Recursive Visual Attention in Visual Dialog

Yulei Niu[1,2] Hanwang Zhang[2] Manli Zhang[1] Jianhong Zhang[1] Zhiwu Lu[1*] Ji-Rong Wen[1]
[1]Beijing Key Laboratory of Big Data Management and Analysis Methods
School of Information, Renmin University of China, Beijing 100872, China
[2]Nanyang Technological University, Singapore 639798
{niu, manlizhang, jianhong, luzhiwu, jrwen}@ruc.edu.cn, hanwangzhang@ntu.edu.sg

Abstract

*Visual dialog is a challenging vision-language task, which requires the agent to answer multi-round questions about an image. It typically needs to address two major problems: (1) How to answer visually-grounded questions, which is the core challenge in visual question answering (VQA); (2) How to infer the co-reference between questions and the dialog history. An example of visual co-reference is: pronouns (*e.g.*, "they") in the question (*e.g.*, "Are they on or off?") are linked with nouns (*e.g.*, "lamps") appearing in the dialog history (*e.g.*, "How many lamps are there?") and the object grounded in the image. In this work, to resolve the visual co-reference for visual dialog, we propose a novel attention mechanism called Recursive Visual Attention (RvA). Specifically, our dialog agent browses the dialog history until the agent has sufficient confidence in the visual co-reference resolution, and refines the visual attention recursively. The quantitative and qualitative experimental results on the large-scale VisDial v0.9 and v1.0 datasets demonstrate that the proposed RvA not only outperforms the state-of-the-art methods, but also achieves reasonable recursion and interpretable attention maps without additional annotations. The code is available at https://github.com/yuleiniu/rva.*

1. Introduction

Vision and language understanding has become an attractive and challenging interdisciplinary field in computer vision and natural language processing. Thanks to the rapid development of deep neural networks and the high quality of large-scale real-world datasets, researchers have achieved inspiring progress in a range of vision-language tasks, including visual relation detection [21, 19, 39], image captioning [36, 8, 38, 3], referring expression grounding [24, 25, 40], and visual question answering (VQA) [6, 33, 11, 32]. However, comprehension and reasoning in vision and natural language are still far from being resolved, especially when the AI agent interacts with human in a continuous communication, such as vision-and-language navigation [4] and visual dialog [9].

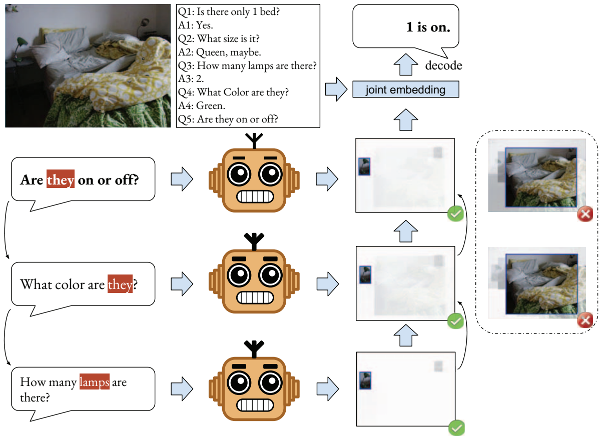

Figure 1. Illustration of the intuition of Recursive Visual Attention in visual dialog. When our dialog agent meets an ambiguous question (*e.g.*, "Are they on or off?"), it will recursively review the dialog history (see the first column) and refine the visual attention (see the third column), until it can resolve the visual co-reference (*e.g.*, How many lamps are there?). The attention maps tagged with green check mark represent reasonable recursive visual attention, while those tagged with red cross mark in the dashed box represent false question-guided visual attention.

Visual dialog is one of the prototype tasks introduced in recent years [9, 10]. It can be viewed as the generalization of VQA, which requires the agent to answer the question about an image [6] or video [33] after comprehending and reasoning out of visual and textual contents. Different from one-round VQA, visual dialog is a multi-round conversation about an image. Therefore, one of the key challenges in visual dialog is visual co-reference resolution, since 98% of dialogs and 38% of questions in the large-scale VisDial dataset have at least one pronoun (*e.g.*, "it", "they", "this", "he", "she") [9]. For example, as illustrated in Figure 1,

[*] Corresponding author.

questions “Are they on or off?” and “What color is it?” contain pronouns that need to be resolved before answering. Recently, researchers have attempted to resolve the visual co-reference using attention memory [29] at a sentence level, or applying the neural module networks [18] at a word level. Specifically, an attention memory [29] is established to store the image attention map at each round, while a reference pool [18] is utilized to keep all the entities recognized from the dialog history. They both apply a soft attention over all the stored visual attentions for refinement. However, humans rarely remember all their previous visual attentions, and only review the topic-related dialog history when they are confused with the ambiguous question.

We expect our dialog agent to *selectively* review the dialog history like us humans during the conversation. For example, as illustrated in Figure 1, “Are they on or off?” is an ambiguous question and the dialog agent needs to resolve “they” before watching the image. The agent then *recursively* browses the dialog history and computes visual attention until it meets the unambiguous description “How many lamps are there?”. One may argue that a natural language parser can achieve this goal by detecting whether there exists a pronoun in the question. However, not all pronouns are needed to be resolved, *e.g.*, “Is it sunny?”. Some abbreviate sentences without context are also ambiguous, *e.g.*, “What color?”. It is thus impractical to exhaust all cases using a natural language parser.

In this work, we formulate visual co-reference resolution in visual dialog as Recursive Visual Attention (RvA). As shown in Figure 1, the agent first infers whether it can ground the visual content based on the current question. If not, the agent will recursively review the topic-related dialog history and refine the visual attention. The recursion termination is that the agent feels “confident” in visual grounding, or it has backtracked to the beginning of dialog history. Thanks to the Gumbel-Max trick [12] and its continuous softmax relaxation [15, 23], our agent can be end-to-end trained when making discrete decisions. In addition, we design two types of language features for different purposes. The *reference-aware* language feature helps with visual grounding and inference of reviewing dialog history, while the *answering-aware* language feature controls which attributes of the image feature should be activated for question answering.

