*. The advantages of such hierarchical framework is that it can explicitly reveal the progressive feature selection mode and preserve fine-grained information as much as possible.

1) Visual Module: This module is presented on the top of Figure 2. Based on the constructed scene graph introduced in Section III-A1, we aim to select question-relevant relation information and object information. For relation information, we propose a relation-based graph attention network to enrich the object representations with question-aware relationships. It mainly consists of two units: *Question-Guided Relation Attention* highlights the critical relationships and *Question-Guided Graph Convolution* enriches the object features by its relation-critical neighbors. For object information, we highlight the most informative objects to answer the question. Finally, the clues of objects and relationships are further fused in *Object-Relation Information Fusion* to obtain the question-relevant visual content.

a) Question-guided relation attention: The question-guided relation attention examines all the relationships to highlight the ones most relevant to the question. As shown in Figure 4, we first select question-relevant information from the dialogue history to merge into the question representation via a gate operation, which is defined as:

$$gate_t^q = \sigma(\mathbf{W}_q[\widetilde{H}_t, \widetilde{Q}_t] + b_q) \tag{1}$$

$$\widetilde{Q}_t^g = \mathbf{W}_1(gate_t^q \circ [\widetilde{H}_t, \widetilde{Q}_t]) + b_1 \tag{2}$$

where "$[\cdot,\cdot]$" denotes concatenation, "$\circ$" denotes the element-wise product. Each word is represented by concatenating the hidden states extracted from pre-trained GloVe and ELMo models. Then dialogue history H_t and the current question Q_t are separately encoded with the final hidden states of two different LSTMs, denoted as $\widetilde{H}_t$ and $\widetilde{Q}_t$, respectively.

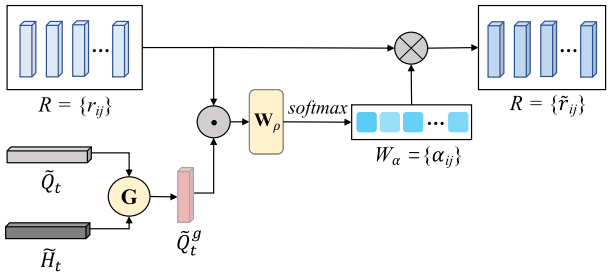


Fig. 4. The illustration of *Question-Guided Relation Attention*, where "G" represents gate operation.

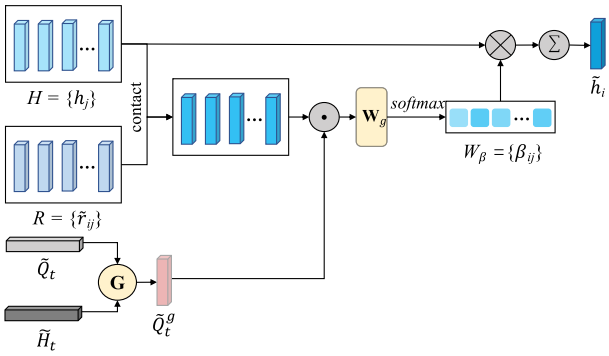


Fig. 5. The illustration of *Question-Guided Graph Convolution*, where "G" represents gate operation.

$gate_t^q$ is a vector of gate values over $\widetilde{H}_t$ and $\widetilde{Q}_t$, $\mathbf{W}_1$ (as well as $\mathbf{W}_2, \ldots, \mathbf{W}_7$ mentioned below) is the linear transformation layer and $\widetilde{Q}_t^g$ is the question features which fused with history information.

The attention weights α_{ij} of the visual relationship r_{ij} is calculated under the guidance of the question $\widetilde{Q}_t^g$:

$$\alpha_{ij} = softmax(\mathbf{W}_\rho(\mathbf{W}_2\widetilde{Q}_t^g \circ \mathbf{W}_3 r_{ij}) + b_r) \tag{3}$$

Each relation embedding is updated based on the attention importance. Formally defined as:

$$\widetilde{r}_{ij} = \alpha_{ij} r_{ij} \tag{4}$$

where $\widetilde{r}_{ij}$ is the question-guided relation embedding.

b) Question-guided graph convolution: This module further updates each object's representation under the guidance of questions by aggregating information from its neighborhood and the corresponding relationships. As shown in Figure 5, given the feature h_j of object v_j and its relation embedding $\widetilde{r}_{ij}$, the attention value of v_j w.r.t. v_i is calculated as:

$$\beta_{ij} = softmax(\mathbf{W}_g(\widetilde{Q}_t^g \circ (\mathbf{W}_4[h_j, \widetilde{r}_{ij}])) + b_g) \tag{5}$$

The obtained attention values for all the neighbors of v_i are used to compute a linear combination of their features, which serves as the updated representation $\widetilde{h}_i$ for v_i:

$$\widetilde{h}_i = \sum_{j=1}^{N} \beta_{ij} h_j \tag{6}$$

Since the scene graph is a fully connected graph, the number of neighbors N for each object is equal to the number of objects detected in each image.

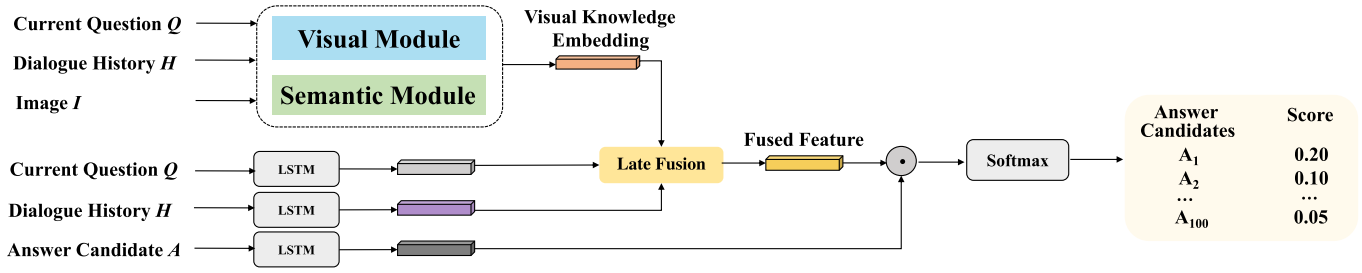


Fig. 6. Overview structure of the DualVD-LF model for visual dialogue.

