

Building Task-Oriented Visual Dialog Systems Through Alternative Optimization Between Dialog Policy and Language Generation

Mingyang Zhou Josh Arnold Zhou Yu

Department of Computer Science

University of California, Davis

{minzhou, jarnold, joyu}@ucdavis.edu

Abstract

Reinforcement learning (RL) is an effective approach to learn an optimal dialog policy for task-oriented visual dialog systems. A common practice is to apply RL on a neural sequence-to-sequence (seq2seq) framework with the action space being the output vocabulary in the decoder. However, it is difficult to design a reward function that can achieve a balance between learning an effective policy and generating a natural dialog response. This paper proposes a novel framework that alternatively trains a RL policy for image guessing and a supervised seq2seq model to improve dialog generation quality. We evaluate our framework on the Guess-Which task and the framework achieves the state-of-the-art performance in both task completion and dialog quality.

1 Introduction

Visually-grounded conversational artificial intelligence (AI) is an important field that explores the extent intelligent systems are able to hold meaningful conversations regarding visual content. Visually-grounded conversational AI can be applied to a wide range of real-world tasks, including assisting blind people to navigate their surroundings, online recommendation systems, and analysing mass amounts of visual media through natural language. Current approaches to these tasks involve an end-to-end framework that maps the multi-modal context to a deep vector and in order to decode a natural dialog response. This framework can be trained through supervised learning (SL) with the objective to maximize the distribution of the response given a human-human dialog history. Given a large conversational data, the neural end-to-end system can effectively learn to generate coherent and natural language.

While much success has been achieved by applying neural sequence to sequence models to

open visual grounding conversation, the visual dialog system also needs to learn an optimal strategy to efficiently accomplish an external goal through natural conversations. To address this issue, various image guessing tasks such as Guess-Which (Chattopadhyay et al., 2017) and Guess-What (de Vries et al., 2016) are proposed to evaluate a visual-grounded conversational agent on its ability to retrieve visual content via conversing in natural language. To obtain an optimal dialog policy, reinforcement learning (RL) is introduced to enable the neural end-to-end framework to model a more effective action distribution by exploring different dialog strategies. With the application of RL, the visual dialog system can generate more consistent responses and achieve a higher level of engagement in the conversation when compared to a dialog system trained via SL (Das et al., 2017b; Zhang et al., 2017). A typical way to apply RL on a dialog system is to assign a task-related reward to influence the utterance generation process by treating each output word as the action step. A significant limitation of this approach is that it is difficult to achieve an optimal dialog policy that can both effectively complete the external goal and generate natural utterances (Zhao et al., 2019; Das et al., 2017b). As there is no ground truth reference during the RL training stage, the dialog system can only leverage the reward signal when generating the response. However, this approach often deviates from natural language as it is challenging to define a comprehensive reward that considers all aspects of the dialog quality, and in addition, assigns appropriate rewards to the large word vocabulary action space.

In this paper we propose a novel learning curriculum to address the challenge of joint learning between the dialog policy and language generation for task-oriented dialog systems. In our framework, we separate the training of the image re-

trieval policy from dialog generation by applying RL, with the goal of achieving an optimal policy for guessing the target image at every turn. In addition, we apply a language model objective function to optimize the utterance generator to mitigate language degeneration. We specifically study this framework in the image guessing task GuessWhich, where a conversational agent attempts to guess a target image by asking a series of questions. When compared to state-of-art RL visual dialog systems, our method achieves superior performance in both task-accomplishment and dialog quality.

2 Related Work

2.1 Visual Dialog System

Visual dialog systems are an emerging area of interdisciplinary research that attracts both the vision and language communities due to the potential applications. [Das et al. \(2017a\)](#) proposed a visual dialog task in which a conversational agent attempts to answer questions regarding an assigned image based on a dialog history. To approach this task, they initially collected data by having two people chat about an image with one person acting as the questioner and the other as the answerer. GuessWhich ([Chattopadhyay et al., 2017](#)) extends VisDial with the goal to build an agent that learns how to identify a target image through question and answers. ([de Vries et al., 2016](#)) additionally introduced a game in which a series of yes-or-no questions are asked by an agent in order to locate an object in an image. Many researchers approached these tasks via reinforcement learning (RL), with the goal of obtaining an optimal dialog policy. [Zhang et al. \(2017\)](#), for example, designed three rewards with respect to the goals of task achievement, efficiency, and question informativeness, in order to help the agent to achieve an effective question generation policy for GuessWhat game. [Das et al. \(2017b\)](#) applies reinforcement learning in the GuessWhich task and demonstrates a moderate improvement in accuracy compared to the supervised learning approach. Both methods apply RL on a neural end-to-end pipeline to jointly influence the language generation and dialog policy. Due the challenge of designing an appropriate reward for language generation, these methods generate responses that deviate from human natural language. [Zhang et al. \(2018\)](#), proposed an approach involving hierarchical reinforcement

