

Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning

Qi Wu¹, Peng Wang², Chunhua Shen¹, Ian Reid¹, and Anton van den Hengel¹

¹Australian Centre for Robotic Vision, The University of Adelaide, Australia

²Northwestern Polytechnical University, China

Abstract

The Visual Dialogue task requires an agent to engage in a conversation about an image with a human. It represents an extension of the Visual Question Answering task in that the agent needs to answer a question about an image, but it needs to do so in light of the previous dialogue that has taken place. The key challenge in Visual Dialogue is thus maintaining a consistent, and natural dialogue while continuing to answer questions correctly. We present a novel approach that combines Reinforcement Learning and Generative Adversarial Networks (GANs) to generate more human-like responses to questions. The GAN helps overcome the relative paucity of training data, and the tendency of the typical MLE-based approach to generate overly terse answers. Critically, the GAN is tightly integrated into the attention mechanism that generates human-interpretable reasons for each answer. This means that the discriminative model of the GAN has the task of assessing whether a candidate answer is generated by a human or not, given the provided reason. This is significant because it drives the generative model to produce high quality answers that are well supported by the associated reasoning. The method also generates the state-of-the-art results on the primary benchmark.

1. Introduction

The combined interpretation of vision and language has enabled the development of a range of applications that have made interesting steps towards Artificial Intelligence, including Image Captioning [11, 34, 37], Visual Question Answering (VQA) [1, 22, 38], and Referring Expressions [10, 12, 41]. VQA, for example, requires an agent to answer a previously unseen question about a previously unseen image, and is recognised as being an AI-Complete problem [1]. Visual Dialogue [5] represents an extension to the VQA problem whereby an agent is required to engage in



Question	Human-like Responses	Machine-like
Are there any large building nearby?	No tall buildings but large one or two story buildings, and one clock is in front of looks like church of.	Yes there are.
With the clock does it look expensive?	Yes, I think so because it's made by stained glass.	I don't know.
Do you see any signs for church?	Yes, there is a sign with light on, but not clear enough.	Yes there are.

Figure 1: Human-like vs. Machine-like responses in a visual dialog. The human-like responses clearly answer the questions more comprehensively, and help to maintain a meaningful dialogue.

a dialogue about an image. This is significant because it demands that the agent is able to answer a series of questions, each of which may be predicated on the previous questions and answers in the dialogue. Visual Dialogue thus reflects one of the key challenges in AI and Robotics, which is to enable an agent capable of acting upon the world, that we might collaborate with through dialogue.

Due to the similarity between the VQA and Visual Dialog tasks, VQA methods [19, 40] have been directly applied to solve the Visual Dialog problem. The fact that the Visual Dialog challenge requires an ongoing conversation, however, demands more than just taking into consideration the state of the conversation thus far. Ideally, the agent should be an engaged participant in the conversation, cooperating towards a larger goal, rather than generating single

word answers, even if they are easier to optimise. Figure 1 provides an example of the distinction between the type of responses a VQA agent might generate and the more involved responses that a human is likely to generate if they are engaged in the conversation. These more human-like responses are not only longer, they provide reasoning information that might be of use even though it is not specifically asked for.

Previous Visual Dialog systems [5] follow a neural translation mechanism that is often used in VQA, by predicting the response given the image and the dialog history using the maximum likelihood estimation (MLE) objective function. However, because this over-simplified training objective only focus on measuring the word-level correctness, the produced responses tend to be generic and repetitive. For example, a simple response of ‘yes’, ‘no’, or ‘I don’t know’ can safely answer a large number of questions and lead to a high MLE objective value. Generating more comprehensive answers, and a deeper engagement of the agent in the dialogue, requires a more engaged training process.

A good dialogue generation model should generate responses indistinguishable from those a human might produce. In this paper, we introduce an adversarial learning strategy, motivated by the previous success of adversarial learning in many computer vision [3, 21] and sequence generation [4, 42] problems. We particularly frame the task as a reinforcement learning problem that we jointly train two sub-modules: a sequence generative model to produce response sentences on the basis of the image content and the dialog history, and a discriminator that leverages previous generator’s memories to distinguish between the human-generated dialogues and the machine-generated ones. The generator tends to generate responses that can fool the discriminator into believing that they are human generated, while the output of the discriminative model is used as a reward to the generative model, encouraging it to generate more human-like dialogue.

Although our proposed framework is inspired by generative adversarial networks (GANs) [9], there are several technical contributions that lead to the final success on the visual dialog generation task. First, we propose a sequential co-attention generative model that aims to ensure that attention can be passed effectively across the image, question and dialog history. The co-attended multi-modal features are combined together to generate a response. Secondly, and significantly, within the structure we propose the discriminator has access to the attention weights the generator used in generating its response. Note that the attention weights can be seen as a form of ‘reason’ for the generated response. For example, it indicates which region should be focused on and what dialog pairs are informative when generating the response. This structure is important as it allows the discriminator to assess the quality of the response, given

the reason. It also allows the discriminator to assess the response in the context of the dialogue thus far. Finally, as with most sequence generation problems, the quality of the response can only be assessed over the whole sequence. We follow [42] to apply Monte Carlo (MC) search to calculate the intermediate rewards.