c) Object-relation feature fusion: In visual dialogue, the object appearance and the visual relationships will contribute to inferring the answer, but with different contributions. In this module, we adaptively fuse question-relevant object features from both original object feature h_i and relation-aware object feature $\widetilde{h}_i$ again by a gate, which is defined by:

$$gate_i^v = \sigma(\mathbf{W}_v[h_i, \widetilde{h}_i] + b_v) \tag{7}$$

$$\widetilde{h}_i^g = \mathbf{W}_5(gate_i^v \circ [h_i, \widetilde{h}_i]) + b_5 \tag{8}$$

where $\widetilde{h}_i^g$ is the updated representation of object v_i. The whole image representation $\widetilde{I}$ is obtained as the weighted sum of the object representations. In order to strengthen the influence of the current question Q_t and the original object features on the retrieved visual clues, we calculate the attention value γ_i^v for h_i under the guidance of Q_t:

$$\gamma_i^v = softmax(\mathbf{W}_s(Q_t \circ (\mathbf{W}_6 h_i)) + b_s) \tag{9}$$

Then the whole representation of the image $\widetilde{I}$ can be updated by:

$$\widetilde{I} = \sum_{i=1}^{N} \gamma_i^v \widetilde{h}_i^g \tag{10}$$

2) Semantic Module: This module aims to select and merge question-relevant semantic information from global and local captions with a *Question-Guided Semantic Attention* module and a *Global-Local Information Fusion* module. The semantic module is located in the middle of Figure 2.

a) Question-guided semantic attention: The semantic attention mechanism highlights relevant captions at both global-level and local-level. This type of attention is guided by the current question which is enhanced with corresponding information from the dialogue history (as introduced above). According to the attention distribution, we enrich the caption representations in order to better adapt to the question. The attention value for each caption in $m_i \in \{\widetilde{C}, \widetilde{z}_1, \widetilde{z}_2, \ldots, \widetilde{z}_k\}$ is calculated as follows:

$$\delta_i^q = softmax((\mathbf{W}_{z1}\widetilde{Q}_t^g + b_{z1})^T(\mathbf{W}_{z2}m_i + b_{z2})) \tag{11}$$

The caption representation for $\widetilde{C}$ and $\widetilde{Z}$ will be updated to $\widetilde{C}^q$ and $\widetilde{Z}^q$:

$$\widetilde{C}^q = \delta_1^q \widetilde{C} \tag{12}$$

$$\widetilde{Z}^q = \sum_{i=2}^{k+1} \delta_i^q \widetilde{z}_{i-1} \tag{13}$$

b) Global-local feature fusion: Some questions are global-related while others are local-related. This step adaptively selects the information from the global caption $\widetilde{C}^q$ and local caption $\widetilde{Z}^q$ via a gate operation, which assigns probability to each dimension of $\widetilde{C}^q$ and $\widetilde{Z}^q$ as computed below:

$$gate^c = \sigma(\mathbf{W}_c[\widetilde{C}^q, \widetilde{Z}^q] + b_c) \tag{14}$$

$$\widetilde{T} = \mathbf{W}_7(gate^c \circ [\widetilde{C}^q, \widetilde{Z}^q]) + b_7 \tag{15}$$

where $\widetilde{T}$ is the textural representations for the abstract visual semantics.

3) Selective Visual-Semantic Fusion: According to the dual-coding theory of human cognitive process, when asked to answer a question, human will retrieve either the visual information or the semantic information individually, or both simultaneously. In this module, we design a gate operation to decide the contributions of the two modalities on the answer prediction. The gate operation and the final visual knowledge representation S are calculated as:

$$gate^s = \sigma(\mathbf{W}_s[\widetilde{I}, \widetilde{T}] + b_s) \tag{16}$$

$$S = gate^s \circ [\widetilde{I}, \widetilde{T}] \tag{17}$$

C. DualVD-LF Based on Late Fusion Framework

In this section, we first apply the DualVD module in the typical Late fusion framework [7] to highlight the advantages of the novel visual representation. This model is named as DualVD-LF. It consists of the Late Fusion (LF) encoder and discriminative (softmax) decoder as shown in Figure 6. The encoder first embeds each part in a dialogue tuple $D = \{I, H_t, Q_t\}$. The image I is encoded by the proposed DualVD introduced in Section III-A and Section III-B. The dialogue history H_t is treated as a long string by concatenating the image caption C with previous $t-1$ round dialogues and encoded by an LSTM. The question Q_t is encoded by another LSTM. Then we concatenate the embeddings of $\widetilde{H}_t$ and $\widetilde{Q}_t$ with the visual knowledge representation S (see Equation 17) into a joint input embedding for answer prediction. The decoder ranks all the answers from a set of N candidates $\mathbb{A}$. It first encodes each candidate answer via a shared LSTM. Then a dot product followed by softmax operation is calculated between the joint input embedding and candidates to get the posterior probability over each candicate. We obtain the correct answer by ranking the candidates based on their posterior probabilities.

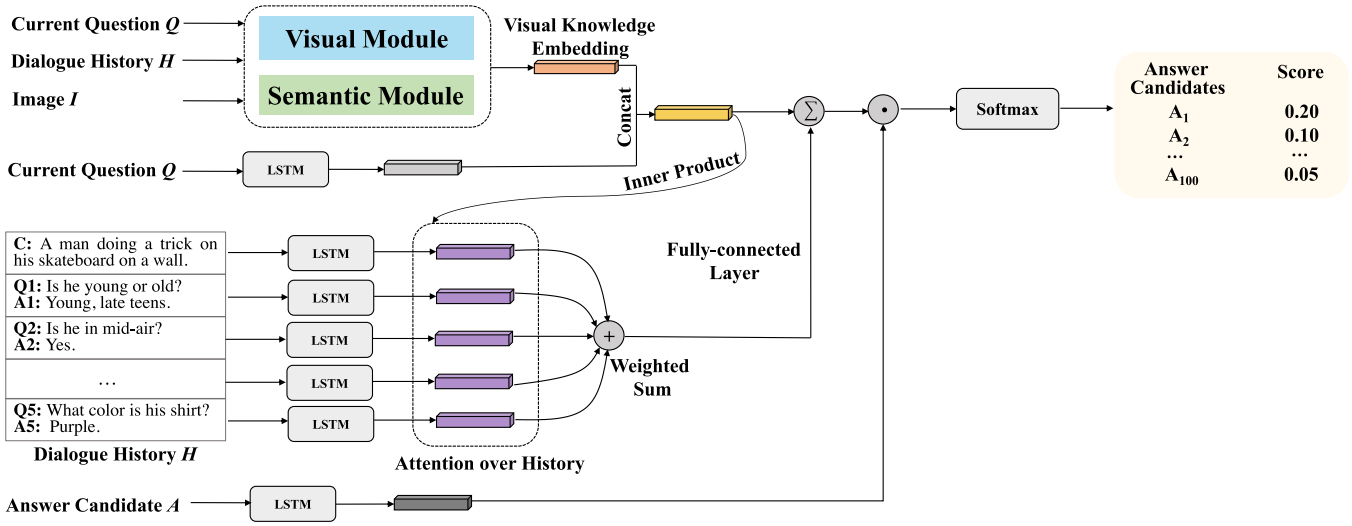


Fig. 7. Overview structure of the DualVD-MN model for visual dialogue.