learning and state-adaptation techniques that enable the agent to learn an optimal and efficient multi-modal policy. The bottleneck of ([Zhang et al., 2018](#))’s method, however, is that the system response is retrieved from a predefined human-written or system-generated utterance. The number of predefined responses are limited, therefore, this method does not easily generalize to other tasks in real-world settings. We address these limitations by applying RL on a reduced, yet more relevant action space, while optimizing the dialog generator in a supervised fashion. We alternatively optimize policy learning to language generation to combine the two tasks together.

2.2 RL on Task-oriented Dialog System

Various RL-based models have been proposed to train task-oriented dialog systems ([Williams and Young, 2007](#)). In order to build a traditional modular-based dialog system, researchers first identify the semantic representation, such as the dialog acts and slots in user utterances. Then they accumulate these semantic representations over time to track the dialog state. Finally they apply RL to learn an optimized dialog policy given the dialog state ([Raux et al., 2005](#); [Shi and Yu, 2018](#)). Such modular-based dialog systems are effective in narrow task domains, such as searching a bus route schedule and reserving a restaurant through natural language, but they fail to generalize to complex settings where the size of the action space increases. Owing to the development of deep learning, RL on neural sequence-to-sequence models has been explored in more complex dialog domains such as open-domain conversation ([Li et al., 2016](#)) and negotiation ([Lewis et al., 2017](#)). However, due to the difficulty of assigning appropriate rewards when operating in a large action space, these frameworks cannot generate fluent dialog utterances. [Zhao et al. \(2019\)](#) proposed a novel latent action RL framework to marry the advantage of a module based approach and sequence-to-sequence approach. They learned the optimal dialog policy in a complex task-oriented dialog domain while achieving decent conversation quality. We study the similar issue in a multi-modal task-oriented dialog scenario. We propose an iterative approach to optimize dialog policy using RL methods and system response generation via SL.

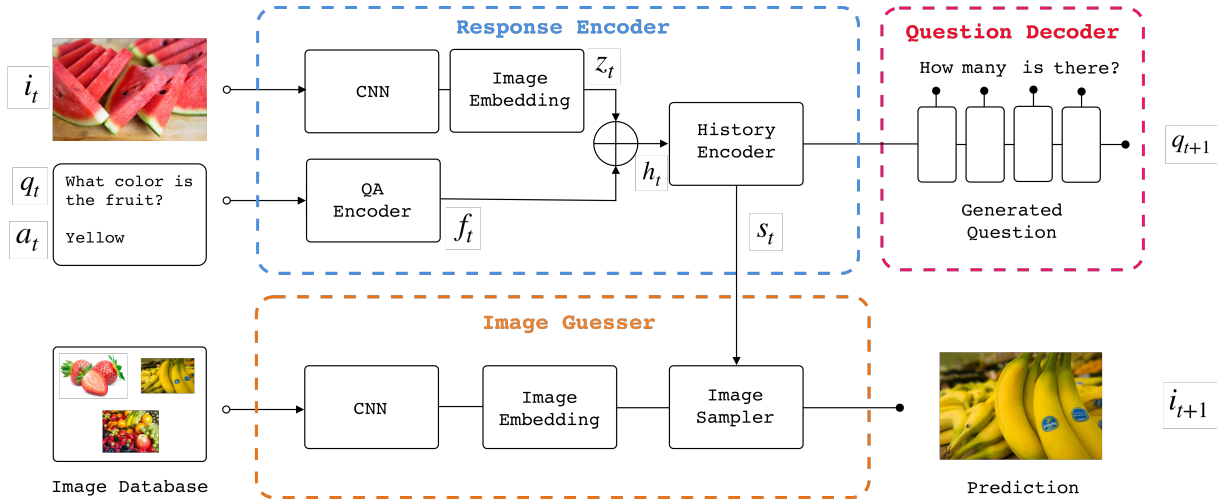


Figure 1: The proposed end-to-end framework of the conversation agent for GuessWhich task-oriented visual dialog task