We evaluate our method on the VisDial dataset [5] and show that it outperforms the baseline methods by a large margin. We also outperform several state-of-the-art methods. Specifically, our adversarial learned generative model outperforms our strong baseline MLE model by 1.87% on recall@5, improving over previous best reported results by 2.14% on recall@5, and 2.50% recall@10. Qualitative evaluation shows that our generative model generates more informative responses and a human study shows that 49% of our responses pass the Turing Test. We additionally implement a model under the discriminative setting (a candidate response list is given) and achieve the state-of-the-art performance.

2. Related work

Visual dialog is the latest in a succession of vision-and-language problems that began with image captioning [11, 34, 37], and includes visual question answering [1, 22, 38]. However, in contrast to these classical vision-and-language tasks that only involve at most a single natural language interaction, visual dialog requires the machine to hold a meaningful dialogue in natural language about visual content. Mostafazadeh *et al.* [20] propose an Image Grounded Conversation (IGC) dataset and task that requires a model to generate natural-sounding conversations (both questions and responses) about a shared image. De Vries *et al.* [7] propose a GuessWhat game style dataset, where one person asks questions about an image to guess which object has been selected, and the second person answers questions in yes/no/NA. Das *et al.* [5] propose the largest visual dialog dataset, VisDial, by pairing two subjects on Amazon Mechanical Turk to chat about an image. They further formulate the task as a ‘multi-round’ VQA task and evaluate individual responses at each round in a retrieval or multiple-choice setup. Recently, Das *et al.* [6] propose to use RL to learn the policies of a ‘Questioner-Bot’ and an ‘Answerer-Bot’, based on the goal of selecting the right images that the two agents are talking, from the VisDial dataset.

Concurrent with our work, Lu *et al.* [18] propose a similar generative-discriminative model for Visual Dialog. However, there are two differences. First, their discriminative model requires to receive a list of candidate responses and learns to sort this list from the training dataset, which means the model only can be trained when such information is available. Second, their discriminator only considers the generated response and the provided list of candidate responses. Instead, we measure whether the generated

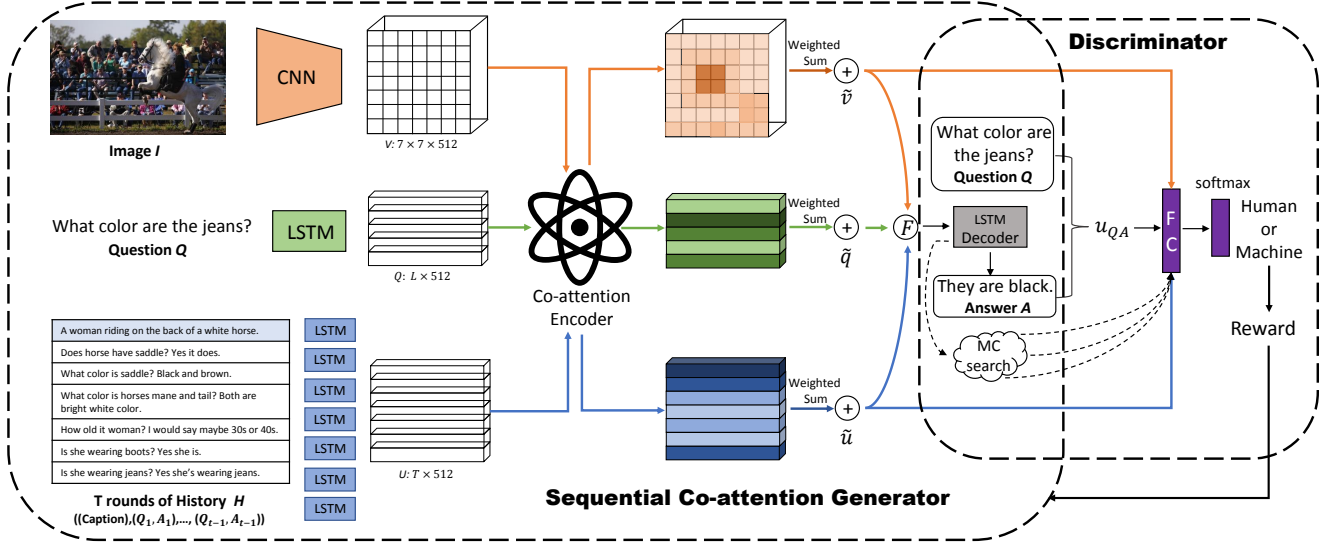


Figure 2: The adversarial learning framework of our proposed model. Our model is composed of two components, the first being a sequential co-attention generator that accepts as input *image*, *question* and *dialog history* tuples, and uses the co-attention encoder to jointly reason over them. The second component is a discriminator tasked with labelling whether each answer has been generated by a human or the generative model by considering the attention weights. The output from the discriminator is used as a reward to push the generator to generate responses that are indistinguishable from those a human might generate.

response is valid given the attention weights which reflect both the reasoning of the model, and the history of the dialogue thus far. As we show in our experiments in Sec. 4, this procedure results in our generator producing more suitable responses.

Dialog generation in NLP Text-only dialog generation [15, 16, 23, 30, 39] has been studied for many years in the Natural Language Processing (NLP) literature, and has led to many applications. Recently, the popular ‘XiaoIce’ produced by Microsoft and the ‘Its Alive’ chatbot created by Facebook have attracted significant public attention. In NLP, dialog generation is typically viewed as a sequence-to-sequence (Seq2Seq) problem, or formulated as a statistical machine translation problem [23, 30]. Inspired by the success of the Seq2Seq model [32] in the machine translation, [26, 33] build end-to-end dialog generation models using an encoder-decoder model. Reinforcement learning (RL) has also been applied to train a dialog system. Li *et al.* [15] simulate two virtual agents and handcraft three rewards (informativity, coherence and ease of answering) to train the response generation model. Recently, some works make an effort to integrate the Seq2Seq model and RL. For example, [2, 31] introduce real users by combining RL with neural generation.