Our main contributions are concluded as follows. First, we propose a novel Recursive Visual Attention (RvA) strategy for the visual co-reference resolution in visual dialog. Second, we carry out extensive experiments on VisDial v0.9 and v1.0 [9], and achieve state-of-the-art performances compared to other methods. Third, the qualitative results indicate that our dialog agent obtains reliable visual and language attention during the reasonable and history-aware recursive process.

2. Related Work

Visual Dialog. Visual dialog is a current vision and language task, which requires the agent to understand the dialog history, ground visual object, and answer the question. Recently, two popular dialog datasets were crowd-sourced on Amazon Mechanical Turk (AMT) [7]. De Vries *et al*. [10] collected GuessWhat dataset from a cooperative two-player game. Given the whole picture and its caption, one player asks questions to locate the selected object, while the other player replies in yes/no/NA. However, the questions are constrained to closed-ended questions. In comparison, Das *et al*. [9] collected VisDial dataset by a different two-person chat style. During the live chat, the “questioner” asks questions to imagine the visual content in the picture based on the caption and chat history, while the “answerer” watches the picture and answer in a free-form way. We apply the second setting in this paper.

Visual Co-reference Resolution. The task of visual co-reference resolution is to link expressions, typically pronoun and noun phrases referring to the same entity, and ground the referent in the visual content. Co-reference resolution has been used to improve visual comprehension in many tasks, such as visual grounding [14], action recognition [27, 28], and scene understanding [17]. Recently Lu *et al*. [22] proposed a history-conditioned attention mechanism to implicitly resolve the visual co-reference. Seo *et al*. [29] used attention memory to store previous image attentions at a sentence level. Furthermore, neural module networks [5] were applied to recognize entities in all the history at a word level [18]. Different from recent works that proposed a soft attention mechanism over all the memorized attention maps [29] or all the grounded entities [18], our proposed recursion predicts *discrete* attention over topic-related history, which is more intuitive and explainable.

3. Approach

In this section, we formally introduce the visual dialog task and our proposed Recursive Visual Attention (RvA) approach. The task of visual dialog [9] is defined as follows. The dialog agent is expected to answer the question q_T at round T by ranking a list of 100 candidate answers $A_T = \{a_T^{(1)}, \cdots, a_T^{(100)}\}$ in a discriminative manner, or producing a sentence in a generative manner. The extra information for visual dialog consists of the image I and the dialog history $H = \{\underbrace{c}_{h_0}, \underbrace{(q_1, a_1)}_{h_1}, \cdots, \underbrace{(q_{T-1}, a_{T-1})}_{h_{T-1}}\}$, where c is the image caption and (q, a) is any question-answer pair.

Next, we first provide an overall structure of RvA in Section 3.1, followed by Section 3.2 introducing the Infer, Pair and Att modules of RvA. The training details of RvA are given in Section 3.3.

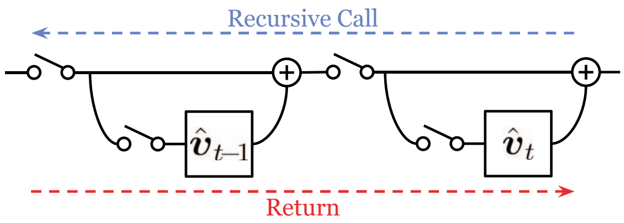

Figure 2. A high-level view of Recursive Visual Attention. The right-to-left direction (dashed blue) represents recursive call, and the left-to-right direction (dashed red) represents visual attention return. The *cond* variable controls the switch on the trunk, while t_p controls the switch on the branch (see Algorithm 1). $\hat{\boldsymbol{v}}_t$ represents the attended feature $\textsc{Att}(\mathcal{V}, \mathcal{Q}, t)$.

3.1. Recursive Visual Attention

Algorithm 1 Recursive Visual Attention

```
function RvA($\mathcal{V}, \mathcal{Q}, \mathcal{H}, t$)
    $cond, \lambda \leftarrow \textsc{Infer}(\mathcal{Q}, t)$
    if $cond$ then
        return $\textsc{Att}(\mathcal{V}, \mathcal{Q}, t)$
    else
        $t_p \leftarrow \textsc{Pair}(\mathcal{Q}, \mathcal{H}, t)$
        return $(1-\lambda) \cdot \textsc{RvA}(\mathcal{V}, \mathcal{Q}, \mathcal{H}, t_p)$
                $+\lambda \cdot \textsc{Att}(\mathcal{V}, \mathcal{Q}, t)$
    end if
end function
```

First of all, the overall structure of the proposed Recursive Visual Attention (RvA) method is shown in Algorithm 1. Here $\mathcal{Q} = \{\boldsymbol{q}_0, \boldsymbol{q}_1, \cdots, \boldsymbol{q}_T\}$ represents the question feature set where the caption feature $\boldsymbol{c}$ is added into the question set as $\boldsymbol{q}_0$, $\mathcal{H} = \{\boldsymbol{h}_0, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_{T-1}\}$ represents the history feature set, and $\mathcal{V} = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K\}$ represents the region feature set. Given any question q_t, our dialog agent first *infers* whether it understands the question q_t for visual grounding. If not, our agent will *pair* the current question q_t with its most related history h_{t_p}, and backtrack to the paired round t_p. This process will be kept executing until the agent can understand the current traced question, or the dialog agent has backtracked to the beginning of the dialog. As a result, our dialog agent recursively modifies the visual attention by adding the question-guided *attended* visual attention at round t and the recursive visual attention at paired round t_p, weighted by a learnable non-negative weight λ. For the question q_T, the output visual attention is formulated by $\boldsymbol{\alpha}_T = \textsc{RvA}(\mathcal{V}, \mathcal{Q}, \mathcal{H}, T)$. The attended visual feature is further calculated by a weighted sum over all the region features $\hat{\boldsymbol{v}}_T = \sum_i \alpha_i \boldsymbol{v}_i$.

In addition, we give a high-level view of Recursive Visual Attention (RvA) in Figure 2. Intuitively, all the switches on both the trunk and branches are initially open (*i.e.*, turned off). Our RvA is recursively called from *present* to *past*, closing (*i.e.*, turning on) the switch on the trunk until the recursion terminates. The switch of the question-guided visual feature $\boldsymbol{v}_{t_p}$ on the branch is closed if the history h_{t_p} is paired with the current traced question q_t. When the recursion termination condition is met, we unroll the process from past to present and finally obtain the recursive visual feature.