3 Method

3.1 Problem Setting

In the GuessWhich problem, we aim to build an agent (Q-Bot) that attempts to guess an image i_{tgt} that another agent (A-Bot) knows by asking it a series of questions. At the beginning of the conversation, the Q-Bot is primed with a short caption c of the target image that is only known by A-Bot. At every round t , the Q-Bot generates a question q_t to elicit as much information as possible about the target image and the A-Bot provides an appropriate answer a_t with regard to q_t and the target image. In the end, the agent guesses the target image among a set of images considering the entire conversation.

In addition, our dialog system also guesses a candidate image i_t out of an image database $\mathcal{I} = \{i_k\}_{k=0}^m$ at every turn. This action models the process of sequentially updating the visual belief state on the target image based on the latest dialog history. Conditioned on the current guessed image and the prior dialog contexts, the system will generate an optimal question in order to get the maximum information from A-Bot that can strengthen the system’s belief on the target image. At the end of the conversation, our Q-Bot will guess the target image based on the multimodal contexts $s_n = (q_{1:n}, a_{1:n}, i_{1:n}, c)$ consisting of the dialog history and the trajectory of guessed images.

3.2 Model Architecture

Our Q-Bot is constructed on top of a hierarchical encoder-decoder framework (Serban et al., 2015),

which consists of three major components: The **Response Encoder**, the **Question Decoder**, and the **Image Guesser**. We introduce each component as follows:

Response Encoder The goal of the response encoder is to append the question q_t , the answer a_t , and the guessed image i_t received at current round to the dialog history and obtain an updated vector representation of the multimodal context s_t . The image i_t is encoded with a pre-trained convolutional neural network VGG-16 (Simonyan and Zisserman, 2015) followed by a linear embedding layer and the image feature vector denoted as z_t . For the question and answer pair at the current round (q_t, a_t) , we map them to a hidden state vector f_t through the LSTM based *QA Encoder*. We then apply a linear projection on the concatenation of f_t and z_t in order to obtain the multi-modal context vector h_t for the current round. The context vector is then passed through another LSTM encoder: *History Encoder* generates an updated dialog history representation $s_t = \text{HistoryEnc}(h_t, s_{t-1})$. We denote the trainable parameters for Response Encoder as θ_e .

Question Decoder The question decoder is a two-layer LSTM network initialized with the most updated dialog history representation vector s_t from the response encoder. It will sequentially sample the words to come up with the next question q_t . The learned parameters for question decoder are denoted as θ_d .

Image Guesser The Image Guesser attempts to identify the candidate image that best aligns with the dialog history. Given an image database $\mathcal{I} = \{i_k\}_{k=0}^m$ where we sample the candidate image, we first extract the image feature representations $\{z_k\}_{k=0}^m$ for all candidate images with the convolutional neural network and image embedding layer defined in response encoder. Then, we can sample a candidate image i_k for the current turn based on the euclidean distance $d(z_k, s_t)$ between the image feature of the candidate image and the current dialog history vector. The image with the smallest euclidean distance is selected as the guess i_t at the current round. The associated parameters for image guesser are defined as θ_g .

3.3 Training Dialog System

We follow a two-stage training fashion as introduced in many previous end-to-end RL dialog systems (Das et al., 2017b; Zhang et al., 2017; Zhao et al., 2019), where we first pre-train the dialog framework with a supervised objective then apply reinforcement learning to learn an optimal policy to retrieve the target image. The Supervised pre-training is a critical step that facilitates an effective policy exploration for RL training, as it is difficult to explore a complex action space with limited prior knowledge. During RL training, we introduce an alternative learning method between dialog policy exploration and natural utterance generation that addresses the issue of language degeneration in previous RL based visual dialog systems (Das et al., 2017b). We introduce each training method as follows.