Li *et al.* in [16] were the first to introduce GANs for dialogue generation as an alternative to human evaluation. They jointly train a generative (Seq2Seq) model to produce response sequences and a discriminator to distinguish between human, and machine-generated responses. Although we also introduce an adversarial learning framework to the

visual dialog generation in this work, one of the significant differences is that we need to consider the visual content in both generative and discriminative components of the system, where the previous work [16] only requires textual information. We thus designed a sequential co-attention mechanism for the generator and an attention memory access mechanism for the discriminator so that we can jointly reason over the visual and textual information. Critically, the GAN we proposed here is tightly integrated into the attention mechanism that generates human-interpretable reasons for each answer. It means that the discriminative model of the GAN has the task of assessing whether a candidate answer is generated by a human or not, given the provided reason. This is significant because it drives the generative model to produce high quality answers that are well supported by the associated reasoning. More details about our generator and discriminator can be found in Sections 3.1 and 3.2 respectively.

Adversarial learning Generative adversarial networks [9] have enjoyed great successes in a wide range of applications in Computer Vision, [3, 21, 24], especially in image generation tasks [8, 43]. The learning process is formulated as an adversarial game in which the generative model is trained to generate outputs to fool the discriminator, while the discriminator is trained not to be fooled. These two models can be jointly trained end-to-end. Some recent works have applied the adversarial learning to sequence generation, for example, Yu *et al.* [42] backpropagate the error from the discriminator to the sequence generator by using policy gradient reinforcement learning. This

model shows outstanding performance on several sequence generation problems, such as speech generation and poem generation. The work is further extended to more tasks such as image captioning [4, 28] and dialog generation [16]. Our work is also inspired by the success of adversarial learning, but we carefully extend it according to our application, *i.e.* the Visual Dialog. Specifically, we redesign the generator and discriminator in order to accept multi-modal information (visual content and dialog history). We also apply an intermediate reward for each generation step in the generator, more details can be found in Sec. 3.3.

3. Adversarial Learning for Visual Dialog Generation

In this section, we describe our adversarial learning approach to generating natural dialog responses based on an image. There are several ways of defining the visual based dialog generation task [7, 20]. We follow the one in [5], in which an image I , a ‘ground truth’ dialog history (including an image description C) $H = (C, (Q_1, A_1), \dots, (Q_{t-1}, A_{t-1}))$ (we define each Question-Answer (QA) pair as an utterance U_t , and $U_0 = C$), and the question Q are given. The visual dialog generation model is required to return a response sentence $\hat{A} = [a_1, a_2, \dots, a_K]$ to the question, where K is the length (number of words) of the response answer. As in VQA, two types of models may be used to produce the response — generative and discriminative. In a generative decoder, a word sequence generator (for example, an RNN) is trained to fit the ground truth answer word sequences. For a discriminative decoder, an additional candidate response vocabulary is provided and the problem is re-formulated as a multi-class classification problem. The biggest limitation of the discriminative style decoder is that it only can produce a response if and only if it exists in the fixed vocabulary. Our approach is based on a generative model because a fixed vocabulary undermines the general applicability of the model, but also because it offers a better prospect of being extensible to the problem of generating more meaningful dialogue in future.

In terms of reinforcement learning, our response sentence generation process can be viewed as a sequence of prediction actions that are taken according to a policy defined by a sequential co-attention generative model. This model is critical as it allows attention (and thus reasoning) to pass across image, question, and dialogue history equally. A discriminator is trained to label whether a response is human generated or machine generated, conditioned on the image, question and dialog attention memories. Considering here that as we take the dialog and the image as a whole into account, we are actually measuring whether the generated response can be fitted into the visual dialog. The output from this discriminative model is used as a reward to the previous generator, pushing it to generate responses that

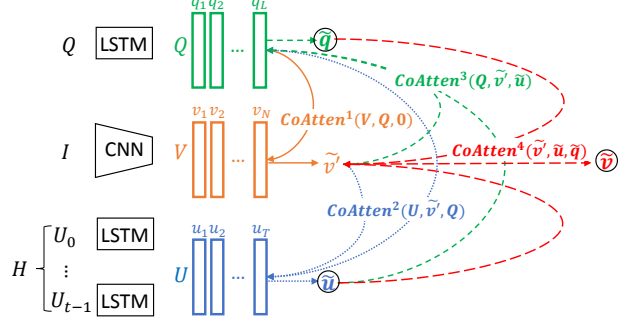


Figure 3: The sequential co-attention encoder. Each input feature is co-attend by the other two features in a sequential fashion, using the Eq.1-3. The number on each function indicates the sequential order, and the final attended features \tilde{u} , \tilde{v} and \tilde{q} form the output of the encoder.

are more fitting with the dialog history. In order to consider the reward at the local (*i.e.* word and phase) level, we use a Monte Carlo (MC) search strategy and the REINFORCE algorithm [36] is used to update the policy gradient. An overview of our model can be found in the Fig. 2. In the following sections, we will introduce each component of our model separately.

