

# GuessWhich? Visual Dialog with Attentive Memory Network

Lei Zhao<sup>a</sup>, Xinyu Lyu<sup>b</sup>, Jingkuan Song<sup>a</sup>, Lianli Gao<sup>a,\*</sup>

<sup>a</sup>*Center of Future Media, School of Computer Science and Engineering,  
University of Electronic Science and Technology of China*

<sup>b</sup>*Rutgers, the State University of New Jersey*

---

## Abstract

Visual dialog is a task that two agents (Q-BOT and A-BOT) communicate in natural language on the situation of information asymmetry. Q-BOT generates questions based on the image caption and historical dialog, while A-Bot answers the questions grounded on images. And we play a cooperative ‘image guessing’ game between Q-BOT and A-Bot, so that Q-BOT can select an unseen image from a set of images. However, due to the insufficient use of image caption and historical dialog, existing methods usually generate irrelevant or homogenous questions, which provide less value to this visual dialog system. To tackle these issues, we propose an Attentive Memory Network (AMN) to fully exploit the image caption and historical dialog information. Specifically, the attentive memory network mainly consists of a memory network and an additional fusion model. The memory network can hold long historical dialog information and give each dialog round a different weight. Aside with the historical dialog information, the fusion model in Q-BOT and A-BOT further uses the image caption and the image information respectively. The caption information can assist the attentive generation of the questions, and the image helps A-BOT produce precise answers. Under their influence, the generated questions are more diverse and concentrated, and the corresponding answers are more accurate. We pre-train our proposed model on VisDial v1.0 dataset

---

\*Corresponding author

Email address: [lianli.gao@uestc.edu.cn](mailto:lianli.gao@uestc.edu.cn) (Lianli Gao)

and then fine-tune it by reinforcement learning. The experimental results on guessing accuracy and dialog quality both greatly exceed the benchmark model.

*Keywords:* Visual dialog, Attentive memory network, Reinforcement learning

---

## 1. Introduction

In recent years, research on the combination of vision and language has gained lots of attention, such as image captioning [1, 2], video captioning [3, 4], and visual question answering [5, 6, 7, 8]. But the development of visual dialog  
5 has only just begun. Visual dialog is a task that two agents (Question-BOT and Answer-BOT, hereinafter referred to as Q-BOT and A-BOT) communicate in natural language on the condition of information asymmetry. This paper focuses on ‘GuessWhich’ based visual dialog [13] which is a combined task of visual dialog and image guessing (image retrieval). These two sub-tasks alternate  
10 iteratively. In the  $t$ -round dialog, Q-BOT generates a question according to the caption and the previous  $t - 1$  historical dialogs. Then A-BOT answers the generated question according to the images and the previous  $t - 1$  historical dialogs. After the interaction between Q-BOT and A-BOT, Q-BOT finally guesses the corresponding image which is seen by A-BOT from the candidate  
15 images according to the dialog content. This task is of great significance in many practical application scenarios. It can assist the visually impaired patients with perceiving the surrounding environment, and aid the intelligent assistant in providing users with more informative visual information. The retrieval function can also be applied to many applications, e.g., online shopping assistant in which  
20 the users can communicate with the recommended agent to get their favorite products.

Previous works [9, 10] on visual dialog just made the agents understand the input image and generate general dialog. It is a simple process from vision to language generation. The quality of generated dialogs is not satisfying enough,  
25 and the applications of this task are limited. Then many methods [11, 12]

about ‘GuessWhat’ based visual dialog were proposed. This task needs the time-consuming annotation of the objects in the image. So ‘GuessWhich’ based cooperative visual dialog that can effectively boost the communication rationality becomes a more and more valuable research topic. There are still many challenges. The most serious issue is that the repeated and invalid interaction appears multiple times. The homogenous dialogs are produced by the simple use of the historical information. The irrelevant questions are caused by the insufficient use of the caption which is the only visual information for Q-BOT.

Inspired by that, we apply the memory network to store historical dialogs to generate the diverse questions, and use an additional fusion model which fuse the caption information and image to improve the accuracy of the generated dialogs. We finally proposed an attentive memory network composed by the memory network and the fusion model. Concretely, each historical question-answer pair in Q-BOT and A-BOT is embedded to make up a memory. In every dialog round, the memory is queried to produce the weighted history. Then the fusion model of the attentive memory network in Q-BOT uses the caption information, the embedded fact, and the weighted history to generate the next question and the guessing images. Meanwhile, the fusion model of the attentive memory network in A-BOT takes advantage of the encoded question, the image, and the weighted history to produce the answer. Under the influence of the attentive memory network, Q-BOT can provide diverse and effective questions. Meanwhile, A-BOT can produce accurate answers. The most intuitive performance is that the generated dialogs significantly reduce the repetitions, and the discrete action space becomes larger. Thanks to the improvement of dialogue quality and the effective use of the caption information, Q-BOT can complete image retrieval precisely. The contributions of this paper are summarized as follows:

- We use memory network in the cooperative ‘GuessWhich’ game between Q-BOT and A-BOT. For the current dialog round, memory network can learn the different weights of the historical question-answer pairs at dif-

ferent times. The weighted history information can reduce the repetition of the generated dialogs and make image retrieval efficient. The memory network is simple but effective enough.

- We propose a novel Attentive Memory Network that adds a fusion model to the memory network. The fusion model can effectively use the manually labeled caption and the image. Thus the historical information is focused under the impact of multivariate information fusion, and the generated dialogs and the predicted image representation can be visually grounded.
- Experiments conducted on VisDial 1.0 datasets demonstrate that our generated dialogs are natural and precise, and the results exceed the state-of-the-art ‘GuessWhich’ based visual dialog algorithms. Extensive image retrieval experiments prove that our method can also generate more accurate results compared to the benchmark models.

The rest of this paper is organized as follows. Related work is presented in Section 2. The details of our proposed method are presented in Section 3 mainly including the framework and training strategy. Extensive experiment is conducted in Section 4. At last, a conclusion is drawn in Section 5.

## 2. Related Work

In this section, three categories of related works including visual dialog, goal-oriented visual dialog and memory networks are described.