D. DualVD-MN Based on Memory Network Framework

The dialogue history is an important part in visual dialogue and it has attracted much attention in the literature [8]–[10]. We believe that our work for the visual part and existing works for the dialogue history are complementary with each other. To prove the complementary advantages, we further integrate the DualVD module in the Memory Network (MN) framework [7] so that we can pay more attention on "dialogue history reasoning" as well as the "visual reasoning". This model is named as DualVD-MN and the structure is shown in Figure 7. MN encoder stores previous rounds of dialogues as facts in a memory block and learns to answer the question based on the facts and the image. Specifically, each round of previous dialogue is encoded by a shared LSTM while the question is encoded by another LSTM. We compute the inner product of the concatenation of question embedding and visual knowledge embedding with each round of dialogue embedding followed by a softmax function to obtain the attention probabilities over the dialogue history. We sum up all the embeddings of previous dialogues based on the attention probabilities to get the context representation. We pass the context representation through a fc-layer and add it with the concatenated image-question embedding to get the encoded representation. We use the same discriminative decoder as DualVD-LF introduced in Section III-C.

IV. Experiments

In this section, we first present the quantitative comparisons between our models (*i.e.* DualVD-LF and DualVD-MN) and state-of-the-art works in Section IV-A. Further, an ablation study shows the influence of the key modules in our model in Section IV-B. The interpretability of visual module and semantic module for answering different types of questions are analyzed by visualization and empirical study in Section IV-C. Afterwards, to further prove the benefit of DualVD on more complex decoder, we incorporate it with the generative decoder and show the performance in Section IV-D. Finally, we analyze the influence of essential parameters in our model in Section IV-E.

Datasets: We conducted extensive experiments on three datasets, including VisDial v0.9 [7], VisDial v1.0 [7] and VisDial-Q [39]. For VisDial v0.9 and VisDial v1.0 datasets, the dialogues are collected from the chatting history about images in MSCOCO dataset by two users in Amazon Mechanical Turk (AMT). The dialogues are split into "train", "val" and "test" sets. In "train" and "val", each image is accompanied with a 10-round dialogue, while in "test", each image is followed by a random rounds of question-answer pairs and an on-going question for answer prediction. VisDial v1.0 is an upgraded version of VisDial v0.9. The difference between the two datasets lies in the data source and splits. For VisDial v0.9, all the splits are built on MSCOCO images [40]. For VisDial v1.0, all the splits of VisDial v0.9 serve as "train" (120k), while "val" (2k) and "test" (8k) consist of dialogues on an extra 10k COCO-like images from Flickr.

VisDial v0.9 and VisDial v1.0 datasets are particularly for evaluating the question answering performance of a visual dialogue system. To assess the performance of a visual dialogue system in asking questions, Jain *et al*. further proposed the VisDial-Q dataset [39]. VisDial-Q dataset is built upon VisDial v0.9 and its spliting is the same as VisDial v0.9. Given an image I and its caption C, a dialogue history till round t-1 and $H_t = \{C, (Q_1, A_1), \ldots, (Q_t, A_t)\}$, the VisDial-Q task is to rank a list of 100 candidate questions to select the next round question Q_{t+1}.

Evaluation Metrics: We follow the metrics in [7] and [39] to evaluate the response performance. In the test stage, the model is asked to rank 100 candidate options and evaluated by Mean Reciprocal Rank (MRR), Recall@k ($k = 1, 5, 10$) and Mean Rank of human response (Mean) on all the datasets. For VisDial v1.0, Normalized Discounted Cumulative Gain (NDCG) is added as an extra metric for more comprehensive analysis. Lower value for Mean and higher value for other metrics are desired.

TABLE I
RESULT COMPARISON ON VALIDATION SET OF VISDIAL V0.9

Model	MRR	R@1	R@5	R@10	Mean
LF[7]	58.07	43.82	74.68	84.07	5.78
HRE[7]	58.46	44.67	74.50	84.22	5.72
HREA[7]	58.68	44.82	74.81	84.36	5.66
MN[7]	59.65	45.55	76.22	85.37	5.46
SAN-QI[17]	57.64	43.44	74.26	83.72	5.88
HieCoAtt-QI[42]	57.88	43.51	74.49	83.96	5.84
AMEM[43]	61.60	47.74	78.04	86.84	4.99
HCIAE[44]	62.22	48.48	78.75	87.59	4.81
SF[39]	62.42	48.55	78.96	87.75	4.70
CoAtt[10]	63.98	50.29	80.71	88.81	4.47
CorefMN[45]	**64.10**	**50.92**	80.18	88.81	4.45
VGNN[8]	62.85	48.95	79.65	88.36	4.57
DualVD-LF	62.94	48.64	80.89	89.94	4.17
DualVD-MN	63.12	48.89	**81.11**	**90.33**	**4.12**

TABLE II
RESULT COMPARISON ON TEST-STANDARD SET OF VISDIAL V1.0

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF[7]	55.42	40.95	72.45	82.83	5.95	45.31
HRE[7]	54.16	39.93	70.47	81.50	6.41	45.46
MN[7]	55.49	40.98	72.30	83.30	5.92	47.50
LF-Att[7]	57.07	42.08	74.82	85.05	5.41	40.76
MN-Att[7]	56.90	42.43	74.00	84.35	5.59	49.58
CorefMN[45]	61.50	47.55	78.10	88.80	4.40	54.70
VGNN[8]	61.37	47.33	77.98	87.83	4.57	52.82
RvA[46]	63.03	49.03	80.40	89.83	4.18	55.59
DL-61[21]	62.20	47.90	80.43	89.95	4.17	**57.32**
DualVD-LF	63.23	49.25	80.23	89.70	4.11	56.32
DualVD-MN	**63.38**	**49.35**	**81.05**	**90.38**	**4.07**	57.09