3.1. A sequential co-attention generative model

We employ the encoder-decoder style generative model which has been widely used in the sequence generation problems. In contrast to text-only dialog generation problem that only needs to consider the dialog history, however, visual dialog generation additionally requires the model to understand visual information. And distinct from VQA that only has one round of questioning, visual dialog has multiple rounds of dialog history that need to be accessed and understood. It suggests that an encoder that can combine multiple information sources is required. A naive way of doing this is to represent the inputs - image, history and question separately and then concatenate them to learn a joint representation. We contend, however, that it is more powerful to let the model selectively focus on regions of the image and segments of the dialog history according to the question.

Based on this, we propose a sequential co-attention mechanism [35]. Specifically, we first use a pre-trained CNN [29] to extract the spatial image features $V = [v_1, \dots, v_N]$ from the convolutional layer, where N is the number of image regions. The question features is $Q = [q_1, \dots, q_L]$, where $q_l = LSTM(w_l, q_{l-1})$, which is the hidden state of an LSTM at step l given the input word w_l of the question. L is the length of the question. Because the history H is composed by a sequence of utterance, we extract each utterance feature separately to make up the dialog history features, *i.e.*, $U = [u_0, \dots, u_T]$, where T is the number of rounds of the utterance (QA-pairs). And each u is the last hidden state of an LSTM, which accepts the

utterance words sequences as the input.

Given the encoded image, dialog history and question feature V, U and Q , we use a co-attention mechanism to generate attention weights for each feature type using the other two as the guidance in a sequential style. Each co-attention operation is denoted as $\tilde{x} = \text{CoAtten}(X, g_1, g_2)$, which can be expressed as follows:

$$H_i = \tanh(\mathbf{W}_x x_i + \mathbf{W}_{g_1} g_1 + \mathbf{W}_{g_2} g_2), \quad (1)$$

$$\alpha_i = \text{softmax}(\mathbf{W}^T H_i), \quad i = 1, \dots, M, \quad (2)$$

$$\tilde{x} = \sum_{i=1}^M \alpha_i x_i, \quad (3)$$

where X is the input feature sequence (*i.e.*, V, U or Q), and $g_1, g_2 \in \mathbb{R}^d$ represent guidances that are outputs of previous attention modules. Here d is the feature dimension. $\mathbf{W}_x, \mathbf{W}_{g_1}, \mathbf{W}_{g_2} \in \mathbb{R}^{h \times d}$ and $\mathbf{W} \in \mathbb{R}^h$ are learnable parameters. Here h denotes the size of hidden layers of the attention module. M is the input sequence length that corresponding to the N, L and T for different feature inputs.

As shown in Fig. 3, in our proposed process, the initial question feature is first used to attend to the image. The weighted image features and the initial question representation are then combined to attend to utterances in the dialog history, to produce the attended dialog history (\tilde{u}). The attended dialog history and weighted image region features are then jointly used to guide the question attention (\tilde{q}). Finally, we run the image attention (\tilde{v}) again, guided by the attended question and dialog history, to complete the circle. All three co-attended features are concatenated together and embedded to the final feature F :

$$F = \tanh(\mathbf{W}_{eg}[\tilde{v}; \tilde{u}; \tilde{q}]) \quad (4)$$

where $[\cdot]$ is a concatenation operator. Finally, this vector representation is fed to an LSTM to compute the probability of generating each token in the target using a softmax function, which forms the response \hat{A} . The whole generation process is denoted as $\pi(\hat{A}|V, U, Q)$.

3.2. A discriminative model with attention memories

Our discriminative model is a binary classifier that is trained to distinguish whether the input dialog is generated by humans or machines. In order to consider the visual information and the dialog history, we allow the discriminator to access to the attention memories in the generator. Specifically, our discriminator takes $\{\tilde{v}, \tilde{u}, Q, \hat{A}\}$ as the input, where \tilde{v}, \tilde{u} are the attended image and dialog history features produced in the generative model¹, given the question Q . And \hat{A} is the generated response in the generator. The Q - \hat{A} pair is further sent to an LSTM to obtain a vector

¹we also tested to use the question memory \tilde{q} , but we find the discriminator result is not as good as when using the original question input Q .

representation $u_{Q\hat{A}}$. All three features are embedded together and sent to a 2-way softmax function, which returns the probability distribution of whether the whole visual dialog is human-natural or not:

$$O = \tanh(\mathbf{W}_{ed}[\tilde{v}; \tilde{u}; u_{Q\hat{A}}]) \quad (5)$$

$$P = \text{softmax}(O) \quad (6)$$

The probability of the visual dialog being recognised as a human-generated dialog is denoted as $r(\{\tilde{v}, \tilde{u}, Q, \hat{A}\})$.