### 2.1. Visual Dialog

Visual dialog refers to the task that an AI bot holds a meaningful communication with a human or another bot in natural language about an input image. It was firstly proposed by Abhishek Das *et al.* [9]. Besides the definition of visual dialog, they also developed a large-scale Visual Dialog dataset named VisDial which was the foundation of the development of this research field. Encoder-decoder architectures had usually been exploited to achieve visual dialog, and these models were still active up to now. There were many

repeated dialogs which led to lots of invalid questions. Meanwhile, the answers  
85 were often so conservative that the dialog fell into multiple circles. All of that  
made it lack value for practical applications. Since then, a lot of research works  
had been proposed to alleviate the above problems.

## 2.2. Goal-oriented Visual Dialog

To improve the quality of dialog and increase its actual value in real-life  
90 scenario, many methods [18, 19, 11, 12, 20, 21, 13, 14, 15, 16, 17] about goal-  
oriented visual dialog models were proposed. There are mainly two branches  
including ‘GuessWhat’ [18, 19, 11, 12, 20] and ‘GuessWhich’ [21, 13, 14, 15, 16,  
17].

The ‘GuessWhat’ game is to identify an unknown object in a complex im-  
95 age through a meaningful dialog, while the cooperative bots both know the  
image information. This task needs the object-level annotation which is time-  
consuming, and the answers to the questions were limited to Yes, No and N/A.  
The initial method [18] trained the agents by supervised learning, and it would  
result in repeated and invalid dialogs. Then, Harm de Vries *et al.* [18] added  
100 a decision-making component that decided whether to continue the following  
dialogs. Abbasnejad *et al.* [20] proposed a Bayesian model of uncertainty to  
identify valid information and generated more efficient questions. Ravi Shekhar  
*et al.* [12] jointly trained the mixed process between ‘GuessWhat’ and generating  
questions by reinforcement learning which was more natural for visual dialog,  
105 and proposed a visual-grounded dialog history encoder to integrate multi-modal  
information efficiently. Shukla *et al.* [21] proposed an end-to-end ‘GuessWhat’  
based visual dialog model using a combined learning mechanism between rein-  
forcement learning and regularized information gain. This task could improve  
the quality of the dialog, but the required dataset need many human labors.  
110 Therefore, its actual development would be somewhat limited.

The ‘GuessWhich’ game was firstly proposed in [13]. It is a task of learning  
cooperative visual dialog agents through image retrieval. Q-BOT does not know  
the image, it just generates a series of questions by the caption information and

the historical dialogs. A-BOT can get the image information, and it takes the  
 115 charge of answering the questions depending on the image, the question, and  
 the historical information. At the last of each round, Q-BOT selects the target  
 image from the candidate image pooling. This subtask still faces the problem  
 that there are many repeated dialogs with low quality. Jiasen Lu *et al.* [14]  
 made use of a discriminative dialog model to rank the candidate responses.  
 120 Then the generic responses were reduced effectively. Community based visual  
 dialog model was proposed in [15]. Each agent communicated with multiple  
 agents to learn abundant information. Jiaping Zhang *et al.* [16] developed  
 a framework that utilized the multi-modal state representation by hierarchical  
 decision learning. Moreover, they used state adaptation to combine the relevant  
 125 information. Vishvak Murahari *et al.* [17] added an additional objective that  
 stimulated Q-BOT to ask more attentive and various questions.

### 2.3. Memory Networks

Memory networks [22, 23, 24, 25, 26] refer to the networks which have addi-  
 tional memory where information can be read and written. They were originally  
 130 used especially in the NLP field. Recently, memory networks have been pen-  
 etrated the computer vision field and the intersection of vision and language.  
 It was first applied to visual language in [9]. This method devised a memory  
 bank that maintained historical dialogs. Then weights of the previous rounds  
 were computed by a query operation and a softmax function. Finally, the com-  
 135 bination of the historical vectors was used to generate dialog. Compared with  
 LSTM, memory networks can hold longer history dialogs. Akshat Agarwal *et*  
*al.* [15] also applied memory networks to goal-oriented cooperative dialog. But  
 they just used fully connected layers to act as the state encoder, the questions  
 encoder, and the answer encoder. Their interactive modes were different seri-  
 140 ously.

Our task inspired by [13] belongs to the ‘GuessWhich’ based visual dia-  
 log. Different from ordinary memory networks, the proposed attentive memory  
 network can take more advantage of the visual and language information. It



### 155 3.1. Overview

The overview of our approach is illustrated in Fig. 1 which shows just a round of interaction. It mainly consists of two parts: Q-BOT that is responsible for asking coherent questions, and A-BOT that is in charge of answering corresponding questions by the visual content. Then Q-BOT also needs to retrieve  
160 the target image at the end of each dialog round depending on the predicted feature.

### 3.2. Q-BOT and A-BOT Architecture

**Q-BOT** mainly consists of five components, including question decoder, fact encoder, attentive memory network, history encoder, and feature regression  
165 network. The question decoder is a simple LSTM which outputs a question  $\mathbf{q}_t$  through the previous hidden state  $\mathbf{s}_{t-1}^Q$ . Then A-BOT responds to the question  $\mathbf{q}_t$ , and returns a corresponding answer  $\mathbf{a}_t$  to Q-BOT. The fact encoder is also an LSTM which embeds the concatenated  $(\mathbf{q}_t, \mathbf{a}_t)$  into a 512-dimensional vector  $\mathbf{f}_t^Q$  which is used as a query.