We further design three modules to achieve the recursive visual attention algorithm, *i.e.*, Infer , Pair, and Att (*i.e.*, attend). In overview, Infer module asserts the recursion termination condition and computes visual feature fusion weight, Pair module returns the paired round, and Att module calculates question-guided visual attention.

3.2. Neural Modules

Algorithm 2 Infer Module

```
 1: function Infer($\mathcal{Q}, t$)
 2:     $\boldsymbol{z}_t^I \leftarrow f_q^I(\boldsymbol{q}_t)$
 3:     $\boldsymbol{o}_t^I \leftarrow \mathrm{GS\_Sampler}(W^I \boldsymbol{z}_t^I)$
 4:     $\boldsymbol{\alpha}_t^I \leftarrow \mathrm{softmax}(W^I \boldsymbol{z}_t^I)$
 5:     $cond_1 \leftarrow t \overset{?}{=} 0$
 6:     $cond_2 \leftarrow o_{t,0}^I \overset{?}{=} 1$
 7:     $cond \leftarrow cond_1$ or $cond_2$          ▷ recursion termination
 8:     $\lambda \leftarrow \alpha_{t,0}^I$                  ▷ attention fusion weight
 9:     return $cond, \lambda$
10: end function
```

Infer Module. Infer module is designed to 1) determine whether to review the dialog history, 2) provide a weight to fuse the recursive visual attention and the question-guided visual attention. Specifically, Infer module takes the question feature $\boldsymbol{q}_t$ as input. The outputs include 1) a Boolean *cond* to decide whether to terminate the recursion, and 2) a weight $\lambda \in (0, 1)$ for visual attention fusion.

The recursion will be terminated if at least one of the following conditions is satisfied (see lines 5-7 in Algorithm 2). First, the review backtracks to the very starting point: caption. Second, the question q_t is predicted to be unambiguous. In order to estimate the ambiguity of the question, we use a non-linear transformation [34] $f_q^I(\cdot)$, followed by a Gumbel Sampling operation GS_Sampler for differentiable discrete decision:

$$\boldsymbol{z}_t^I = f_q^I(\boldsymbol{q}_t); \tag{1}$$

$$\boldsymbol{o}_t^I = \mathrm{GS\_Sampler}(W^I \boldsymbol{z}_t^I) \tag{2}$$

where W^I denotes the learnable parameters. GS_Sampler (see Section 3.3.2) outputs a 2-dim one-hot vector $\boldsymbol{o}_t^I$ for discrete decision, where the binary element $o_{t,0}^I$ is encoded as the Boolean output to determine whether q_t is ambiguous. As illustrated in Figure 3, our dialog agent successfully learns the relation between words and recursion termination without additional annotations.

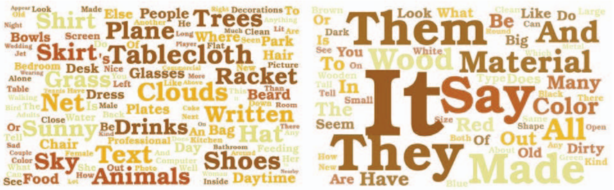

Figure 3. Word cloud visualization of word attention in RvA. For questions that our dialog agent thinks to be unambiguous (left), the word attentions are spread out a variety of nouns (*e.g*., "clouds", "drinks"). For questions that confuse the agent (right), the word attention significantly focuses on pronouns (*e.g*., "it", "they").

Algorithm 3 PAIR Module

```
function PAIR($\mathcal{Q}, \mathcal{H}, t$)
    $\boldsymbol{e}_t^q \leftarrow f_q^P(\boldsymbol{q}_t)$
    for $i \leftarrow 0, \cdots, t-1$ do
        $\boldsymbol{e}_i^h \leftarrow f_h^P(\boldsymbol{h}_i)$
        $z_{t,i}^P \leftarrow \mathrm{MLP}([\boldsymbol{e}_t^q, \boldsymbol{e}_i^h])$
        $\Delta_{t,i} \leftarrow t-i$
    end for
    $\boldsymbol{o}_t^P \leftarrow \mathrm{GS\_Sampler}(W^P[\boldsymbol{z}_t^P, \boldsymbol{\Delta}_t])$
    $t_p \leftarrow \sum_i o_{t,i}^P \cdot i$
    return $t_p$
end function
```

PAIR Module. We observe that an ambiguous question often follows the latest topic. A simple idea is to directly pair the question with its latest history, *i.e*., set t_p as $t-1$ in INFER module. However, the questioner sometimes traces back to an earlier topic, which means that the question has no relationship with its latest history. Therefore, we design a PAIR module to estimate which history is most related with the question q_t.

Algorithm 3 shows the structure of PAIR module. Specifically, PAIR module takes the question feature $\boldsymbol{q}_t$ and the history feature $\mathcal{H} = \{\boldsymbol{h}_0, \cdots, \boldsymbol{h}_{t-1}\}$ as input, and predicts which history is most related to q_t. The PAIR module is formulated as:

$$z_{t,i}^P = \mathrm{MLP}([f_q^P(\boldsymbol{q}_t), f_h^P(\boldsymbol{h}_i)]) \tag{3}$$

$$\boldsymbol{o}_t^P = \mathrm{GS\_Sampler}(W^P[\boldsymbol{z}_t^P, \boldsymbol{\Delta}_t]) \tag{4}$$

$$t_p = \sum_{i=0}^{t-1} o_{t,i}^P \cdot i \tag{5}$$

where $[\cdot]$ is the concatenation operation. The PAIR module considers 1) the matching score between the question q_t and the history h_i, which is denoted as $z_{t,i}^P$; 2) the "sequential distance" between q_t and h_i in the dialog, which is measured by $\Delta_{t,i} = t-i$. Finally, GS_Sampler outputs a t-dim one-hot vector $\boldsymbol{o}_t^P$ for discrete decision (*i.e*., pairing the question with a single history). The question q_t will be paired with the k-th history h_k if $o_{t,k}^P = 1$, *i.e*., the k-th history h_k matches the question q_t better than others.