TABLE III
RESULT COMPARISON ON VALIDATION SET OF VISDIAL-Q

Model	MRR	R@1	R@5	R@10	Mean
LF[7]	18.45	7.80	26.12	40.78	20.42
MN[7]	39.83	25.80	54.76	69.80	9.68
SF-QI[39]	30.21	17.38	42.32	57.16	14.03
SF-QIH[39]	40.60	26.76	55.17	70.39	9.32
VGNN[8]	41.26	27.15	56.47	**71.97**	**8.86**
DualVD-LF	41.31	27.24	56.50	71.51	9.09
DualVD-MN	**41.34**	**27.27**	**56.60**	71.45	9.15

Implementation Details: For the textual part, the maximum sentence length of the dialogue history, dense captions and the current question are all set to 20. The hidden state size of all the LSTM blocks is set to 512. We use Faster-RCNN with the ResNet-101 to detect object regions and extract the 2048-dimensional region features. Since some captions with low confidence are likely to introduce unexpected noise and too many captions will decrease the computation efficiency, we select the top 6 dense captions in our model and the corresponding selection strategy is detailed in Section IV-E. We train all of our models by Adam optimizer with 16 epochs, where the mini-batch size is 15 and the dropout ratio is 0.5. In the training process, we combine warm up strategy and cosine annealing learning strategy together to learn the model parameters. For the setting of learning rate, we first warm up the model for 2 epoches with warm-up factor 0.2 and initial learning rate 1×10^{-3}. Then we adopt cosine annealing learning strategy [41] with initial learning rate $\eta_{max} = 1 \times 10^{-3}$ and termination learning rate $\eta_{max} = 3.4 \times 10^{-4}$. The learning rate η_t is computed as follows:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{T_{cur}}{T_{max}}\pi)) \qquad (18)$$

where T_{max} represents maximum number of iterations, T_{cur} represents the number of current iteration. All the experiments are conducted on an NVIDIA Titan X GPU with 12 GB memory and implemented in PyTorch.

A. State-of-the-Art Comparison

In Table I and Table II, we compare our models (DualVD-LF and DualVD-MN) with state-of-the-art discriminative models on VisDial v0.9 and VisDial v1.0, respectively. Both DualVD-LF and DualVD-MN consistently outperform all the approaches on most metrics, which highlights the importance of visual understanding from visual and semantic modules in visual dialogue. It's obvious that DualVD-MN gains further improvement compared with MN baseline [7] and outperforms other models [7], [45] enhanced by complex attention mechanism over the dialogue history. It proves the complementary contribution of our model on the existing "dialogue history" models. The representative CoAtt and HeiCoAtt-QI are relevant to our model in the sense that they leverage attention mechanism to identify question-relevant visual features. However, they ignore the semantic-rich relationships and language priors. We consider both object-relation visual information and global-local semantic information relevant to the question and achieve richer image understanding.

DL-61 [21] is a two-stage network for candidate selection and re-ranking, which aims to achieve precise ranking of the candidate answers. We haven't done much research on the ranking and we simply ranks all the candidates in the decoder following the typical solutions. This is why NDCG of DL-61 is relatively high compared with our model. To notice that our model and the compared approaches all belong to single-step models. With the success of multi-step reasoning, the recent proposed ReDAN [26] achieves 1% boost over our model on most metrics. We believe that stacking our visual encoder to achieve multi-step visual understanding is a promising future work. Another recent work FGA [25] utilizes complex attention mechanism to operate on all the data parts in visual dialogue to extract details while we mainly focus on fine-grained visual representation. The aforementioned models [21], [25], [26] and our DualVD deal with different problems in visual dialog. Our model for the visual part can incorporate with existing works for the dialogue or answer parts for complementary advantages.

We further compare our models with state-of-the-art discriminative models on the VisDial-Q dataset. We also test the performance of the MN baseline and LF baseline on this dataset to clearly prove the benefits when incorporated with DualVD. In Table III, we can see that both DualVD-LF and DualVD-MN consistently outperform their baseline models at all the metrics, which proves the effectiveness of our model in improving the ability of a visual dialogue system in asking questions. Moreover, DualVD-LF and DualVD-MN outperform other existing approaches at most metrics. The proposed

TABLE IV
ABLATION STUDY ON VALIDATION SET OF VISDIAL V1.0

Model	MRR	R@1	R@5	R@10	Mean	NDCG
ObjRep	63.84	49.83	81.27	90.29	4.07	55.48
RelRep	63.63	49.25	81.01	90.34	4.07	55.12
VisNoRel	63.97	49.87	81.74	90.60	4.00	56.73
VisMod	64.11	50.04	81.78	90.52	3.99	56.67
GlCap	60.02	45.34	77.66	87.27	4.78	50.04
LoCap	60.95	46.43	78.45	88.17	4.62	51.72
SemMod	61.07	46.69	78.56	88.09	4.59	51.10
w/o ELMo	63.67	49.89	80.44	89.84	4.14	56.41
DualVD-LF	**64.64**	**50.74**	**82.10**	**91.00**	**3.91**	**57.30**
DualVD-MN	**64.70**	**50.79**	**82.41**	**91.10**	**3.90**	**58.24**

models are slightly inferior to VGNN [8], which adopts graph neural network to model the semantic dependency in the dialogue history. The advantages of our models are orthogonal to the modeling of dialogue history in VGNN.

B. Ablation Study

We also conduct an ablation study to further exploit the influence of the essential components of DualVD. To be mentioned, we use DualVD-LF as the full model and apply the same descriminative decoder for all the following variations:

Object Representation (ObjRep): this model uses the averaged object features to represent the image. Question-driven attention is applied to enhance the object representations.

Relation Representation (RelRep): this model applies averaged relation-aware object representations as the image representation without fusing with original object features.

Vision Module without Relationships (VisNoRel): this model contains the full Vision Module, differing in that the relation embeddings are replaced by unlabeled edges.

Visual Module (VisMod): this is our full visual module, which fuses objects and relation features.

Global Caption (GlCap): this model uses LSTM to encode the global caption to represent the image.

Local Caption (LoCap): this model uses LSTM to encode the local captions to represent the image.

Semantic Module (SemMod): this is our full semantic module, which fuses global and local features.

w/o ELMo: this is our full model based on late fusion encoder, differing in that the word embedding GloVe+ELMo is replaced by GloVe.