**Attentive memory network  $\text{AMN}^Q$**  is composed of memory reading and information fusion. The memory  $\mathbf{M}_t^Q$  refers to the stored fact  $[\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{t-1}]$ . Because all facts are 512-d vectors, the memory  $\mathbf{M}_t^Q$  is a  $(t-1) \times 512$  matrix. The process that the fact  $\mathbf{f}_t$  queries the memory  $\mathbf{M}_t^Q$  is matrix multiplication essentially. Then an attention vector  $\mathbf{A}^Q$  for the memory  $\mathbf{M}_t^Q$  is produced:

$$\mathbf{A}^Q = \text{Softmax} \left( \mathbf{f}_t^Q \left( \mathbf{M}_t^Q \right)^T \right) = [\mathbf{a}_1, \mathbf{a}_2 \dots \mathbf{a}_{t-1}] \quad (1)$$

Afterwards, the weighted memory  $[\mathbf{a}_1 \mathbf{f}_1, \mathbf{a}_2 \mathbf{f}_2, \dots, \mathbf{a}_{t-1} \mathbf{f}_{t-1}]$  is generated by dot product. An attentive history  $\mathbf{h}_{t-1}^Q$  can be computed as:

$$\mathbf{h}_{t-1}^Q = \mathbf{a}_1 \mathbf{f}_1 + \mathbf{a}_2 \mathbf{f}_2 + \dots + \mathbf{a}_{t-1} \mathbf{f}_{t-1} \quad (2)$$

170 The next part is the information fusion. Different from the common methods of ‘GuessWhich’ based visual dialog which use the caption information only in the first round to initialize the question encoder,  $\text{AMN}^Q$  takes advantage of the caption  $\mathbf{C}$  in all rounds of the interaction. The caption  $\mathbf{C}$  can prompt the



generated questions more relevant on the condition that Q-BOT does not know  
 175 the image. Specifically,  $\text{AMN}^Q$  concatenates the caption  $\mathbf{C}$ , the embedded  $\mathbf{f}_t^Q$ ,  
 and the attentive history  $\mathbf{h}_{t-1}^Q$ . The concatenated information is taken as the  
 input of the history encoder. Moreover, the embedded fact  $\mathbf{f}_t^Q$  will be written  
 in the memory.

The history encoder is an LSTM that outputs the state of the current round  
 180  $\mathbf{s}_t^Q$ , which is used to generate the next question. Meanwhile, feature regression  
 network is a single fully-connected layer that takes  $\mathbf{s}_t^Q$  to produce a representa-  
 tion prediction  $\hat{\mathbf{y}}_t$  of the image  $\mathbf{I}$ .

**A-BOT** is also composed of five parts, including question encoder, atten-  
 tive memory network, history encoder, answer decoder, and fact encoder. The  
 185 question encoder is an LSTM which encodes the received question  $\mathbf{q}_t$  to a 512-d  
 vector which is input into the Attentive Memory Network  $\text{AMN}^A$ .

**Attentive memory network  $\text{AMN}^A$**  consists of memory reading and  
 information fusion like  $\text{AMN}^Q$ . And the memory also refers to the stored fact  
 $\mathbf{M}_t^A = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{t-1}]$ . But the query vector is different from the query in  
 $\text{AMN}^Q$ . In  $\text{AMN}^A$ , the encoded question  $\mathbf{q}_t$  is used to read the memory. The  
 process of generating attentive history  $\mathbf{h}_t^A$  is similar to that of  $\mathbf{h}_t^Q$ . Firstly, the  
 attention vector  $\mathbf{A}^A$  for the memory  $\mathbf{M}_t^A$  is produced:

$$\mathbf{A}^A = \text{Softmax} \left( \mathbf{q}_t (\mathbf{M}_t^A)^T \right) = [\mathbf{a}_1, \mathbf{a}_2 \dots \mathbf{a}_{t-1}] \quad (3)$$

Then the weighted memory  $[\mathbf{a}_1 \mathbf{f}_1, \mathbf{a}_2 \mathbf{f}_2, \dots, \mathbf{a}_{t-1} \mathbf{f}_{t-1}]$ . The attentive history  
 $\mathbf{h}_{t-1}^A$  can be computed as:

$$\mathbf{h}_{t-1}^A = \mathbf{a}_1 \mathbf{f}_1 + \mathbf{a}_2 \mathbf{f}_2 + \dots + \mathbf{a}_{t-1} \mathbf{f}_{t-1} \quad (4)$$

In the part of information fusion, the concatenated information including the  
 encoded question  $\mathbf{q}_t$ , the feature of the input image, and the attentive memory  
 $\mathbf{h}_{t-1}^A$  is input into the history encoder. The image feature is extracted from  
 190 the FC7 layer of VGG. The history encoder is an LSTM that produces a state  
 encoding of the current round  $\mathbf{s}_t^A$ . The answer decoder is also an LSTM that  
 generates an answer  $\mathbf{a}_t$  by natural language. Then the generated answers can

take context information into account. The accuracy of the answers can get a certain degree of promotion.

### 195 3.3. Training Strategy

We follow the training strategy used by [13]. It includes two training process: supervised training and fine-tuning by reinforcement learning. The supervised training aims to pre-train Q-BOT and A-BOT separately making the agents can communicate by natural language. Reinforcement learning helps the pre-trained  
200 model to achieve convergence by fine-tuning. Finally, Q-BOT and A-BOT accomplish interacting with natural language. These two training processes are both essential and interdependent.

**Supervised training.** A-BOT is first to be pre-trained on ImageNet which can guarantee A-BOT recognize some general scenes. Then Q-BOT and A-  
205 BOT are separately pre-trained on the train split of VisDial v1.0 [13] by a supervised manner. A Maximum Likelihood Estimation (MLE) objective is used to optimize the agents. Based on the ground truth dialog which is produced by imitating human conversation, Q-BOT can generate a series of questions, and A-BOT responses to the corresponding question. Considering image retrieval  
210 subtask, the feature regression network in Q-BOT is optimized by minimizing the Mean Squared Error (MSE) loss between the real feature of the image  $\mathbf{y}$  and the representation prediction  $\hat{\mathbf{y}}$ . Specifically, we pre-train the two agents for 15 epochs. These pre-training processes lay a foundation. But only the ground truth questions and answers are used as the history. There is no natural and  
215 meaningful interaction between Q-BOT and A-BOT actually. Thus, fine-tuning the pre-trained model by reinforcement learning is logical and essential.

**Reinforcement learning.** In this training process, the available information is only the image and its caption. Q-BOT and A-BOT interact with each other through self-talk. We firstly introduce five elements of reinforcement  
220 learning including action, state, environment, reward, and policy. And then the specific training procedure will be introduced.