Algorithm 4 ATT Module

```
function ATT($\mathcal{V}, \mathcal{Q}, t$)
    $\boldsymbol{e}_t^q \leftarrow f_q^A(\boldsymbol{q}_t)$
    for $i \leftarrow 1, \cdots, K$ do
        $\boldsymbol{e}_i^v \leftarrow f_v^A(\boldsymbol{v}_i)$
        $\boldsymbol{z}_{t,i}^A \leftarrow \mathrm{L2Norm}(\boldsymbol{e}_t^q \circ \boldsymbol{e}_i^v)$
    end for
    $\boldsymbol{\alpha}_t^A \leftarrow \mathrm{softmax}(W^A \boldsymbol{Z}_t^A)$
    return $\boldsymbol{\alpha}_t^A$
end function
```

ATT Module. ATT module takes visual features of regions $V = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K\}$ and the question feature $\boldsymbol{q}_t$ as input, and outputs question-guided visual attention. As shown in Algorithm 4, the question-guided visual attention is formulated as:

$$\boldsymbol{z}_{t,i}^A = \mathrm{L2Norm}(f_q^A(\boldsymbol{q}_t) \circ f_v^A(\boldsymbol{v}_i)) \tag{6}$$

$$\boldsymbol{\alpha}_t^A = \mathrm{softmax}(W^A \boldsymbol{Z}_t^A) \tag{7}$$

where $f_q^A(\cdot)$ and $f_v^A(\cdot)$ represents non-linear transformations to embed visual and language features into the same space, and $\circ$ denotes Hadamard (element-wise) product for multi-modal feature fusion.

3.3. Training

As mentioned in Section 3.2, our Recursive Visual Attention takes visual and language representations as input, and applies Gumbel sampling for differentiable discrete decision. The details are given as follows.

3.3.1 Feature Representation

Language Feature. Let $W_t^q = \{\boldsymbol{w}_{t,1}^q, \cdots, \boldsymbol{w}_{t,m}^q\}$ be the word embeddings of the question q_t. The word embeddings are passed through the bidirectional LSTM (bi-LSTM):

$$\overrightarrow{\boldsymbol{h}}_{t,i}^q = \mathrm{LSTM}_f^q(\boldsymbol{w}_{t,i}^q, \overrightarrow{\boldsymbol{h}}_{t,i-1}^q) \tag{8}$$

$$\overleftarrow{\boldsymbol{h}}_{t,i}^q = \mathrm{LSTM}_b^q(\boldsymbol{w}_{t,i}^q, \overleftarrow{\boldsymbol{h}}_{t,i+1}^q) \tag{9}$$

$$\boldsymbol{h}_{t,i}^q = [\overrightarrow{\boldsymbol{h}}_{t,i}^q, \overleftarrow{\boldsymbol{h}}_{t,i}^q] \tag{10}$$

where $\overrightarrow{\boldsymbol{h}}_{t,i}^q$ and $\overleftarrow{\boldsymbol{h}}_{t,i}^q$ represent forward and backward hidden state of the i-th word respectively, LSTM_f^q and LSTM_b^q represent the forward and backward LSTMs. We use the concatenation of last hidden states $\boldsymbol{e}_t^q = [\overrightarrow{\boldsymbol{h}}_{t,m}^q, \overleftarrow{\boldsymbol{h}}_{t,1}^q]$ as the encoding of the whole question q_t. Similarly, we can encode the history h_i as $\boldsymbol{e}_i^h$ using the same bi-LSTM with different parameters. In PAIR module, we denote $\boldsymbol{e}_t^q$ as $\boldsymbol{q}_t$ and $\boldsymbol{e}_i^h$ as $\boldsymbol{h}_i$ to calculate the matching score between the question q_t and the history h_i.

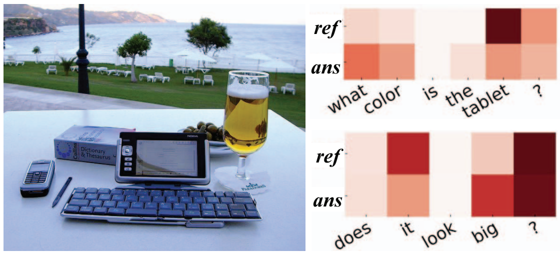

Figure 4. A qualitative example of question attentions. The reference-aware (*ref*) question attention mainly emphasizes nouns (*i.e.*, *"tablet"*) and pronouns (*i.e.*, *"it"*) for recursion termination estimation and visual grounding. The answering-aware (*ans*) question attention highlights property words (*i.e.*, *"what color"*, *"big"*) to record question type and activate specific attributes of visual representation for question answering. Darker color indicates higher weight.

Note that the words contribute differently to the question representation for various purposes. An example is illustrated in Figure 4. On one hand, the words "tablet" and "it" should be emphasized for recursion termination estimation and visual grounding. On the other hand, the phrase "what color" and the word "big" should be highlighted to activate specific attributes of the visual representation for question answering. Therefore, we encode each question using self-attention mechanisms [35] into two forms: *reference-aware* question feature $\boldsymbol{q}^{ref}_t$ and *answering-aware* question feature $\boldsymbol{q}^{ans}_t$. Different from prior attention mechanism that uses linear transformation followed by hyperbolic tangent $(\tanh)$ activation, we formulate the self-attention mechanism as:

$$\boldsymbol{z}^{q,*}_{t,i} = \mathrm{L2Norm}(f^{q,*}_q(\boldsymbol{h}^q_{t,i})) \tag{11}$$

$$\boldsymbol{\alpha}^{q,*}_t = \mathrm{softmax}(W^{q,*}\boldsymbol{Z}^{q,*}_t) \tag{12}$$

$$\boldsymbol{q}^*_t = \sum_{i=1}^{m} \alpha^{q,*}_{t,i}\boldsymbol{w}^q_i \tag{13}$$

where $f^{q,*}_q(\cdot)$ is a non-linear transformation function, $W^{q,*}$ is the learnable parameters, and $* \in \{ref, ans\}$. The attended question features $\boldsymbol{q}^{ref}_t$ and $\boldsymbol{q}^{ans}_t$ are calculated by a weighted sum overall the words. In INFER and ATT modules, we denote $\boldsymbol{q}^{ref}_t$ as $\boldsymbol{q}_t$ for recursion termination estimation and visual grounding.