DualVD-LF (full model): this is our full model, which incorporates both the visual module and semantic module.

DualVD-MN: this model is based on the memory network encoder instead of the late fusion encoder in DualVD-LF. Other settings are the same as DualVD-LF.

Table IV shows the ablation results on VisDial v1.0 validation set. Models in the first block are designed to evaluate the influence of key components in the visual module. The limitation for **ObjRep** is that it only mines the pivotal features from isolated objects and ignores the relational information, which achieves worse performance at all metrics compared to VisMod. **RelRep** considers the relationships by introducing relation embedding for aggregating the object features. However, empirical study indicates that enhancing object relationships while weakening object appearance is still not sufficient to represent the visual semantics for better performance. **VisNoRel** takes a step further by adaptively fusing the information from both object appearance and full-connected neighbors, aggregating all the neighborhood features directly without relation semantics. This strategy achieves slight improvement compared with ObjRep. **VisMod** moves a step further by adaptively fusing the information from both object appearance and full-connected neighbors, aggregating all the neighborhood features with relational information, which achieves the best performance compared to above two models.

Orthogonal to visual part, models in the second block are conducted to test the influence of key components in the semantic part. The overall performance of either **GlCap** or **LoCap** decreases by 1% and 0.15% respectively, compared to their integrated version **SemMod**, which adaptively selects and fuses the task-specific descriptive clues from both global-level and local-level captions.

We compare the performance of VisMod and SemMod with DualVD-LF. By adaptively select information from the visual and the semantic module, **DualVD-LF** results in a great boost compared to SemMod and a relatively slight boost compared to VisMod. This unbalanced boost indicates that visual module provides comparatively richer clues than semantic module. Combining the two modules together gains an extra boost, because of the complementary information derived from different modalities. By paying more attention on "dialogue history reasoning" with Memory Network encoder, **DualVD-MN** achives further improvement compared with DualVD-LF. This also suggests that our DualVD module is complementary with our improvements in visual dialogue and can be plugged into existing visual dialogue models. The model **w/o ELMo** aims to remove ELMo from DualVD-LF to evaluate how much the model benefits from the pre-training knowledge. The results decrease slightly compared with DualVD-LF, which proves that the improvement of the performance mainly comes from the contribution of the novel visual representation.

C. Interpretability

A critical advantage of DualVD lies in its interpretability: DualVD is capable to predict the attention weights in the visual module, semantic module and the gate values in visual-semantic fusion. It supports explicit visualization and can reveal DualVD's mode in information selection. We show the visualization for success cases and failure cases of DualVD-LF model in Figure 8 and Figure 9, respectively. Four meaningful observations of our model are presented as follows:

1) Comprehensive Visual-Semantic Clues: The visual features at object-level, relationship-level, and semantic-level are preserved in the framework of DualVD, which enables the DualVD-LF model to answer a wide range of visually grounded questions through the dialogue. For instance, in Figure 8, the third example (third and fourth rows in Figure 8) shows the attention visualization of three round of dialogues about an image, which depicts "*2 boys playing*