3.3. Adversarial REINFORCE with an intermediate reward

In adversarial learning, we encourage the generator to generate responses that are close to human generated dialogs, or, in our case, we want the generated response can fit into the visual dialog as good as possible. The policy gradient methods are used here to achieve the goal. The probability of the visual dialog being recognised as a human-generated dialog by the discriminator (*i.e.*, $r(\{\tilde{v}, \tilde{u}, Q, \hat{A}\})$) is used as a reward for the generator, which is trained to maximize the expected reward of generated response using the REINFORCE algorithm [36]:

$$J(\theta) = \mathbf{E}_{\hat{A} \sim \pi(\hat{A}|V, U, Q)}(r(\{\tilde{v}, \tilde{u}, Q, \hat{A}\})|\theta) \quad (7)$$

Given the input visual information (V), question (Q) and dialog history utterances (U), the generator generates an response answer \hat{A} by sampling from the policy. The attended visual (\tilde{v}) and dialog (\tilde{u}) memories with the Q and generated answer \hat{A} are concatenated together and fed to the discriminator. We further use the likelihood ratio trick [36] to approximate the gradient of Eq. 7:

$$\begin{aligned} \nabla J(\theta) &\approx \nabla \log \pi(\hat{A}|V, U, Q) \cdot [r(\{\tilde{v}, \tilde{u}, Q, \hat{A}\}) - b] \\ &= \nabla \sum_k \log p(a_k|V, U, Q, a_{1:k-1}) \cdot [r(\{\tilde{v}, \tilde{u}, Q, \hat{A}\}) - b] \end{aligned} \quad (8)$$

where p is the probability of the generated responses words, a_k is the k -th word in the response. b denotes the baseline value. Following [16], we train a critic neural network to estimate the baseline value b by given the current state under the current generation policy π . The critic network takes the visual content, dialog history and question as input, encodes them to a vector representation with our co-attention model and maps the representation to a scalar. The critic neural network is optimised based on the mean squared loss between the estimated reward and the real reward obtained from the discriminator. The entire model can be trained end-to-end, with the discriminator updating synchronously. We use the human generated dialog history and answers as the positive examples and the machine generated responses as negative examples.

Intermediate reward An issue in the above vanilla REINFORCE is it only considers a reward value for a finished sequence, and the reward associated with this sequence is used for all actions, *i.e.*, the generation of each token. However, as a sequence generation problem, rewards for intermediate steps are necessary. For example, given a question ‘Are they adults or babies?’, the human-generated answer is ‘I would say they are adults’, while the machine-generated answer is ‘I can’t tell’. The above REINFORCE model will give the same low reward to all the tokens for the machine-generated answer, but a proper reward assignment way is to give the reward separately, *i.e.*, a high reward to the token ‘I’ and low rewards for the token ‘can’t’ and ‘tell’.

Considering that the discriminator is only trained to assign rewards to fully generated sentences, but not intermediate ones, we propose to use the Monte Carlo (MC) search with a roll-out (generator) policy π to sample tokens. An N-time MC search can be represented as:

$$\{\hat{A}_{1:K}^1, \dots, \hat{A}_{1:K}^N\} = \text{MC}^\pi(\hat{A}_{1:k}; N) \quad (9)$$

where $\hat{A}_{1:k}^n = (a_1, \dots, a_k)$ and $\hat{A}_{k+1:K}^n$ are sampled based on the roll-out policy π and the current state. We run the roll-out policy starting from the current state till the end of the sequence for N times and the N generated answers share a common prefix $\hat{A}_{1:k}$. These N sequences are fed to the discriminator, the average score

$$r_{a_k} = \frac{1}{N} \sum_{n=1}^N r(\{\tilde{v}, \tilde{u}, Q, \hat{A}_{1:K}^n\}) \quad (10)$$

of which is used as a reward for the action of generating the token a_k . With this intermediate reward, our gradient is computed as:

$$\nabla J(\theta) = \nabla \sum_k \log p(a_k | V, U, Q, a_{1:k-1}) \cdot [r_{a_k} - b] \quad (11)$$

where we can see the intermediate rewards for each generation action are considered.

Teacher forcing Although the reward returned from the discriminator has been used to adjust the generation process, we find it is still important to feed human generated responses to the generator for the model updating. Hence, we apply a teacher forcing [14, 16] strategy to update the parameters in the generator. Specifically, at each training iteration, we first update the generator using the reward obtained from the sampled data with the generator policy. Then we sample some data from the real dialog history and use them to update the generator, with a standard maximum likelihood estimation (MLE) objective. The whole training process is reviewed in the Alg. 1.

Algorithm 1 Training Visual Dialog Generator with REINFORCE

Require: Pretrained generator Gen and discriminator Dis

- 1: **for** Each iteration **do**
- 2: # Train the generator Gen
- 3: **for** $i=1$, steps **do**
- 4: Sample (I, H, Q, A) from the real data
- 5: Sample $(\tilde{v}, \tilde{u}, \hat{A}) \sim Gen_\pi(\cdot | I, H, Q)$
- 6: Compute Reward r for $(\tilde{v}, \tilde{u}, Q, \hat{A})$ using Dis
- 7: Evaluate $\nabla J(\theta)$ with Eq. 8 or 11 depends on whether the intermediate reward (Eq. 10) is used
- 8: Update Gen parameter θ using $\nabla J(\theta)$
- 9: Update baseline parameters for b
- 10: Teacher-Forcing: Update Gen on (I, H, Q, A) using MLE
- 11: # Train the discriminator Dis
- 12: Sample (I, H, Q, A) from the real data
- 13: Sample $(\tilde{v}, \tilde{u}, \hat{A}) \sim Gen_\pi(\cdot | I, H, Q)$
- 14: Update Dis using $(\tilde{v}, \tilde{u}, Q, A)$ as positive examples and $(\tilde{v}, \tilde{u}, Q, \hat{A})$ as negative examples

4. Experiments

We evaluate our model on a recently published visual dialog generation dataset, VisDial [5]. Images in Visdial are all from the MS COCO [17], which contain multiple objects in everyday scenes. The dialogs in Visdial are collected by pairing 2 AMT works (a ‘questioner’ and an ‘answerer’) to chat with each other about an image. To make the dialog measurable, the image remains hidden to the questioner and the task of the questioner is to ask questions about this hidden image to imagine the scene better. The answerer sees the image and his task is to answer questions asked by the questioner. Hence, the conversation is more like multi-rounds of visual based question answering and it only can be ended after 10 rounds. There are 83k dialogs in the COCO training split and 40k in the validation split, for totally 1,232,870 QA pairs, in the Visdial v0.9, which is the latest available version thus far. Following [17], we use 80k dialogs for train, 3k for val and 40k as the test.