Visual Feature. Spatial image features with attention mechanism have been widely used in many vision and language tasks, such as image captioning and visual question answering. Recently, a bottom-up attention mechanism [3] is proposed based on the Faster R-CNN framework. The ResNet model is utilized as backbone and trained on Visual Genome [19] dataset to predict attributes and classes. In this paper, we apply the bottom-up attention mechanism and select top-K region proposals from each image, where K is simply fixed as 36.

After obtaining the visual feature $\hat{\boldsymbol{v}}_T$ using Recursive Visual Attention, we further refine the visual feature using the answering-aware question feature $\boldsymbol{q}^{ans}_T$. The motivation is that only question-related attributes of visual content are useful for answering questions (*e.g.*, "What color is the tablet?", "Does it look big?" in Figure 4). Motivated by the gating operations within LSTMs and GRUs, we further refine visual feature as:

$$\tilde{\boldsymbol{v}}_T = \hat{\boldsymbol{v}}_T \circ f^v_q(\boldsymbol{q}^{ans}_T) \tag{14}$$

where the output of non-linear transformation $f^v_q(\cdot)$ works as a "visual feature filter" to deactivate the information unrelated to answering questions in the visual representation $\hat{\boldsymbol{v}}_t$.

Joint Embedding. Considering that the dialog history reflects prior knowledge of visual content, we obtain the "fact" embedding by attending to all the history as

$$\boldsymbol{z}^h_{T,i} = \mathrm{L2Norm}(f^h_q(\boldsymbol{e}^q_T) \circ f^h_h(\boldsymbol{e}^h_i)) \tag{15}$$

$$\boldsymbol{\alpha}^h_T = \mathrm{softmax}(W^h\boldsymbol{Z}^h_T) \tag{16}$$

$$\boldsymbol{h}^f_T = \sum_{i=0}^{T-1} \alpha^h_{T,i}\boldsymbol{e}^h_i \tag{17}$$

where f^h_h and f^h_q are non-linear transformation functions. The "fact" embedding $\boldsymbol{h}^f_T$ is calculated by a weighted sum over all the history encodings.

Since we have obtained the filtered visual feature $\tilde{\boldsymbol{v}}_T$, the answering-aware question feature $\boldsymbol{q}^{ans}_T$, and the fact embedding $\boldsymbol{h}^f_T$ for question q_T, we concatenate these features and use a linear transform followed by a tangent activation to obtain the final joint embedding:

$$\boldsymbol{e}^J_T = \tanh(W^J[\tilde{\boldsymbol{v}}_T, \boldsymbol{q}^{ans}_T, \boldsymbol{h}^f_T]) \tag{18}$$

where $[\cdot]$ denotes the concatenation operation. The joint embedding is further fed into the answering decoder.

3.3.2 Gumbel Sampling

Our dialog agent needs to make a discrete decision in some cases, *e.g.*, estimating whether to review the history and which history should be paired. In addition, we hope that the gradients can be back propagated through discrete decision making for end-to-end training. In order to achieve these goals, we utilize the Gumbel-Max trick [12] with its continuous softmax relaxation [15, 23]. Specifically, the samples $\boldsymbol{z}$ can be drawn from a categorical distribution with $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_c\}$ as:

$$\boldsymbol{z} = \mathrm{one\_hot}\left(\underset{k\in\{1,\ldots,c\}}{\arg\max}\left(\log(\pi_k) + g_k\right)\right) \tag{19}$$

where $\boldsymbol{g} = -\log(-\log(\boldsymbol{u}))$ with $\boldsymbol{u} \sim \mathrm{unif}[0, 1]$.

The softmax relaxation of Gumbel-Max trick is to replace non-differentiable $\arg\max$ operation with the continuous $\mathrm{softmax}$ function:

$$\hat{\boldsymbol{z}} = \mathrm{softmax}\left((\log(\boldsymbol{\pi}) + \boldsymbol{g})/\tau\right) \tag{20}$$

where the temperature of the softmax function τ is empirically set as 1 in our work. During the training stage, we obtain an one-hot vector $\boldsymbol{z}$ as the discrete sample from Eq. 19 for forward propagation, and compute gradients w.r.t. $\boldsymbol{\pi}$ in Eq. 20 for back propagation. At the test stage, we greedily draw the sample with the largest probability without Gumbel samples $\boldsymbol{g}$.

4. Experiments

Our proposed model is evaluated on two real-world datasets: VisDial v0.9 and v1.0 [9]. In this section, we first introduce the datasets, evaluation metrics, and implementation details. We then compare our method with the state-of-the-art models and provide qualitative results.

4.1. Datasets and Setup

The VisDial v0.9 [9] dataset was collected based on MS-COCO [20] images and captions. In a two-player chat game, one player attempts to learn about an unseen image and asks questions based on the previous dialog, while the other player watches the image and replies with free-form answers. The whole chat lasts for 10 rounds for each image. As a result, the VisDial v0.9 dataset contains 83k dialogs on MS-COCO training images and 40k dialogs on validation images. Recently, the VisDial v1.0 [9] dataset was released, including additional 10k dialogs on Flickr images. The collection of dialogs on Flickr images is similar to that on MS-COCO images. Overall, the new train split consists of 123k dialogs on MS-COCO images, which is the combination of train and validation splits from VisDial v0.9. The validation and test splits have 2k and 8k dialogs on Flickr images, respectively. Different from val split in VisDial v0.9 where each image is associated with a 10-round dialog, the dialogs in VisDial v1.0 test split have a random length within 10 rounds.

4.2. Metrics

As in [9], we evaluated the responses at each round in VisDial v0.9 and the last round in VisDial v1.0 in a retrieval setting. Specifically, at test stage, each question is linked with a list of 100 candidate answers. The model is expected to rank over the candidates and return a ranked list for further evaluation. The metrics for retrieval performance evaluation are: 1) mean rank of human response (**Mean**); 2) recall@k (**R@k**), which is the existence of the human response in the top-k responses; 3) mean reciprocal rank (**MRR**) of the human response in the returned ranked list. As for VisDial v1.0, we also used the newly introduced normalized discounted cumulative gain (**NDCG**), which penalizes the lower rank of answers with high relevance.