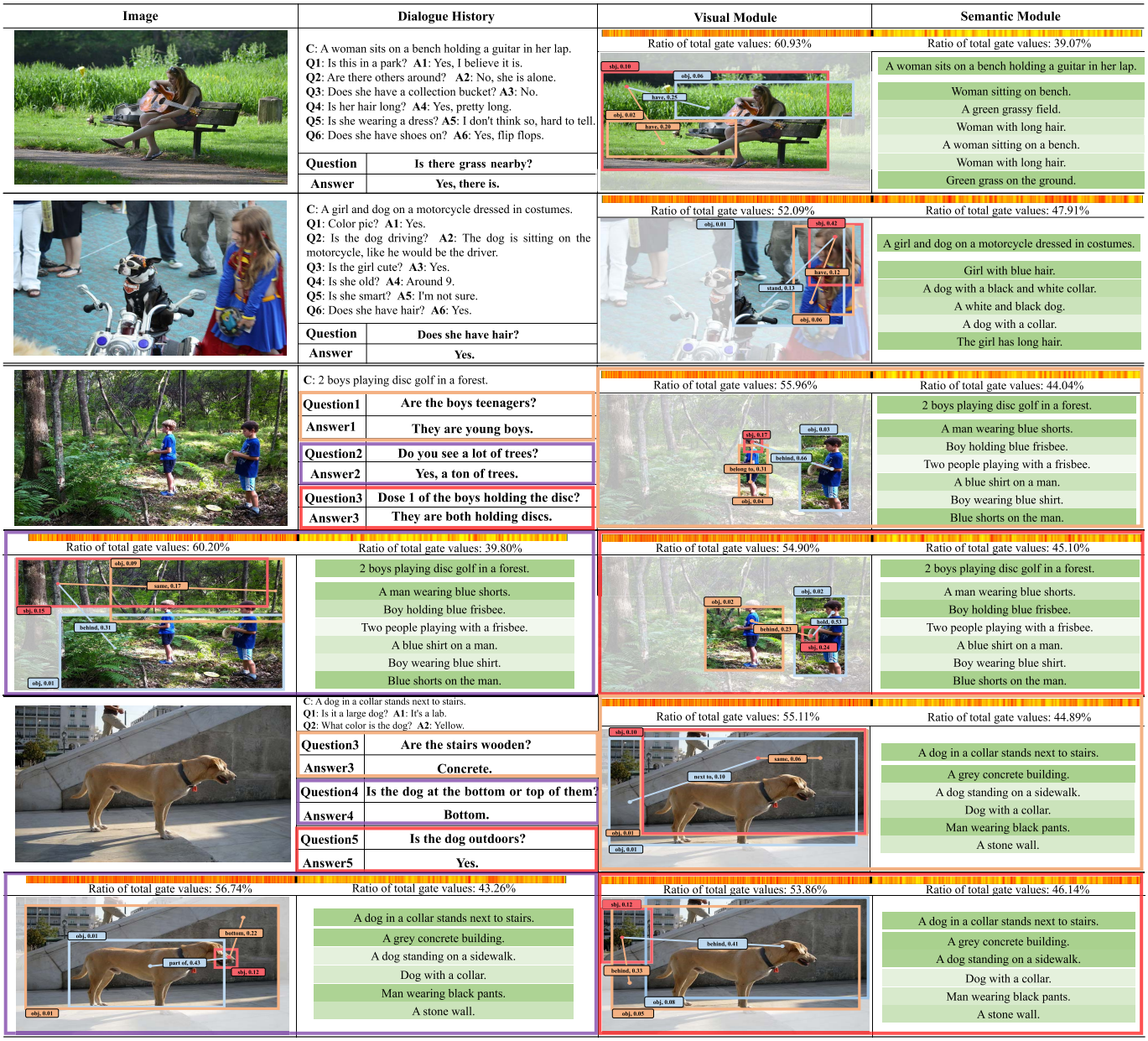


Fig. 8. Visualization for success cases. Visual module highlights the most relevant subject (red box) according to attention weights of each object (γ_i^v in Eq. 9) and the objects (orange and blue boxes) with the top two attended relationships (β_{ij} in Eq. 5). Semantic module shows the attention distribution (δ_i^q in Eq. 11) over the global caption (first row) and the local captions (rest rows), where darker green color indicates bigger attention weight. The yellow thermogram on the top visualizes the gate values ($gate^s$ in Eq. 16) of the visual embedding (left) and the caption embedding (right) in visual-semantic fusion.

disc golf in a forest". The three questions in this example respectively refer to the foreground (the boys), the background (the trees) and the relationships (holding). As we can see, in each round of dialogue, the model is capable to capture the most relevant visual and semantic information regarding the current question. For example, in the first question "*Are the boys teenagers?*", the visual module highlights the face of a boy and the relationships to his body and the other boy, while the semantic module puts more attention on the captions describing the two boys, which all provide useful clues to infer the correct answer. In the second and third round of dialogues, the model respectively attends to the whole trees and the discs accurately. We have the same observation for the fourth example (fifth and sixth rows in Figure 8). These examples prove that DualVD is capable to adaptively focus on question-relevant information through the dialogue and this explains why the correct answer is selected.

2) Information Selection Mode: By selecting visual and semantic information by gating mechanism, the DualVD-LF model can reveal the mode in information selection for answering the current question by visualizing the gate values. We observe that the amount of information derived from each module highly depends on the complexity of the question and the relevance of the module content. More information will come from the semantic module when the question involves complex relationships or the semantic module explicitly

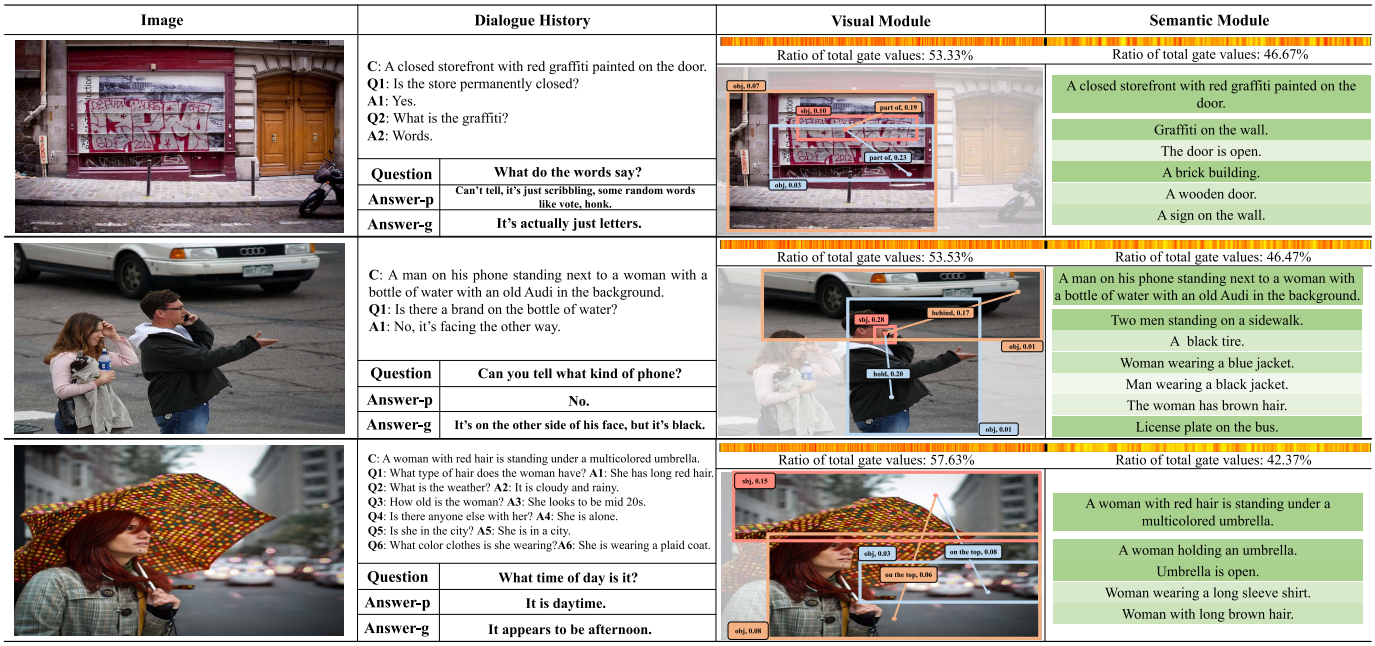


Fig. 9. Visualization for three different failure cases. "Answer-p" represents the predicted answer via DualVD and "Answer-g" represents the ground truth answer provided by VisDial v1.0 dataset. The attention weights and gate values are visualized in the same way as that in Figure 8.

contains question-relevant clues. We show two examples with a current question and the dialogue history (first two rows in Figure 8) to reveal DualVD's mode in information selection. In Figure 8, *ratio of total gate values* reveals the amount of information derived from each module. We denote the sum of all the gate values for the visual dimensions and semantic dimensions in Equation 16 as G_v and G_s, respectively. We denote the sum of all the gate values for all the dimensions in Equation 16 as G. The *ratio of total gate values* for the visual module and semantic module is defined as $\frac{G_v}{G}$ and $\frac{G_s}{G}$, respectively. In the first example (first row in Figure 8), more visual information is required compared to semantic information. Similar observation exists for the second question in the third example. Such questions that referring to object appearance depend more clues from the visual module. In the second example (second row in Figure 8), the current question is about the relationship between the girl and the hair. The amount of semantic information remarkably increases since there exists explicit evidence "*The girl has long hair*". This observation holds for the third question in the third example. Since language is a higher-level encoding of the visual content after complex reasoning involved with prior knowledge, it is able to provide more useful clues for semantic-level questions.