Action space is composed of all possible sequential English words. In order to ensure questions and answers are non-redundant and precise, the action space should be large and diverse enough. Moreover, Q-BOT needs to make a representation prediction of the input image in each dialog round. The state between Q-BOT and A-BOT is different. The state of Q-BOT  $\mathbf{s}_t^Q$  contains the caption  $\mathbf{C}$  and the historical question-answer pairs. The state of A-BOT  $\mathbf{s}_t^A$  consists of the image  $\mathbf{I}$ , the current question  $\mathbf{q}_t$ , and the historical question-answer pairs. We define the parameters of those LSTMs and A-BOT as  $\theta^Q$  and  $\theta^A$ . The policy of Q-BOT aims to generate questions depending on the state  $\mathbf{s}_t^Q$  and the parameters of the LSTMs  $\theta^Q$  in Q-BOT. The policy of A-BOT aims to produce answers relying on the state  $\mathbf{s}_t^A$  and the parameters of the LSTMs  $\theta^A$ . Moreover, the policy about the feature regression network aims to output the representation prediction according to the state  $\mathbf{s}_t^Q$ , the answer  $\mathbf{a}_t$  and the parameters of the regression network  $\theta^f$ . The environment is just the input image and the caption. The reward of every round is defined as follow:

$$\mathbf{r}_t \left( \mathbf{s}_t^Q, (\mathbf{q}_t, \mathbf{a}_t, \mathbf{y}_t) \right) = l(\mathbf{y}_{t-1}, \mathbf{y}^{\text{gt}}) - l(\mathbf{y}_t, \mathbf{y}^{\text{gt}}) \quad (5)$$

where  $\ell(\cdot, \cdot)$  is the L-2 distance between the generated representation prediction and the feature of the real image. If  $\mathbf{r}_t \left( \mathbf{s}_t^Q, (\mathbf{q}_t, \mathbf{a}_t, \mathbf{y}_t) \right)$  is positive, the representation prediction is close to the true image, and the following retrieval accuracy can be improved. Inversely, if  $\mathbf{r}_t \left( \mathbf{s}_t^Q, (\mathbf{q}_t, \mathbf{a}_t, \mathbf{y}_t) \right)$  is negative, the quality of the representation prediction becomes bad, and the retrieval accuracy will also be reduced. Policy parameters  $\theta^Q$ ,  $\theta^A$ ,  $\theta^f$  are updated by REINFORCE [27] algorithm.

The supervised learning and reinforcement learning are combined as a curriculum. Firstly, we train the model by supervised learning for the  $N$  rounds, and then fine-tune the pre-trained model by reinforcement learning for the rest  $10 - K$  rounds. Then the value of  $K$  gradually decreases to 0. At last, the model converges depending on this training strategy.

## 4. Experiments

235 In this section, we firstly introduce the detailed implementations of the proposed model. And then our proposed method is evaluated on the VisDial v1.0 dataset and compared with the previous works. An ablation study is also conducted to test the effect of the important components in our model.

### 4.1. Implementation Details

240 We implement the proposed model using Pytorch and build on top of the publicly available implementations of [13]. The feature extractors for image, caption, question, and answer are all VGG-16 [28]. Specifically, the FC-7 features are selected. All the LSTMs in Q-BOT and A-BOT are single layer with  $512 - d$  hidden state. A-BOT is initialized by the model pre-trained on ImageNet. Then Q-BOT and A-BOT are separately pre-trained on VisDial v1.0  
245 with supervised learning for 15 epochs. The mini-batch size, learning rate, and weight decay are set as 20,  $1e - 3$  and 0.99 separately. The dropout rate used to alleviating over-fitting is set as 0.5. A-BOT produces 100 answers at each round, and the beam-search is set as 5 (only 5 answers are feedback to Q-BOT).  
250 The parameters in the reinforcement learning process are the same as the pre-training. All the experiments are conducted on two NVIDIA GeForce GTX 1080 Tian GPUs.

### 4.2. Dataset

**VisDial dataset v1.0** is the most popular dataset about visual dialog currently. It totally consists of  $130K$  dialogs and  $1.3M$  question-answer pairs on  
255 about  $130k$  images from the COCO dataset [29]. Specifically, the training set includes about  $120k$  dialogs and  $1.2M$  question-answer pairs on 123287 images from COCO-trainval, the validation and test set composed of  $100K$  question-answer pairs on  $10K$  images from Flickr are divided by the proportion of  $2 : 8$ .  
260 It is important to note that each dialog in the test set contains a random length within 10 rounds, while every dialog in training and validation set consists of

10 rounds. The distribution of the additional Flickr image and its caption can be well matched to the COCO dataset.

### 4.3. Results

265 There are two subtasks in our task, including visual dialog and image retrieval. We evaluate the generated visual dialog on the valid split of VisDial v1.0, and image retrieval on the test split of VisDial v1.0 separately.

#### 4.3.1. Results of Visual Dialog

The evaluation metrics for visual dialog are Mean Reciprocal Rank (MRR),  
 270 Mean Rank, and Recall@k. MRR and Mean Rank are the universal mechanisms to evaluate search algorithms, especially the field that including many results. In this task, each question in the 10 rounds produces 100 candidate answers. Mean Rank computes their average rank assigned to the ground truth answer over the whole dialog, and MRR computes the average reciprocal of their rank.  
 275 Recall@k replies to the percentage of the generated answers whose rank is less than  $k$ . Specifically, the value of  $K$  is set to 1, 5, 10. The three measures as mentioned above can comprehensively evaluate the performance of our model. Higher is better for MRR and Recall@k, and lower is better for Mean Rank.

We compare our method with 3 previous architectures, i.e., ADQ [17], RL  
 280 [13], and CR [15]. The comparison results are shown in Tab. 1. Our proposed method almost outperforms all of the existing algorithms. Especially, for the most important evaluation metric recall@K, our method is significantly better than the others. Moreover, we emphasize again that the task is the ‘Guess-Which’ based visual dialog, so we do not compare the methods that only focus  
 285 on the generation of visual dialog. The visualization results are shown as Fig. 2. Compared with RL [13], we can find that there are fewer repetitive question-answers. Especially in the last example, the number of repetitions dropped from five to two. The generated dialogs are more diverse and precise compared with the baseline.

model	Recall@1	Recall@5	Recall@10	Mean Rank	MRR
ADQ [17]	36.31	56.26	62.53	19.35	0.465
RL [13]	-	53.67	60.48	21.13	0.437
Ours	<b>42.17</b>	<b>63.24</b>	<b>68.65</b>	<b>15.82</b>	<b>0.613</b>

Table 1: The quality comparison of the generated dialog with other algorithms on VisDial v1.0. For recall@k and MRR, the higher the result, the better the performance. The Mean Rank is opposite.