4.3. Implementation Details

Language Model. We pre-processed the text data as follows. As in [9], we first lowercased all questions and answers, converted digits to words, and removed contractions, before tokenizing using the Python NLTK toolkit [1]. The captions, questions, and answers were then padded or truncated to 40, 20 and 20, respectively. We kept words to those that occur at least 5 times in the training split, resulting in a vocabulary of 9,795 words for VisDial v0.9 and 11,336 words for VisDial v1.0. Our word embeddings are 300-dim vectors, initialized with pre-trained GloVe [26] embeddings and shared across captions, questions and answers. The dimension of hidden states in all LSTMs is set to 512 in this work.

Training Details. We minimized the standard cross-entropy loss for the discriminative training, and the maximum likelihood estimation (MLE) loss for generative training. We used Adam [16] with the learning rate of 1×10^{-3}, multiplied by 0.5 after every epoch, decreasing to 5×10^{-5}. We also applied Dropout [31] with a ratio of 0.5 before each fully-connected layer. Other settings are default in PyTorch [2].

4.4. Comparing Methods

We compared our proposed Recursive Visual Attention (**RvA**) model with the state-of-the-art methods in both discriminative and generative settings. Based on the design of encoders, these methods can be grouped into:

Fusion-based Models. Early methods simply fuse image, question, and history features at different stages. These early methods include **LF** [9] and **HRE** [9].

Attention-based Models. Furthermore, some methods establish attention mechanisms over image, question, and history. The attention-based methods include **HREA** [9], **MN** [9], **HCIAE** [22], and **CoAtt** [37].

VCoR-based models. Recent works have focused on explicit visual co-reference resolution (VCoR) in visual dialog. We compared our method with VCoR-based models including **AMEM** [29] and **CorefNMN** [18].

Ablative Models. In addition, we evaluate the individual contribution of following features and components in our method: 1) **RPN**: we replaced the region proposal network with VGG-16 [30] model, and used the spatial grids of *pool5* feature map as regions. 2) **Bi-LSTM**: we replaced bidirectional LSTM with the vanilla LSTM. 3) **Rv**: we only considered the termination condition of RvA, and replaced the recursive attention with question-guided attention. 4) **FL**: we withdrew the "visual feature filter" $f_q^v(\cdot)$ in Eq. 14, which controls the activation of visual attributes.

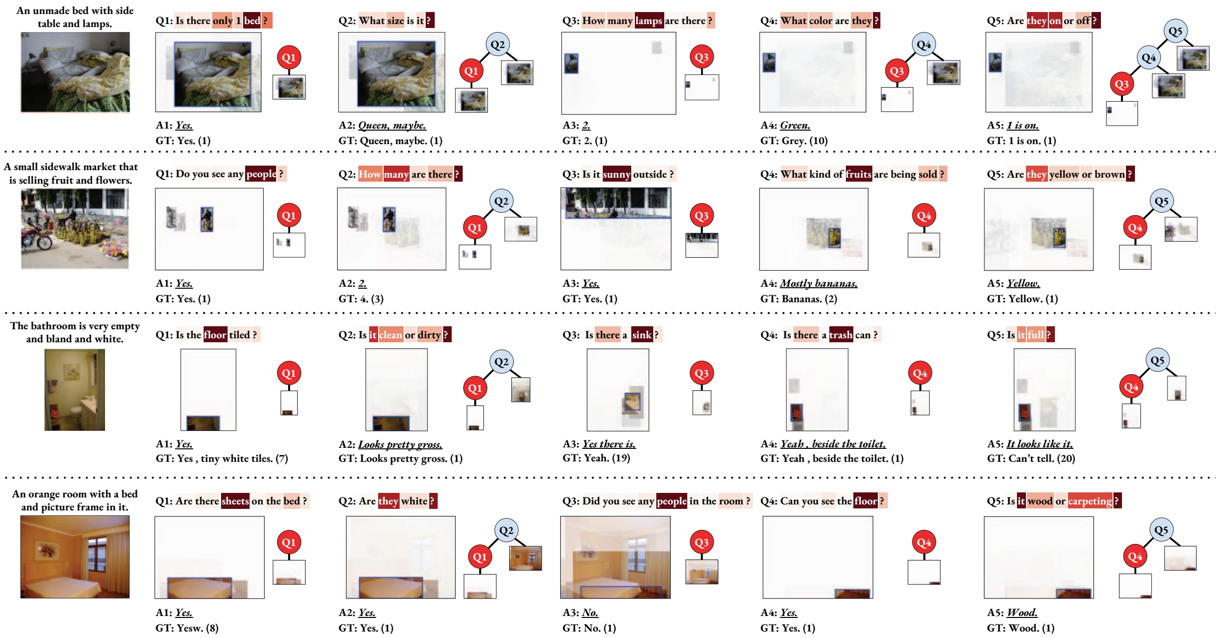
An unmade bed with side table and lamps.
Q1: Is there only 1 bed ?
Q1
A1: Yes.
GT: Yes. (1)
Q2: What size is it ?
Q2
Q1
A2: Queen, maybe.
GT: Queen, maybe. (1)
Q3: How many lamps are there ?
Q3
A3: 2.
GT: 2. (1)
Q4: What color are they ?
Q4
Q3
A4: Green.
GT: Grey. (10)
Q5: Are they on or off ?
Q5
Q4
Q3
A5: 1 is on.
GT: 1 is on. (1)
A small sidewalk market that is selling fruit and flowers.
Q1: Do you see any people ?
Q1
A1: Yes.
GT: Yes. (1)
Q2: How many are there ?
Q2
Q1
A2: 2.
GT: 4. (3)
Q3: Is it sunny outside ?
Q3
A3: Yes.
GT: Yes. (1)
Q4: What kind of fruits are being sold ?
Q4
A4: Mostly bananas.
GT: Bananas. (2)
Q5: Are they yellow or brown ?
Q5
Q4
A5: Yellow.
GT: Yellow. (1)
The bathroom is very empty and bland and white.
Q1: Is the floor tiled ?
Q1
A1: Yes.
GT: Yes , tiny white tiles. (7)
Q2: Is it clean or dirty ?
Q2
Q1
A2: Looks pretty gross.
GT: Looks pretty gross. (1)
Q3: Is there a sink ?
Q3
A3: Yes there is.
GT: Yeah. (19)
Q4: Is there a trash can ?
Q4
A4: Yeah , beside the toilet.
GT: Yeah , beside the toilet. (1)
Q5: Is it full ?
Q5
Q4
A5: It looks like it.
GT: Can't tell. (20)
An orange room with a bed and picture frame in it.
Q1: Are there sheets on the bed ?
Q1
A1: Yes.
GT: Yesw. (8)
Q2: Are they white ?
Q2
Q1
A2: Yes.
GT: Yes. (1)
Q3: Did you see any people in the room ?
Q3
A3: No.
GT: No. (1)
Q4: Can you see the floor ?
Q4
A4: Yes.
GT: Yes. (1)
Q5: Is it wood or carpeting ?
Q5
Q4
A5: Wood.
GT: Wood. (1)