4.1. Evaluation Metrics

Different from the previous language generation tasks that normally use BLEU, MENTOR or ROUGE score for evaluation, we follow [17] to use a retrieval setting to evaluate the individual responses at each round of a dialog. Specifically, at test time, besides the image, ground truth dialog history and the question, a list of 100 candidates answers are also given. The model is evaluated on retrieval metrics: (1) rank of human response, (2) existence of the human response in top- k ranked responses, *i.e.*, recall@ k and (3) mean reciprocal rank (MRR) of the human response. Since we focus on evaluating the generalization ability of our generator, we simply rank the candidates by the generative model’s log-likelihood scores.




Image+Caption	Question	Human Answer	CoAtt-G-MLE	Ours
 <p>A bathroom with a white bath tub, sink and large window.</p>	What color is the bathroom?	The walls are gray	White	Most white
	Are there any people in there?	No	No	No
	Are there towels hanging?	No folded up	No	No, on the floor
	Is there any soap on the sink?	I do n't think so	No soap	I do n't think so
	What color are the towels?	White	White	White
	What kind of bathtub is it?	A fancy rectangular	It 's a tub	It 's a shower tub with a shower
	Can you see anything out the bathroom window?	No	No	No, just the wall
	Are there curtains on the window?	No	No	No curtains
	Is the bathroom light on?	Yes	Yes	Yes
	Is there anything else on the sink?	No	No	No
 <p>A motorcycle, moped and a bus parked by the street.</p>	What color is the motorcycle?	It is black and white	White and blue	It's black and white
	Is this on a busy street with shops and people?	It looks like it is not	No	No it is not
	Is it daylight or night time?	Daytime	It 's daytime	It is daytime
	Is the photo in color?	Yes it is	Yes	Yes
	What color are the other cars?	I see a white van and a blue	white and black	One is blue and the other is white
	Are there any people walking?	Not that i can see	no	no, there are no people
	Can you tell what shops businesses they are?	Not really	i 'm not sure	I 'm not sure , they are in the background
	Do you see any traffic lights?	No, i do not	No i do n't	No i do n't
	Do you think the motorcycle should be parked on the sidewalk?	Yes	No	No, it looks like it 's parked
	Do you see any signs?	One, but only a picture	Yes	I see a sign on the side of road
 <p>A man in a wet suit carrying a surfboard by some rocks.</p>	Is the photo in color?	Yes	Yes	Yes
	How old does the man appear to be?	I would estimate late 30s	20 's	I would say 20 's
	What color wetsuit?	Dark blue	Black	Black
	What color surfboard?	White and red	White with red	It 's white with red
	Do the rocks appear to be smooth or sharp?	I would guess they are smooth	Smooth	They look smooth
	Is he close to the water?	Moderately close	No	Yes
	Does it appear to be a beach or private section?	Private area	I ca n't tell	I ca n't tell
	What color is the water dark or light blue?	It is blurry so it appears black	light blue	It 's light blue
	Does he have any shoes on?	I ca n't see his feet	I ca n't see his feet	I ca n't see his feet
	Does he appear to be wet or dry?	Dry	Dry	He looks dry

Figure 4: Qualitative results of our model (CoAtt-GAN-w/ R_{inte} -TF) comparing to human ground-truth answer and our baseline model.

4.2. Implementation Details

To pre-process the data, we first lowercase all the texts, convert digits to words, and remove contractions, before tokenizing. The captions, questions and answers are further truncated to ensure that they are no longer than 40, 20 and 20, respectively. We then construct the vocabulary of words that appear at least 5 times in the training split, giving us a vocabulary of 8845 words. The words are represented as one-hot vector and 512-d embeddings for the words are learned. These word embeddings are shared across question, history, decoder LSTMs. All the LSTMs in our model are 1-layered with 512 hidden states. The Adam [13] optimizer is used with the base learning rate of 10^{-3} , further decreasing to 10^{-5} . We use 5-time Monte Carlo (MC) search for each token. The co-attention generative model is pre-trained using the ground-truth dialog history for 30 epochs. We also pre-train our discriminator (for 30 epochs), where the positive examples are sampled from the ground-truth dialog, the negative examples are sampled from the dialog generated by our generator. The discriminator is updated after every 20 generator-updating steps.