Image & Caption	Ground Truth	Baseline	Ours
 <p>Caption: a man in the stance to throw a frisbee.</p>	<p>Is the photo in color? Yes  Any other people? No  Is the man old or young? Young  What color is the man? white  What color is his hair? Black  Is he wearing glasses? No  Is there a decal on the frisbee? Can't tell only can see the bottom  Is it daytime? yes  What color is the frisbee? White  Is this indoors or outdoors? outdoors</p>	<p>Is the photo in color? Yes  Is the person a man? Yes  Is the man wearing glasses? No  Is the person a woman? Yes  Is the man wearing glasses? No  Can you see any trees?  Is the man wearing glasses? No  Can you see any trees?  Is the man wearing a hat?  Can you see any trees?</p>	<p>Is the photo in color? Yes, it is in color.  How many people are there? Just 1 person.  Is the person male or female? No idea  Is the person wearing glasses? No, it is not visible.  What color are the glasses? I can't tell  Is this indoors or outdoors? It is outdoors  Can you see the ground? Yes, a little bit  Is the person wearing a hat? No, I can't see the hat.  What color is the frisbee? White  What is the man doing? He is throwing a white frisbee</p>
 <p>Caption: a portrait of 2 sheep looking directly at the camera.</p>	<p>What color are the sheep? Beige  Where are they? Hill  Is there a fence? Rock fence  Is there water? No  Do the sheep have a lot of wool? Yes  Do they look healthy? yes  Are they dirty? No  Are there trees? Yes  What color is the grass? Green  How is the weather? Is it sunny</p>	<p>Is this a color picture? Yes  Is it sunny? Yes  Is there a fence? No  Is the grass tall? No  Is the grass tall? No  Is the grass tall? No  Is the grass tall? No  Can you see any trees? Yes  Is the grass tall? No  Can you see any trees? Yes</p>	<p>What color are the sheep? They are both in gray.  Do you see a fence? Yes, rock fence.  Do the sheep have horns? I can't tell  Are they on the grass? Yes, the sheep are on the grass.  Can you see the sky? Not at all  What color is the grass? Green grass  Is the grass short or tall? Short  Is it sunny? Not really  Can you see the sky? I can't tell, but I think so  Do you see any trees? No</p>
 <p>Caption: a red double-decker bus UNK for passengers.</p>	<p>How many people do you see? 2  Are the windows tinted? No  Is the bus full? No  Is the road busy? No  Are there any other cars on the road? No  Is it sunny? No  Is it raining? Yes  Is it cloudy? Yes  Is there a number on the bus? Yes  What is the number? UNK</p>	<p>Is this a color photo? Yes  Is it a double-decker bus? No  Is the bus a double-decker? No  Can you see the driver? Yes  Is the bus a double-decker? No  Can you see the buildings? Yes  Can you see the buildings? Yes  Can you see the buildings? Yes  Can you see the buildings? Yes  Can you see the buildings? Yes</p>	<p>Is the photo in color? The picture is in color.  Is there a bus? Yes, a red bus  How many people on it? No, I can't tell  Can you see the buildings? No, I can't tell  Is there a bus? A red double-decker bus  Can you see the sky? Not at all  Is it sunny? I can't tell  Is it sunny? I can't tell  Is the bus double-decker? Yes, red  What color is the bus? Red</p>

Figure 2: The qualitative results of our method for visual dialog.

#### 290 4.3.2. Results of Image Retrieval

The evaluation metric for retrieval function is Image Retrieval Percentile that is computed by the rank of the ground truth representation in the sorted test image, according to their distance to the predicted feature. Then we compute the mean percential rank of the input image. If a percential rank is 90%, the ground truth image is closer to the prediction than 90% images in the test set. The image retrieval experiment is conducted on the test split of VisDial v1.0 as described in the dataset part. At the end of each dialog round, A-BOT feedbacks the responses selected from 100 candidate answers, and then Q-BOT generates

representation prediction. Finally, the image retrieval function is implemented.

300 In this section, we compared our method with several algorithms, including RL [13], ADQ [17], ML [15], LS [30]. The comparison results are shown in Tab. 2. And Fig. 3 shows the mean percetile rank of the input image for our proposed method and four baselines. Experiment results show that the image retrieval capability of our proposed method is better than all the baselines. The great  
305 contribution is that the attentive memory network can help our model make the most use of the history information and the caption of the source image. There are less repeated and invalid generated questions. Meanwhile, the answers also become more precise. Visual dialog and image retrieval can affect each other positively. Its visualization is shown in Fig. 4. In the last row, the retrieved  
310 images are similar to the input of image in general. Almost all of the images consist of a plate of food and a table. But there are some differences in more details, such as the attribute of the tables. Some tables in the retrieved images are not wooden. The reason is that the extracted features about the background is somewhat inadequate. This can be relieved by richer visual features.

#### 315 4.4. Ablation Study

In order to comprehensively and deeply analyze the proposed method and further demonstrate the effect of the attentive memory network, we present an extensive ablation study on VisDial v1.0 to evaluate the individual contributions of its major components. Thus, we consider different variants of the proposed  
320 method and compare their effect on the two subtasks separately.

**Ablation study on visual dialog.** As mentioned before, there are so many repeated questions that the interaction is inefficient in previous methods. So the additional caption and weighted history information are added to the process of question generation in each dialog round. Moreover, the attentive history is  
325 also used to boost the quality of the answers. Then the role of memory network is verified. We consider four AMN variants about the memory network: *Q-BOT (no MN)*, *A-BOT (no MN)*, *Q-BOT (MN)*, *A-BOT (no MN)*, *Q-BOT (no MN)*, *A-BOT (MN)*, *Q-BOT (MN)*, *A-BOT (MN)*. Their results are shown in Tab. 3.