Figure 5. Qualitative results of our RvA model on VisDial dataset. The number in the bracket is the rank of ground-truth (GT) answer in the returned sorted list. Our model successfully obtains interpretable reference-aware question attention (represented by the highlight color of words, darker color indicates higher weight), reliable recursive image attention (represented by the attention map below the question), and reasonable recursions (represented by the recursion tree). The root nodes in the recursion tree represent the questions to be answered, the red nodes denote the questions terminating recursions, and the leaf nodes represent question-guided visual attention.

Model	MRR	R@1	R@5	R@10	Mean
LF [9]	0.5807	43.82	74.68	84.07	5.78
HRE [9]	0.5846	44.67	74.50	84.22	5.72
HREA [9]	0.5868	44.82	74.81	84.36	5.66
MN [9]	0.5965	45.55	76.22	85.37	5.46
HCIAE [22]	0.6222	48.48	78.75	87.59	4.81
AMEM [29]	0.6227	48.53	78.66	87.43	4.86
CoAtt [37]	0.6398	50.29	80.71	88.81	4.47
CorefNMN [18]	0.641	50.92	80.18	88.81	4.45
RvA w/o RPN	0.6436	50.40	81.36	89.59	4.22
RvA w/o Rv	0.6551	51.81	82.35	90.24	4.07
RvA w/o FL	0.6598	52.35	82.76	90.54	3.98
RvA	**0.6634**	**52.71**	**82.97**	**90.73**	**3.93**

Table 1. Retrieval performance of discriminative models on the validation set of VisDial v0.9. RPN, Rv and FL indicate the usage of region proposal network, recursive image attention, and visual feature filter, respectively.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF [9]	0.5542	40.95	72.45	82.83	5.95	0.4531
HRE [9]	0.5416	39.93	70.45	81.50	6.41	0.4546
MN [9]	0.5549	40.98	72.30	83.30	5.92	0.4750
CorefNMN† [18]	0.615	47.55	78.10	88.80	4.40	0.547
RvA w/o RPN	0.6060	46.25	77.88	87.83	4.65	0.5176
RvA w/o Rv	0.6226	47.95	79.75	89.08	4.37	0.5319
RvA w/o FL	0.6294	48.68	80.18	89.03	4.31	0.5418
RvA	**0.6303**	**49.03**	**80.40**	**89.83**	**4.18**	**0.5559**

Table 2. Retrieval performance of discriminative models on the test-standard split of VisDial v1.0. † indicates that the model uses ResNet-152 features.

4.5. Quantitative Results

Table 1 reports the retrieval performances of our model and the comparing methods under the discriminative setting on VisDial v0.9. Overall, our RvA model outperforms the state-of-the-art methods across all the metrics. Specifically, our RvA model achieves approximately 2 points improvement on R@k, and 2% increase on MRR. In addition, the performance of our model drops significantly without recursive attention (*i.e.*, Rv) or region proposal network (*i.e.*, RPN), which demonstrates their substantial contributions to visual dialog. The similar conclusions can also be drawn on VisDial v1.0 in Table 2. Furthermore, Table 3 shows a more comprehensive ablation (the component-wise and the feature-wise). It can be seen that by using the proposed recursive attention, any ablative method can be improved regardless of the usage of visual and language representations. Furthermore, our dialog agent could occupy the third place based on the VisDial v1.0 leaderboard[1], while the team DL-61 [13] has achieved the best NDCG record 0.5788 in a two-stage fashion.

We also evaluated the retrieval performance of our model under generative setting on VisDial v0.9. As shown in Table 4, our approach obtains an approximately 2 points higher R@k compared to the visual co-reference solution model CorefNMN [18]. In addition, our RvA model outperforms nearly all state-of-the-art methods except CoAtt [37], which is trained using reinforcement learning.

[1] https://evalai.cloudcv.org/web/challenges/challenge-page/103/leaderboard/298

RPN	Bi-LSTM	Rv	MRR	R@1	R@5	R@10	Mean
			0.6377	49.67	80.86	89.14	4.35
		✓	**0.6418**	**50.17**	**81.17**	**89.37**	**4.29**
	✓		0.6396	49.83	81.16	89.34	4.30
	✓	✓	**0.6436**	**50.40**	**81.36**	**89.59**	**4.22**
✓			0.6534	51.78	82.28	90.21	4.09
✓		✓	**0.6626**	**52.69**	**82.97**	**90.71**	**3.95**
✓	✓		0.6551	51.81	82.35	90.24	4.07
✓	✓	✓	**0.6634**	**52.71**	**82.97**	**90.73**	**3.93**

Table 3. Ablations of discriminative models on the validation set of VisDial v0.9. RPN, Bi-LSTM and Rv indicate the usage of region proposal network, bidirectional LSTM, and recursive image attention, respectively.