3) The Critical Role of Visual Information: From the visualization results, we find that the visual information is more important than the semantic information in question answering. In all the testing cases, the ratio of gate values of the visual module is larger than that of the semantic module and also that the differences between the two are not greatly disproportionate. This demonstrates that more comprehensive and accurate clues come from the visual information, though the semantic information can also provide useful clues. This observation is quite different from previous work [20], which has shown that VQA is dominated by language information. In [20], QANet aims to incorporate external knowledge to answer questions beyond the image. By representing the image as text, QANet unifies the representations of the image and the knowledge facts into the natural language space, which avoids multimodal feature fusion and can straightforwardly incorporate image-relevant knowledge for answer prediction. In our task, the questions in visual dialogue are mostly visual-grounded, which requires detailed visual information as the main evidence. We believe that using our DualVD model to achieve both semantic-level and visual-level advantages in knowledge-based VQA tasks is a promising future work.

4) Failure Case Analysis: Since the visualization results explicitly show the visual and semantic clues the model focused on, it provides us evidence to analysis the reasons for the failure cases. In Figure 9, we show three failure cases from three typical failure modes. We can see that our model can accurately attend to the visual and the semantic information that relevant to the questions in all the cases. In the first example, the ground-truth answer is similar to the predicted answer. This mode is so-called *one-to-many problem* in the dialogue system. A second failure mode is due to the incomplete visual information in the image. In the second example, the phone is obscured by the face, which provides limited clues to predict an exact answer for the question "Can you tell what kind of phone?". The third mode is due to the lack of visual clues for more accurate answer. In the third example, our model accurately focuses on the "*sky*" and predicts "*It is daytime*". However, the image doesn't provide extra evidence to infer more precise time, i.e. "*afternoon*" in the ground-truth answer.

TABLE V
GENERATIVE MODEL COMPARISON ON VALIDATION SPLIT OF VISDIAL V1.0

Model	MRR	R@1	R@5	R@10	Mean	NDCG
MN-G[7]	47.83	38.01	57.49	64.08	18.76	56.99
HCIAE-G[44]	49.07	39.72	58.23	64.73	18.43	59.70
CoAtt-G[10]	49.64	**40.09**	59.37	65.92	17.86	59.24
ReDAN-G[26]	49.60	39.95	59.32	65.97	17.79	59.41
DualVD-LF-G	**49.78**	39.96	**59.96**	**66.62**	**17.49**	**60.08**

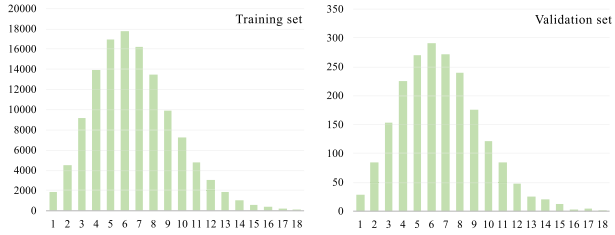


Fig. 10. Analysis of dense caption distribution on "train"and "val" splits of VisDial v1.0. x-axis represents the number of dense captions while y-axis represents the number of images generating corresponding number of captions.

TABLE VI
RESULT OF DIFFERENT T_l OF DUALVD ON VISDIAL 1.0

Model	MRR	R@1	R@5	R@10	Mean	NDCG
$T_l = 4$	64.38	50.44	81.79	90.58	3.97	**57.34**
$T_l = 6$	**64.64**	**50.74**	**82.10**	**91.00**	**3.91**	57.30
$T_l = 8$	64.35	50.28	81.94	90.69	3.94	56.28
$T_l = 10$	64.34	50.37	81.74	90.71	3.97	57.09

D. Performance on Generative Decoder

To further prove the generalization of our model on more complex decoder, we report the performance of DualVD-LF and the state-of-the-art models under the same generative decoder [7], which is more practical in realistic applications. The generative decoder is an LSTM language model taking the encoded embedding as the initial state. During the training process, we maximum the log-likelihood of the encoded representation of the ground truth answer. In the testing process, we use the log-likelihood scores to rank the candidate answers. The experiments were conducted on the validation split of VisDial v1.0 for fair comparison with tate-of-the-art models, including MN-G [7], HCIAE-G [44], CoAtt-G [10] and ReDAN-G (T=1, T is the number of infer step) [26]. As shown in Table V, our model DualVD-LF-G outperforms all the existing approaches at all the metrics, with merely slight inferiority at R@1 compared with CoAtt-G. It proves the effectiveness of the novel visual representation DualVD when incorporated with the descriminative decoder.

E. Parameter Analysis

We conducted experiments to exploit the influence of the number of local captions on the visual dialogue performance, since local captions with low confidence are likely to introduce unexpected noise and too many captions will decrease the computation efficiency. We analyzed the distribution of the number of local captions on the "train" and "val" split of the VisDial v1.0 dataset. As shown in Figure 10, in both "train" and "val" splits, the distribution of dense caption approximates normal distribution. The mean value of the distributions approximates 6. Thus, we range the number of local captions T_l from 4 to 10 with step length 2 and show the experimental performance of DualVD-LF on the validation split of VisDial v1.0 in Table VI. The model gains superior performance when $T_l = 6$. This because that the semantic information is not sufficient when $T_l = 4$ while the semantic information may contains unexpected noise when $T_l = 8, 10$. Thus, we set the number of local captions to 6 in our models.

V. CONCLUSION

In this paper, inspired by the dual-coding theory in cognitive science, we proposed a novel DualVD module for visual dialogue. DualVD mainly consists of a visual module and a semantic module, which encodes image information at appearance-level and semantic-level, respectively. Desired clues for answer inference are adaptively selected from the two modules via gate mechanism. To evaluate the effectiveness of our module, we integrate DualVD with the Late Fusion framework and Memory Network framework for the visual dialogue task. Results from extensive experiments on three benchmarks demonstrate that deriving visual information from visual-semantic representations can achieve superior performance compared to other state-of-the-art approaches. Another major advantage of DualVD lies in its interpretability via progressive visualization. It can give us insights of how information from different modalities is used for inferring the answers. How to incorporate DualVD with various existing models, such as transformer-based models, to further prove its effectiveness will be our future work.

ACKNOWLEDGMENT

The authors would like to thank Zilong Zheng, in the Department of Computer Science at UCLA, for his insightful advice on the experiments.