Round	RL [13]	ADQ [17]	ML [15]	LS [30]	Ours
0	90.72	95.29	88.13	95.50	<b>96.84</b>
1	93.49	95.46	88.28	95.77	<b>97.15</b>
2	93.76	95.49	88.30	96.00	<b>97.25</b>
3	93.90	95.34	88.29	96.29	<b>97.33</b>
4	93.75	95.25	88.28	96.47	<b>97.40</b>
5	93.93	95.26	88.29	96.56	<b>97.46</b>
6	93.79	95.28	88.21	96.63	<b>97.51</b>
7	93.72	95.18	88.28	96.76	<b>97.55</b>
8	93.79	95.19	88.29	96.93	<b>97.58</b>
9	93.58	95.06	88.29	97.07	<b>97.60</b>
10	93.38	94.99	88.29	97.18	<b>97.63</b>

Table 2: The comparison of image retrieval capability with other algorithms. The higher the percentile, the better the performance

	Recall@1	Recall@5	Recall@10	Mean Rank	MRR
Q-BOT (no MN) A-BOT (no MN)	42.09	53.67	60.48	21.13	0.437
Q-BOT (MN) A-BOT (no MN)	39.82	59.13	63.78	19.92	0.498
Q-BOT (no MN) A-BOT (MN)	41.09	61.746	67.65	17.15	0.515
Q-BOT (MN) A-BOT (MN)	<b>42.172</b>	<b>63.24</b>	<b>68.48</b>	<b>15.82</b>	<b>0.613</b>

Table 3: Memory Network analysis of visual dialog on the validation set of VisDial v1.0.



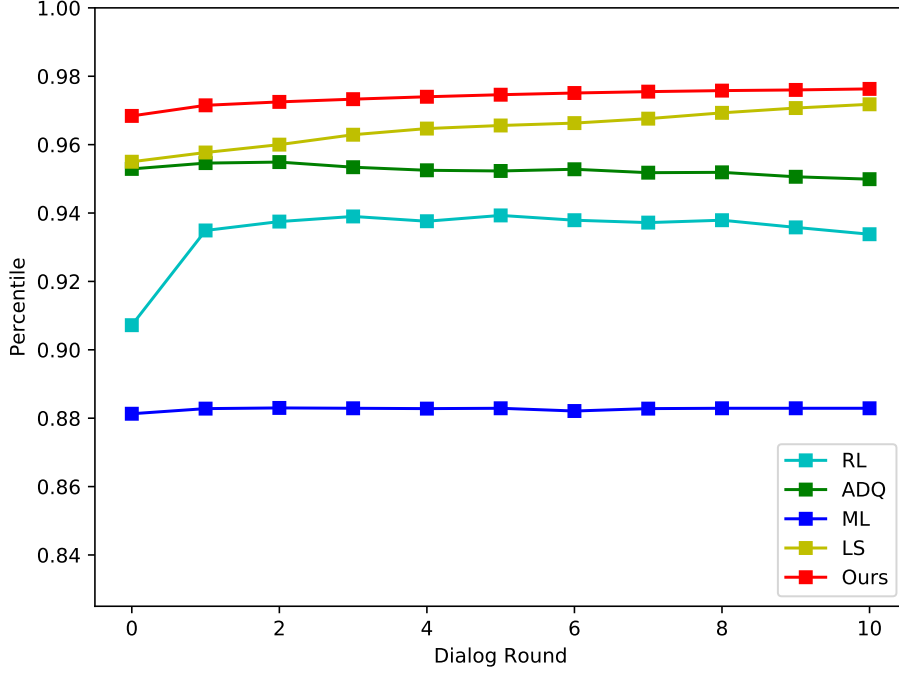


Figure 3: The image retrieval evaluation.

It can be observed that both the memory network in Q-BOT and A-BOT can significantly boost the quality of the generated visual dialog. *Q-BOT (MN)* and *A-BOT (MN)* performs better than *Q-BOT (no MN)*, *A-BOT (no MN)* by 13% on Recall@10 and 33% on Mean Rank. These results prove the importance of the memory network for visual dialog.

**Ablation study on image retrieval.** As emphasized before, this work focuses on the ‘GuessWhich’ based visual dialog. Thus, the effect of different components in AMN on image retrieval is evaluated in this part. We respectively verify the impact of memory network and caption information on image retrieval. The design of the variants is the same as the ablation study of visual dialog. The corresponding results about memory network analysis are shown in Tab. 4. Moreover, we also draw a more intuitive line chart as shown in Fig. 5. We have the following observations: 1) whether the memory network is in A-BOT or B-BOT, it can increase the performance of image retrieval on