Model	MRR	R@1	R@5	R@10	Mean
LF [9]	0.5199	41.83	61.78	67.59	17.07
HRE [9]	0.5237	42.29	62.18	67.92	17.07
HREA [9]	0.5242	42.28	62.33	68.17	16.79
MN [9]	0.5259	42.29	62.85	68.88	17.06
CorefNMN [18]	0.535	43.66	63.54	69.93	15.69
HCIAE [22]	0.5386	44.06	63.55	69.24	16.01
CoAtt [37]	0.5411	44.32	63.82	69.75	16.47
CoAtt‡ [37]	**0.5578**	**46.10**	**65.69**	71.74	14.43
RvA w/o RPN	0.5417	43.75	64.21	71.85	11.18
RvA	0.5543	45.37	65.27	**72.97**	**10.71**

Table 4. Retrieval performance of generative models on the validation set of VisDial v0.9. ‡ indicates that the model is trained using reinforcement learning.

4.6. Qualitative Results

The qualitative results shown in Figure 5 and 6 demonstrate the following advantages of our RvA model:

Reasonable Recursions. Our RvA model achieves reasonable recursions represented by the recursive trees. These recursions can also be regarded as topic-aware dialog clips. Thanks to the reference-aware language feature, our RvA model is able to handle unambiguous sentences with pronouns (*e.g.*, "Is it sunny outside?") and ambiguous sentences without pronouns (*e.g.*, "How many are there?"). Note that it is hard to exhaust all these special cases using a natural language parser.

Reliable Visual Attention. Our dialog agent successfully focuses on the correct region using recursive visual attention. In contrast, the question-guided visual attention sometimes fails due to the ambiguous question. On the validation set of VisDial v1.0, we observed that: 1) 56% of question-guided visual attention and 89% of recursive attention are reasonable for ambiguous questions; 2) 62% of dialogs require at least one accurate co-reference resolution. Since the recursive visual attention relies heavily on historical visual attention, our dialog agent needs to establish a robust visual attention mechanism. If it were otherwise, the agent would distrust historical visual attention and tend to learn more bias from generic language information, which would hurt the visual dialog system.

History-aware Skipping Pairing. One may argue that PAIR module can be replaced with referring all the ambiguous questions to their last history (*i.e.*, setting t_p as $t-1$ in INFER module) for simplicity. However, our PAIR module is able to *skip* the irrelevant dialog history and produce *history-aware* recursions. As illustrated in Figure 6 (a), the dialog agent concludes from the dialog history that "there is no cord" in the image. Therefore, the agent skips the second history when pairing the ambiguous question "Is it black?". If we replace the second answer "no" with "yes" to make a fake history (see Figure 6 (b)), the third question will be directly paired with its last history. The visual attention and predicted answer are also influenced.

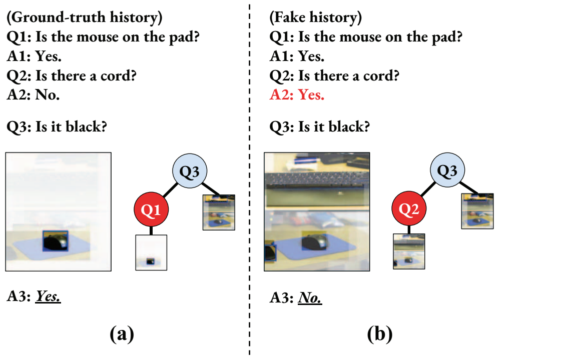

Figure 6. An qualitative example of the history-aware recursion using PAIR module to resolve "it" in Q3. (a) represents the recursion obtained by our model using the ground-truth history. Our dialog agent skips the unrelated second history, and pairs the ambiguous question Q3 with the first history. (b) represents the recursion obtained by our model with the fake history, where the second answer "No" is replaced with "Yes". In this case, our dialog agent pairs the third question with its last history.

5. Conclusions

In this paper, we formulated the visual co-reference resolution in visual dialog as Recursive Visual Attention (RvA), which consists of three simple neural modules that determine the recursion at run-time. Our dialog agent recursively reviews topic-related history to refine visual attention, and can be end-to-end trained when making discrete decisions of module assembling. Experimental results on the large-scale real-world datasets VisDial v0.9 and v1.0 demonstrate that our proposed model not only achieves state-of-the-art performance, but also obtains explainable recursion and attention maps. Moving forward, we are going to incorporate in-depth language parsing modules into RvA for more accurate recursive decisions.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (61573363 and 61832017), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (15XNLQ01), and NTU-Alibaba JRI.

References

[1] NLTK. http://www.nltk.org/. 6

[2] PyTorch. https://pytorch.org/. 6

[3] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 1, 5

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 1

[5] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016. 2

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 1

[7] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011. 2

[8] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5659–5667, 2017. 1

[9] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017. 1, 2, 6, 7, 8

[10] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017. 1, 2

[11] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1

[12] Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954. 2, 5

[13] Dalu Guo, Chang Xu, and Dacheng Tao. Image-question-answer synergistic network for visual dialog. *arXiv preprint arXiv:1902.09774*, 2019. 7

[14] De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5948–5957, 2018. 2

[15] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 2, 5

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3558–3565, 2014. 2

[18] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of European Conference on Computer Vision*, pages 153–169, 2018. 2, 6, 7, 8

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1, 5

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision*, pages 740–755. Springer, 2014. 6

[21] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of European Conference on Computer Vision*, pages 852–869. Springer, 2016. 1

[22] 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 *Advances in Neural Information Processing Systems*, pages 314–324, 2017. 2, 6, 7, 8

[23] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 2, 5

[24] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016. 1

[25] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Proceedings of European Conference on Computer Vision*, pages 792–807. Springer, 2016. 1

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. 6

[27] Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *Proceedings of European Conference on Computer Vision*, pages 95–110. Springer, 2014. 2

[28] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2017. 2

[29] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. In *Advances in Neural Information Processing Systems*, pages 3719–3729, 2017. 2, 6, 7

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[31] Nitish Srivastava, Geoffrey 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. 6

[32] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. *arXiv preprint arXiv:1812.01880*, 2018. 1

[33] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4631–4640, 2016. 1

[34] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2018. 3

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 5

[36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 1

[37] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 6106–6115, 2018. 6, 7, 8

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of International Conference on Machine Learning*, pages 2048–2057, 2015. 1

[39] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: weakly supervised visual relation detection via parallel pairwise r-fcn. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4233–4241, 2017. 1

[40] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018. 1