3.3.1 Supervised Pre-training

During the supervised pre-training process, we jointly optimize the objective to generate questions and also predict target image features from dialog contexts. The task of question generation is optimized by maximizing the log conditional probability of the next question dependent on a ground truth dialog for every round of the conversation. For the image feature prediction, we minimize the mean square error (MSE) between the target image feature z_{tgt} and the dialog context vector s_t at each round. The joint loss function for

supervised pre-training is:

$$\mathcal{L}_{SL}(\theta_r, \theta_d, \theta_g) = \alpha \sum_{t=0}^n \log p(q_t | s_t) + \beta \sum_{t=0}^n \text{MSE}(z_{tgt}, s_t) \quad (1)$$

Where α and β are weights assigned to the objective function of each task in the joint objective function. With SL pre-training process, the dialog system is facilitated with the ability to estimate a visual object and emit a natural language sentence given a dialog context.

3.3.2 Reinforcement Learning on Image Retrieval

In our framework, we treat the sequence of image guess through out the conversation as a partially observable markov decision process and train a policy network via RL to obtain an optimal strategy to retrieve the target image. We formally describe state, policy, action, rewards, and the training procedures in our pipeline.

State The dialog states in our framework consist of a combination of multimodal contexts, including the image caption c , the dialog history with A-Bot $[q_1, a_2, \dots, q_t, a_t]$, and the image guessing trajectories $[i_1, i_2, \dots, i_t]$.

Policy The dialog policy $\pi_{\theta_r, \theta_g}(i_t | S_t)$ is a stochastic policy that samples the candidate image to guess from an image set based on the previous dialog histories. The policy is learned from response encoder and image generator which is parameterized via θ_r and θ_g .

Action The full action space is the number of images in the database that we can sample to guess an image. As the pre-trained process already enables the system to approximate a target image feature z_{tgt} with the dialog history representation vector s_t , we reduce the action space to the top-K nearest images, s_t , based upon the euclidean distance. The probability to sample an image i_j is gained by applying a softmax function over the top-K candidates on their distance to s_t : $\pi(j) = \frac{e^{-d_j}}{\sum_{k=1}^K e^{-d_k}}$. d_j represents the mean-square-distance between the j -th image and the dialog history state vector s_t .

Rewards We use the ranking percentile of the target image with respect to the dialog history vector s_t as the reward signal to credit the guess at each turn. The goal is to maximize the expectation value of the discounted return $\mathbb{E}[\sum_{t=1}^n \gamma^t r_t]$ over the n -round conversation. r_t is the ranking percentile of target image at round t and γ is the discounted factor between (0, 1).

Training Procedure Inspired from the RL training process on the iterative image retrieval framework (Guo et al., 2018), we apply the policy improvement theory (Sutton and Barto, 1998) to estimate an improved policy $\pi^*(s_t)$ from an existing policy $\pi(s_t)$ obtained from the pre-trained dialog system. Given a dialog state s_t and the action a_t derived from the existing policy, the value estimated by the current policy for taking the action a_t is $Q_\pi(s_t, a_t) = \mathbb{E}[\sum_{t'=t}^n \gamma^{t'} r_{t'}]$. To improve this, we explore a different action $a_t^* \neq a_t$ such that a larger policy value $Q_\pi(s_t, a_t^*) > Q_\pi(s_t, a_t)$ estimated with the current policy is achieved. Then we can adjust the existing policy $\pi(s_t)$ to a new policy $\pi^*(s_t)$ that executes that optimal action a_t^* given the current dialog state. The parameters of the policy can be effectively optimized via a cross entropy loss function conditioned on the derived optimal action a_t^* :

$$\mathcal{L}_{RL}(\theta_r, \theta_g) = \mathbb{E}\left[-\sum_{t=1}^n \log(\pi_{\theta_r, \theta_g}(a_t^* | s_t))\right] \quad (2)$$

Compared to the previous RL visual-grounded conversational agent, (Das et al., 2017b), there are several advantages of conducting policy learning on the action level of guessing the image. First, the action space of the top- k nearest neighbors are much smaller compared to the vocabulary size of the output words which reduces the difficulty to explore optimal strategies. Second, only the parameters of response encoder and image generator will be optimized during the RL training stage. The question decoder stays intact so that it is less likely for the dialog system to suffer from language deviation.