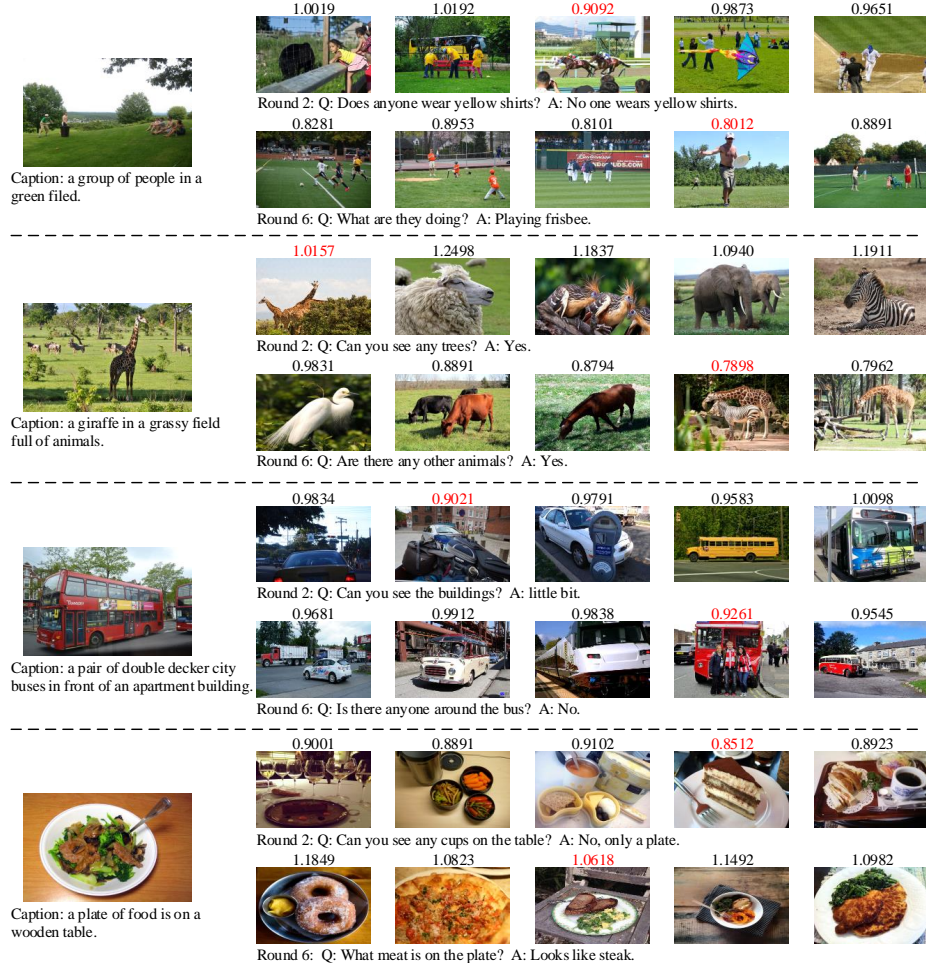


Figure 4: The qualitative results of our method for image retrieval. We randomly choose the round 2 and round 6 in the cooperative interaction. Those numbers represent L-2 distance between the predicted representation and the real image in fc7 space. The lower the value is, the more similar the image is. The retrieval results are intuitively accurate.

Round	Q-BOT (no MN)	Q-BOT (MN)	Q-BOT (no MN)	Q-BOT (MN)
	A-BOT (no MN)	A-BOT (no MN)	A-BOT (MN)	A-BOT (MN)
0	90.79	91.98	95.46	<b>96.83</b>
1	93.48	94.81	95.78	<b>97.15</b>
2	93.76	94.88	95.92	<b>97.25</b>
3	93.98	94.92	96.05	<b>97.33</b>
4	93.76	94.94	96.15	<b>97.40</b>
5	93.93	94.95	96.21	<b>97.46</b>
6	93.79	94.95	96.28	<b>97.51</b>
7	93.72	94.94	96.31	<b>97.55</b>
8	93.75	94.54	96.35	<b>97.58</b>
9	93.59	94.93	96.39	<b>97.60</b>
10	93.39	94.88	96.40	<b>97.63</b>

Table 4: Memory Network analysis of image retrieval on the test spilt of VisDial v1.0.

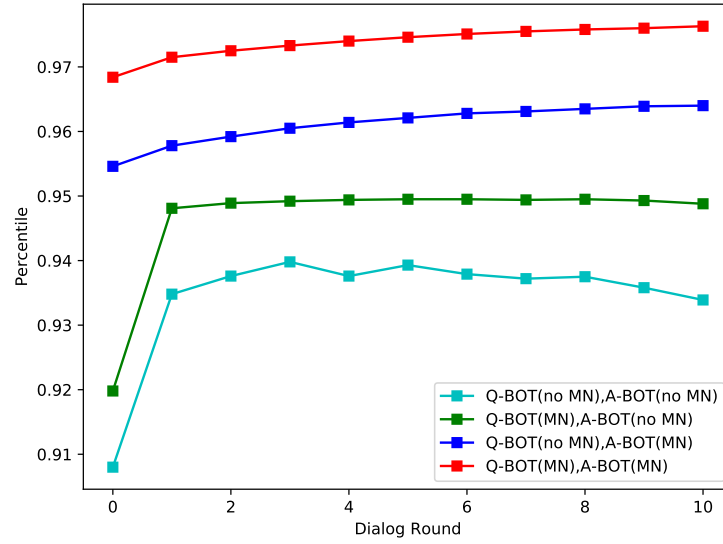


Figure 5: The comparison of the variants about memory network.

Round	Q-BOT (no caption)	Q-BOT (caption)
0	50.39	<b>95.69</b>
1	50.49	<b>95.74</b>
2	50.99	<b>95.77</b>
3	51.39	<b>96.08</b>
4	51.38	<b>96.15</b>
5	52.07	<b>96.23</b>
6	51.47	<b>97.13</b>
7	52.39	<b>97.01</b>
8	52.98	<b>97.02</b>
9	53.40	<b>97.02</b>
10	53.46	<b>97.03</b>

Table 5: Caption analysis of image retrieval on the test split of VisDial v1.0.

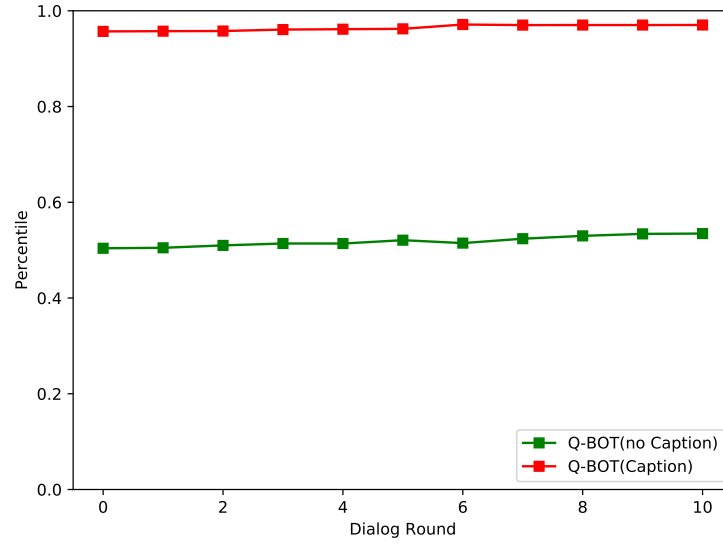


Figure 6: The comparison of the variants about caption information.

percentile. Compared with *Q-BOT (no MN)*, *A-BOT (no MN)*, *Q-BOT (MN)*, *A-BOT (MN)* improves retrieval quality about 4%. 2) *Q-BOT (no MN)*, *A-BOT (MN)* outperforms *Q-BOT (MN)*, *A-BOT (no MN)* by more than 2% on percentile. The memory network can help A-BOT generate more accurate answers which are important for retrieval quality, and the attentive history plays a more significant role in answers generation. The caption analysis and its line chart are shown in Tab. 5 and Fig. 6. Because the A-BOT can see the real image which contains more visual information, there is only a single variant of Q-BOT about the caption. The performance of the model without caption confusion in Q-BOT declines violently, which proves the importance of the caption information for image retrieval.