4.3. Experiment results

Baselines and comparative models We compare our model with a number of baselines and state-of-the-art models. **Answer Prior** [5] is a naive baseline that encodes answer options with an LSTM and scored by a linear classifier, which captures ranking by frequency of answers in the training set. **NN** [5] finds the nearest neighbor images and questions for a test question and its related image. The op-

Model	MRR	R@1	R@5	R@10	Mean
Answer Prior [5]	0.3735	23.55	48.52	53.23	26.50
NN [5]	0.4274	33.13	50.83	58.69	19.62
LF [5]	0.5199	41.83	61.78	67.59	17.07
HRE [5]	0.5237	42.29	62.18	67.92	17.07
HREA [5]	0.5242	42.28	62.33	68.17	16.79
MN [5]	0.5259	42.29	62.85	68.88	17.06
HCIAE [18]	0.5386	44.06	63.55	69.24	16.01
CoAtt-G-MLE	0.5411	44.32	63.82	69.75	16.47
CoAtt-GAN-w/o R_{inte}	0.5415	44.52	64.17	70.31	16.28
CoAtt-GAN-w/ R_{inte}	0.5506	45.56	65.16	71.07	15.30
CoAtt-GAN-w/ R_{inte} -TF	0.5578	46.10	65.69	71.74	14.43

Table 1: Performance of generative methods on VisDial v0.9. Higher is better for MRR and recall@k, while lower is better for mean rank.

tions are then ranked by their mean-similarity to answers to these questions. **Late Fusion (LF)** [5] encodes the image, dialog history and question separately and later concatenated together and linearly transformed to a joint representation. **HRE** [5] applies a hierarchical recurrent encoder [27] to encode the dialog history and the **HREA** [5] additionally adds an attention mechanism on the dialogs. **Memory Network (MN)** [5] maintains each previous question and answer as a ‘fact’ in its memory bank and learns to refer to the stored facts and image to answer the question. A concurrent work [18] proposes a **HCIAE** (History-Conditioned Image Attentive Encoder) to attend on image and dialog features.

From Table 1, we can see our final generative model **CoAtt-GAN-w/ R_{inte} -TF** performs the best on all the evaluation metrics. Comparing to the previous state-of-the-art model MN [5], our model outperforms it by 3.81% on R@1. We also produce better results than the HCIAE [18] model,

Model	MRR	R@1	R@5	R@10	Mean
LF [5]	0.5807	43.82	74.68	84.07	5.78
HRE [5]	0.5846	44.67	74.50	84.22	5.72
HREA [5]	0.5868	44.82	74.81	84.36	5.66
MN [5]	0.5965	45.55	76.22	85.37	5.46
SAN-QI [40]	0.5764	43.44	74.26	83.72	5.88
HieCoAtt-QI [19]	0.5788	43.51	74.49	83.96	5.84
AMEM [25]	0.6160	47.74	78.04	86.84	4.99
HCIAE-NP-ATT [18]	0.6222	48.48	78.75	87.59	4.81
Ours	0.6398	50.29	80.71	88.81	4.47

Table 2: Performance of discriminative methods on VisDial v0.9. Higher is better for MRR and recall@k, while lower is better for mean rank.

which is the previous best results that without using any discriminative knowledges. Figure 4 shows some qualitative results of our model. More results can be found in the supplementary material.

Ablation study Our model contains several components. In order to verify the contribution of each component, we evaluate several variants of our model.

- **CoAtt-G-MLE** is the generative model that uses our co-attention mechanism shown in Sec. 3.1. This model is trained only with the MLE objective, without any adversarial learning strategies. Hence, it can be used as a baseline model for other variants.
- **CoAtt-GAN-w/o R_{inte}** is the extension of above CoAtt-G model, with an adversarial learning strategy. The reward from the discriminator is used to guide the generator training, but we only use the global reward to calculate the gradient, as shown in Equ. 8.
- **CoAtt-GAN-w/ R_{inte}** uses the intermediate reward as shown in the Equ. 10 and 11.
- **CoAtt-GAN-w/ R_{inte} -TF** is our final model which adds a ‘teacher forcing’ after the adversarial learning.

Our baseline **CoAtt-G-MLE** model outperforms the previous attention based models (HREA, MN, HCIAE) shows that our co-attention mechanism can effectively encode the complex multi-source information. **CoAtt-GAN-w/o R_{inte}** produces slightly better results than our baseline model by using the adversarial learning network, but the improvement is limited. The intermediate reward mechanism contributes the most to the improvement, *i.e.*, our proposed **CoAtt-GAN-w/ R_{inte}** model improves over our baseline by average 1%. The additional Teacher-Forcing model (our final model) brings the further improvement, by average 0.5%, achieving the best results.

Discriminative setting We additionally implement a model for the discriminative task on the Visdial dataset [5]. In this discriminative setting, there is no need to generate a string, instead, a pre-defined answer set is given and the problem is formulated as a classification problem. We modify our model by replacing the response generation LSTM (can be treated as a multi-step classification process) as a single-step classifier. **HCIAE-NP-ATT [18]** is the origi-

	MN [5]	CoAtt-G-MLE	Ours
M1: Percentage of responses that pass the Turing Test	0.39	0.46	0.49
M2: Percentage of responses that are evaluated as better or equal to human responses.	0.36	0.42	0.45

Table 3: Human evaluation on 1000 sampled responses on VisDial v0.9

nal HCIAE model with a n-pair discriminative loss and a self-attention mechanism. **AMEM [25]** applies a more advanced memory network to model the dependency of current question on previous attention. Additional two VQA models [19, 40] are used for comparison. Table 2 shows that our model outperforms the previous baseline and state-of-the-art models on all the evaluation metrics.