3.3.3 Alternating Policy Learning and Language Generation

Although the parameters of the decoder won't be impacted during the RL training stage, the shared response encoder of the dialog context is still optimized with policy learning. The language distribution captured by both the response encoder

and question decoder will gradually be differentiated from the original human dialog distribution. To prevent the potential language degeneration behavior, we alternatively optimize the dialog system with a policy learning objective in equation 2 and the language model objective function in equation 1 at every other epoch. It assures the dialog system maintains a good estimation of the human language distribution while also effectively exploring various dialog actions in order to achieve the task of guessing the right image.

4 Experiments

4.1 AI-AI Image Guessing Game

We evaluate the performance of our task-oriented dialog system by playing the image guessing game, GuessWhich with an automatic answer bot. Our conversational agent's goal is to locate the target image out of the 9,628 test images by interacting with the other player in five conversation exchanges. We evaluate agent on both goal achievement and utterance generation quality using two automatic evaluation metrics Percentile Mean Rank (PMR) and perplexity respectively. PMR estimates how good the agent can rank the target image against other candidates in the test database based on its current dialog state. Perplexity estimates the closeness of the generated response to a reference utterance given a dialog context from the VisDial dataset.

4.2 Human-AI Image Guessing Game

To evaluate the ability of our task-oriented dialog system in a realistic conversational scenario, we also make our agent play the image guessing game with human users. The games are set up as 20-image guessing games where the agent attempts to guess a target image outside of a pool of 20 candidate images by asking a human player 5 rounds of questions. The objective of the human player is to play the role of answer bot and answer agent's question with respect to the target image.

In this setting, the performance of the agent on task accomplishment is evaluated by the game win rates. The quality of the dialogs are manually rated on four criteria: fluency, comprehension, diversity and relevance. Fluency defines the naturalness and readability of the generated question in English. Comprehension represents the consistency of the generated question with respect to the previous dialog context. Diversity evaluates the uniqueness


Image	RL-Q-IG	SL-Q-IG	RL-Q
	Q: are bananas in bunch?	Q: what color are planes?	Q: is this outside?
	A: yes it's in bunch	A: no planes there	A: yes it is
	Q: are they on table?	Q: are they in city?	Q: is there any people?
	A: yes they are	A: yes it is	A: yes ...
	Q: is this in kitchen?	Q: are they in baskets?	Q: is this in zoo?
	A: I'm not sure	A: yes they are	A: no it is not
	Q: any people?	Q: any people?	Q: are there any people?
	A: Yes there are several people	A: yes there are	A: yes there are
	Q: any other fruits?	Q: animals?	Q: is there any people?
	A: NO, only bananas	A: no animals	A: yes ...

Table 1: A dialog example with the ground truth caption: **bunches of bananas hang on a wall and arranged for sale**. **blue** indicates ideal relevant questions and **orange** indicates less relevant questions.

of the questions generated within one game. Relevance presents how well the asked question is related to the target image and the given caption.

4.3 Comparative Models

We compare the performance of our model with state-of-the-art task-oriented visual dialog systems. Meanwhile we also perform an ablation study to evaluate the contribution of different designs in our framework. We introduce each model as follows:

SL-Q: The dialog agent from (Das et al., 2017b), which is trained with a joint supervised learning objective function for language generation and image prediction.

RL-Q: The dialog agent from (Das et al., 2017b) which is fine-tuned on a trained SL-Q by applying RL to the action space of output word vocabulary.

SL-Q-IG: The dialog agent from this framework is build on top of the SL-Q. Compared to SL-Q, SL-Q-IG has an additional image guesser module that makes a guess on target image at every round. SL-Q-IG also has an image encoder which fuses the guessed candidate image into the dialog history tracker. We only train this model with the supervised learning objective introduced equation 1.

RL-Q-IG: We use RL method to fine-tune SL-Q-IG. The RL method used is applied on action space of guessing candidate image. We alternate the model to optimize towards dialog policy learning and language generation.

RL-Q-IG-NA: We fine-tune SL-Q-IG by applying RL to the action space of guessing candidate image and only optimized with policy learning objective function alone.

RL-Q-IG-W: The dialog agent from our framework, which is fine-tuned on a trained SL-Q-IG by applying reinforcement learning on output word vocabulary. It follows the same training procedures as RL-Q to conduct policy learning.