## 5. Conclusion

In this paper, we play a cooperative ‘image guessing’ game between Q-BOT and A-BOT to achieve effective visual dialog, and propose a novel Attentive Memory Network (AMN) which makes full use of historical dialog and caption information. The generated dialogs of existing methods are repeated frequently, and their interactions are often invalid. Moreover, the retrieval performance descends after some rounds. The AMN can help the Q-BOT to produce more valuable questions by memory reading and information fusion. And the attentive history dialogs also can improve the quality of answers. Meanwhile, the accuracy and stability of image retrieval will also be improved. The model is firstly pre-trained on VisDial v1.0, and then fine-tuned by reinforcement learning. Extensive experimental results have shown the superiority of our methods over the state-of-the-art ‘GuessWhich’ based visual dialog methods.

## References

- [1] L. Gao, X. Li, J. Song, H. T. Shen, Hierarchical lstms with adaptive attention for visual captioning, *IEEE transactions on pattern analysis and machine intelligence* (2019).

- [2] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 3156–3164.
- 375 [3] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, H. T. Shen, From deterministic to generative: Multimodal stochastic rnns for video captioning, IEEE transactions on neural networks and learning systems 30 (10) (2018) 3047–3058.
- 380 [4] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, Sequence to sequence-video to text, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4534–4542.
- [5] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 677–691.
- 385 [6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, VQA: visual question answering, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, 2015, pp. 2425–2433.
- 390 [7] Q. Wu, D. Teney, P. Wang, C. Shen, A. R. Dick, A. van den Hengel, Visual question answering: A survey of methods and datasets, Computer Vision and Image Understanding 163 (2017) 21–40.
- [8] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, C. Gan, Beyond rnns: Positional self-attention with co-attention for video question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8658–8665.
- 395 [9] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, D. Batra, Visual dialog, in: 2017 IEEE Conference on Com-

- puter Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA,  
 400 July 21-26, 2017, 2017, pp. 1080–1089.
- [10] D. Guo, H. Wang, M. Wang, Dual visual attention network for visual dialog,  
 in: Proceedings of the 28th International Joint Conference on Artificial  
 Intelligence, AAAI Press, 2019, pp. 4989–4995.
- [11] R. Shekhar, T. Baumgärtner, A. Venkatesh, E. Bruni, R. Bernardi, R. Fer-  
 405 nández, Ask no more: Deciding when to guess in referential visual dialogue,  
 in: Proceedings of the 27th International Conference on Computational  
 Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26,  
 2018, 2018, pp. 1218–1233.
- [12] R. Shekhar, A. Venkatesh, T. Baumgärtner, E. Bruni, B. Plank,  
 410 R. Bernardi, R. Fernández, Beyond task success: A closer look at jointly  
 learning to see, ask, and guesswhat, in: Proceedings of the 2019 Conference  
 of the North American Chapter of the Association for Computational Lin-  
 guistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis,  
 MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp.  
 415 2578–2587.
- [13] A. Das, S. Kottur, J. M. F. Moura, S. Lee, D. Batra, Learning cooperative  
 visual dialog agents with deep reinforcement learning, in: IEEE Interna-  
 tional Conference on Computer Vision, ICCV 2017, Venice, Italy, October  
 22-29, 2017, 2017, pp. 2970–2979.
- [14] J. Lu, A. Kannan, J. Yang, D. Parikh, D. Batra, Best of both worlds:  
 420 Transferring knowledge from discriminative learning to a generative visual  
 dialog model, in: Advances in Neural Information Processing Systems 30:  
 Annual Conference on Neural Information Processing Systems 2017, 4-9  
 December 2017, Long Beach, CA, USA, 2017, pp. 314–324.
- [15] A. Agarwal, S. Gurumurthy, V. Sharma, M. Lewis, K. P. Sycara, Com-  
 425 munity regularization of visually-grounded dialog, in: Proceedings of the

18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019, 2019, pp. 1042–1050.

- 430 [16] J. Zhang, T. Zhao, Z. Yu, Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, Melbourne, Australia, July 12-14, 2018, 2018, pp. 140–150.
- [17] V. Murahari, P. Chattopadhyay, D. Batra, D. Parikh, A. Das, Improving  
435 generative visual dialog by answering diverse questions, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, 2019, pp. 1449–1454.
- 440 [18] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, A. C. Courville, Guesswhat?! visual object discovery through multi-modal dialogue, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 4466–4475.
- 445 [19] H. Hu, X. Wu, B. Luo, C. Tao, C. Xu, W. Wu, Z. Chen, Playing 20 question game with policy-based reinforcement learning, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, 2018, pp. 3233–3242.
- [20] E. Abbasnejad, Q. Wu, Q. Shi, A. v. d. Hengel, What’s to know? uncertainty as a guide to asking goal-oriented questions, in: Proceedings of the  
450 IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4155–4164.
- [21] P. Shukla, C. Elmadjian, R. Sharan, V. Kulkarni, M. Turk, W. Y. Wang, What should i ask? using conversationally informative rewards for goal-oriented visual dialog, arXiv preprint arXiv:1907.12021 (2019).  
455



- [22] J. Weston, S. Chopra, A. Bordes, Memory networks, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
- 460 [23] S. Sukhbaatar, J. Weston, R. Fergus, et al., End-to-end memory networks, in: Advances in neural information processing systems, 2015, pp. 2440–2448.
- [24] C. Xiong, S. Merity, R. Socher, Dynamic memory networks for visual and textual question answering, in: International conference on machine learning, 2016, pp. 2397–2406.
- 465 [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
- [26] I. Danihelka, G. Wayne, B. Uria, N. Kalchbrenner, A. Graves, Associative long short-term memory, arXiv preprint arXiv:1602.03032 (2016).
- 470 [27] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine Learning 8 (1992) 229–256.
- [28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
- 475 [29] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, in: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, 2014, pp. 740–755.
- 480 [30] S. Lee, T. Gao, S. Yang, J. Yoo, J. Ha, Large-scale answerer in questioner’s mind for visual dialog question generation, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019.