REFERENCES

[1] A. Agrawal *et al.*, "VQA: Visual question answering," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 4–31, May 2017.
[2] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. V. D. Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1960–1968.
[3] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4565–4574.
[4] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 203–212.
[5] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improved visual-semantic embeddings with hard negatives," in *Proc. BMVC*, 2018, pp. 2–10.
[6] J. Yu *et al.*, "Modeling text with graph convolutional network for cross-modal information retrieval," in *Proc. PCM*, 2018, pp. 223–234.
[7] A. Das *et al.*, "Visual dialog," in *Proc. CVPR*, Jul. 2017, pp. 1080–1089.
[8] Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6669–6678.

[9] T. Yang, Z.-J. Zha, and H. Zhang, "Making history matter: history-advantage sequence training for visual dialog," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2561–2569.

[10] Q. Wu, P. Wang, C. Shen, I. Reid, and A. V. D. Hengel, "Are you talking to me? Reasoned visual dialog generation through adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6106–6115.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[13] A. Paivio, *Imagery and Verbal Process*. New York, NY, USA: Holt, Rinehart and Winston, 1971.

[14] X. Jiang *et al.*, "Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue," in *Proc. AAAI*, 2020, pp. 11125–11132.

[15] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. NeurIPS*, 2015, pp. 2953–2961.

[16] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.

[17] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[18] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[19] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10313–10322.

[20] H. Li, P. Wang, C. Shen, and A. V. D. Hengel, "Visual question answering as reading comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6319–6328.

[21] D. Guo, C. Xu, and D. Tao, "Image-question-answer synergistic network for visual dialog," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10434–10443.

[22] F. Chen, F. Meng, J. Xu, P. Li, B. Xu, and Z. Jie, "Dmrm: A dual-channel multi-hop reasoning model for visual dialog," in *Proc. AAAI*, 2020, pp. 7504–7511.

[23] D. Guo, H. Wang, H. Zhang, Z.-J. Zha, and M. Wang, "Iterative context-aware graph inference for visual dialog," in *Proc. CVPR*, Jun. 2020, pp. 10055–10064.

[24] J. Qi, Y. Niu, J. Huang, and H. Zhang, "Two causal principles for improving visual dialog," in *Proc. CVPR*, Jun. 2020, pp. 10857–10866.

[25] I. Schwartz, S. Yu, T. Hazan, and A. G. Schwing, "Factor graph attention," in *Proc. CVPR*, Jun. 2019, pp. 2039–2048.

[26] Z. Gan, Y. Cheng, A. E. Kholy, L. Li, and J. Gao, "Multi-step reasoning via recurrent dual attention for visual dialog," in *Proc. ACL*, 2019, pp. 6463–6474.

[27] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, "Multi-class segmentation with relative location prior," *Int. J. Comput. Vis.*, vol. 80, no. 3, pp. 300–316, Dec. 2008.

[28] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. ECCV*, 2016, pp. 852–869.

[29] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proc. NeurIPS*, 2017, pp. 4967–4976.

[30] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5831–5840.

[31] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, "Large-scale visual relationship understanding," in *Proc. AAAI*, 2019, pp. 9185–9194.

[32] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[33] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1473–1482.

[34] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[35] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6700–6709.

[36] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[37] M. Peters *et al.*, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Papers)*, vol. 1, 2018, pp. 2227–2237.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] U. Jain, A. Schwing, and S. Lazebnik, "Two can play this game: Visual dialog with discriminative question generation and answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5754–5763.

[40] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. ECCV*. 2014, pp. 740–755.

[41] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, 2016, pp. 1–16.

[42] M. I. Hasan Chowdhury, K. Nguyen, S. Sridharan, and C. Fookes, "Hierarchical relational attention for video question answering," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 289–297.

[43] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal, "Visual reference resolution using attention memory for visual dialog," in *Proc. NeurIPS*, 2017, pp. 3719–3729.

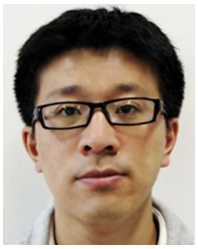
[44] J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra, "Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model," in *Proc. NeurIPS*, 2017, pp. 314–324.

[45] S. Kottur, J. M. Moura, D. Parikh, D. Batra, and M. Rohrbach, "Visual coreference resolution in visual dialog using neural module networks," in *Proc. ECCV*, Sep. 2018, pp. 153–169.

[46] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, and J.-R. Wen, "Recursive visual attention in visual dialog," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6679–6688.

Jing Yu received the B.S. degree in automation science from Minzu University, China, in 2011, the M.S. degree in pattern recognition from Beihang University, China, in 2014, and the Ph.D. degree from the University of Chinese Academy of Sciences, China, in 2019. She is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences. She works on vision and language problems, including visual question answering, visual dialogue, and cross-modal information retrieval.

Xiaoze Jiang received the B.S. degree in automation science from Northeast Forestry University, China, in 2018. He is currently a Graduate Student with Intelligent Computing and Machine Learning Lab, School of Automation Science and Electrical Engineering, Beihang University, China. His research interests include visual dialogue and natural language processing.



Zengchang Qin received the B.S. degree from Heilongjiang University, China, in 2001, and the M.S. and Ph.D. degrees from the University of Bristol, U.K., in 2002 and 2005, respectively. He is currently an Associate Professor with Intelligent Computing and Machine Learning Lab, School of Automation Science and Electrical Engineering, Beihang University, China. His research interests include machine learning, data mining, multimedia learning, and natural language processing.

Weifeng Zhang received the B.S. degree in electronic information engineering from the Beijing University of Technology in 2009, the M.S. degree in pattern recognition from Beihang University in 2012, and the Ph.D. degree in computer science from Hangzhou Dianzi University in 2019. He is currently an Associate Professor with Jiaxing University. His research interests include machine learning and multimedia modeling.

Yue Hu received the Ph.D. degree in computer science from the University of Science and Technology Beijing in 2000. She is currently a Professor and Researcher with the Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include the area of natural language processing and social network analysis.

Qi Wu received the M.Sc. and Ph.D. degrees in computer science from the University of Bath, U.K., in 2011 and 2015, respectively. He is currently an Assistant Professor with The University of Adelaide, where he is also an Associate Investigator with the Australia Centre for Robotic Vision (ACRV). His educational backgrounds primarily include computer science and mathematics. He works on the vision and language problems, including image captioning, visual question answering, and visual dialog. His work has been published in prestigious journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, CVPR, ICCV, AAAI, and ECCV. He is a Fellow of ARC Discovery Early Career Researcher Award (DECRA) from 2019 to 2021.