4.4. Human study

Above experiments verify the effectiveness of our proposed model on the Visdial [5] task. In this section, to check whether our model can generate more human-like dialogs, we conduct a human study.

We randomly sample 1000 results from the test dataset in different length, generated by our final model, our baseline model **CoAtt-G-MLE**, and the Memory Network (**MN**)² [5] model. We then ask 3 human subjects to guess whether the last response in the dialog is human-generated or machine-generated and if at least 2 of them agree it is generated by a human, we say it passed the Turing Test. Table 3 summarizes the percentage of responses in the dialog that passes the Turing Test (M1), we can see our model outperforms both the baseline model and the MN model. We also apply our discriminator model in Sec. 3.2 on these 1000 samples and it recognizes that nearly 70% percent of them as human-generated responses (random guess is 50%), which suggests that our final generator successfully fool the discriminator in this adversarial learning. We additionally record the percentage of responses that are evaluated as better than or equal to human responses (M2), according to the human subjects’ manual evaluation. As shown in Table 3, 45% of the responses fall into this case.

5. Conclusion

Visual Dialog generation is an interesting topic that requires machine to understand visual content, natural language dialog and have the ability of multi-modal reasoning. More importantly, as a human-computer interaction interface for the further robotics and AI, apart from the correctness, the human-like level of the generated response is a significant index. In this paper, we have proposed an adversarial learning based approach to encourage the generator to generate more human-like dialogs. Technically, by combining a sequential co-attention generative model that can

²we use the author provided code and pre-trained model provided on <https://github.com/batra-mlp-lab/visdial>

jointly reason the image, dialog history and question, and a discriminator that can dynamically access to the attention memories, with an intermediate reward, our final proposed model achieves the state-of-art on VisDial dataset. A Turing Test fashion study also shows that our model can produce more human-like visual dialog responses.

References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2425–2433, 2015. [1](#), [2](#)
- [2] N. Asghar, P. Poupart, J. Xin, and H. Li. Online sequence-to-sequence reinforcement learning for open-domain conversational agents. *arXiv preprint arXiv:1612.03929*, 2016. [3](#)
- [3] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016. [2](#), [3](#)
- [4] B. Dai, D. Lin, R. Urtasun, and S. Fidler. Towards diverse and natural image descriptions via a conditional gan. *arXiv preprint arXiv:1703.06029*, 2017. [2](#), [4](#)
- [5] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017. [1](#), [2](#), [4](#), [6](#), [7](#), [8](#)
- [6] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *Proc. IEEE Int. Conf. Comp. Vis.*, 2017. [2](#)
- [7] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. [2](#), [4](#)
- [8] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. [3](#)
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 2672–2680, 2014. [2](#), [3](#)
- [10] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2016. [1](#)
- [11] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3128–3137, 2015. [1](#), [2](#)
- [12] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. [1](#)
- [13] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [7](#)
- [14] A. M. Lamb, A. G. A. P. GOYAL, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, pages 4601–4609, 2016. [6](#)
- [15] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016. [3](#)
- [16] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017. [3](#), [4](#), [5](#), [6](#)
- [17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [6](#)
- [18] J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. *arXiv preprint arXiv:1706.01554*, 2017. [2](#), [7](#), [8](#)
- [19] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 289–297, 2016. [1](#), [8](#)
- [20] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. P. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017. [2](#), [4](#)
- [21] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. [2](#), [3](#)
- [22] M. Ren, R. Kiros, and R. Zemel. Image Question Answering: A Visual Semantic Embedding Model and a New Dataset. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 1, page 5, 2015. [1](#), [2](#)
- [23] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Proc. Conf. Empirical Methods in Natural Language Processing*, pages 583–593. Association for Computational Linguistics, 2011. [3](#)
- [24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. [3](#)
- [25] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal. Visual reference resolution using attention memory for visual dialog. *arXiv preprint arXiv:1709.07992*, 2017. [8](#)
- [26] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784, 2016. [3](#)
- [27] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301, 2017. [7](#)
- [28] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele. Speaking the same language: Matching machine to human captions by adversarial training. *arXiv preprint arXiv:1703.10476*, 2017. [4](#)
- [29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [4](#)

- [30] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015. 3
- [31] P.-H. Su, M. Gasic, N. Mrksic, L. Rojas-Barahona, S. Ultes, D. Vandyke, T.-H. Wen, and S. Young. Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*, 2016. 3
- [32] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 3104–3112, 2014. 3
- [33] O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015. 3
- [34] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3156–3164, 2014. 1, 2
- [35] P. Wang, Q. Wu, C. Shen, and A. v. d. Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 4
- [36] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 4, 5
- [37] Q. Wu, C. Shen, A. v. d. Hengel, L. Liu, and A. Dick. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 203–212, 2016. 1, 2
- [38] Q. Wu, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4622–4630, 2016. 1, 2
- [39] Z. Xu, B. Liu, B. Wang, C. Sun, and X. Wang. Incorporating loose-structured knowledge into lstm with recall gate for conversation modeling. *arXiv preprint arXiv:1605.05110*, 2016. 3
- [40] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked Attention Networks for Image Question Answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 21–29, 2016. 1, 8
- [41] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *Proc. Eur. Conf. Comp. Vis.*, pages 69–85. Springer, 2016. 1
- [42] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proc. Conf. AAAI*, pages 2852–2858, 2017. 2, 3
- [43] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016. 3