All the SL dialog agents are trained on the VisDial Dataset with the default setting from (Das et al., 2017b) for 40 epochs. The RL dialog agents are then fine-tuned on their corresponding SL dialog agents for another 20 epochs. We evaluate every model on AI-AI image guessing games with the same answer bot, trained on the VisDial Dataset with the objective of visual question answering. We only evaluate RL-Q, SL-Q-IG and RL-Q-IG in human evaluation.

4.4 Human-AI Evaluation Implementation

In order to evaluate the effectiveness of the model, we designed three human evaluation tasks. Six college students were recruited to conduct the evaluation. Each student evaluated 100 games using the ground truth captions and 30 games using human generated captions. An additional three evaluators each completed 30 rounds of the relevancy experiment.

Ground Truth Captions We generated 100 image guessing games that used the ground truth captions to ensure a consistent amount of information is supplied across all human evaluators. Each game consists of a randomly selected set of 20 images from the VisDial Dataset, with one image randomly chosen as the target. For each game, we test three different models, each twice, resulting in a total of 600 evaluated games from the 100 generated games. We keep the identity of the models anonymous to the evaluator.

During each game, the human evaluator is pre-

sented with a target image the agent is trying to guess. Five rounds of Q&A take place in which the agent asks a question to elicit information and the human evaluator responds with a relevant truthful answer. At the end of each game, the evaluator is asked to rate the conversation on four criteria: fluency, relevance, comprehension and diversity.

Human Generated Captions In order to distinguish SL-Q-IG and RL-Q-IG in a more natural setting, we generate an additional 30 games, similar to the previous human evaluation task, except when beginning the game, the evaluator is asked to provide the caption for the target image instead of using the ground truth.

Relevance Experiment We noticed that the human evaluators found rating dialogues on the relevance criteria challenging and nuanced. In order to reduce the difficulty of rating dialogues using the relevance criteria, we designed a separate experiment in which, using the conversations obtained from the previous 600 evaluated ground truth games, a human evaluator is presented with three complete conversations side by side at each round. The evaluator then selects the most relevant conversation out of the three that corresponds to the image caption. Each of the three conversations have the same caption, however, correspond to a different model, thus allowing for an effective comparison between the relevancy of each model.

5 Results

5.1 Results on AI-AI Image Guess Game

Image Retrieval It is clear from Fig 2 that our dialog system significantly outperforms the baseline models from (Das et al., 2017b) in terms of PMR on every round of the dialog. PMR estimates how good the agent can rank the target image against other candidates in the test database. The biggest improvement gap is observed between SL-Q-IG and SL-Q. In comparison to SL-Q, SL-Q-IG tracks the additional context from the previously guessed images which leads to a better estimation of the target image. RL-Q-IG has better performance compared to SL-Q-IG in terms of PMR. This suggests that fine-tuning dialog systems with RL can further improve the success of guessing the correct image. The best image retrieval result is achieved by RL-Q-IG-NA, as the objective function of RL-Q-IG-NA is based solely on policy learning without consideration for the dialog generation quality.

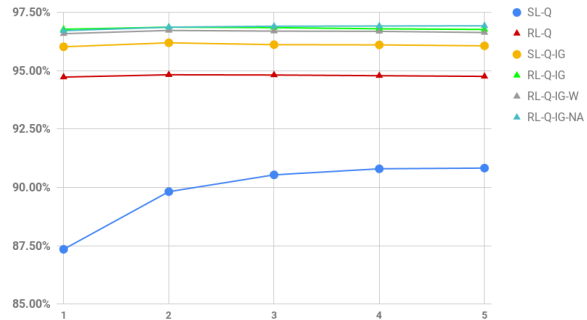


Figure 2: The percentile mean rank (PMR) over the 5-round dialog in the AI-AI image guessing game

Model	PMR	Perplexity
SL-Q	90.07%	79.49
SL-Q-IG	96.09%	61.42
RL-Q	94.78%	544.97
RL-Q-IG	96.81%	54.66
RL-Q-IG-NA	96.88%	363.88
RL-Q-IG-W	96.65%	227.35

Table 2: RL-Q-IG-NA performs best in PMR and RL-Q-IG perform best in perplexity

Although our framework achieved an improved image retrieval accuracy, we observed, however, that there is little improvement gained in PMR after additional rounds of conversation. We suspect this is partially due to the fact that images from MSCOCO are composed of a diverse selection objects and background scenes, thus making images easily distinguishable with a detailed caption. In cases where candidate images are visually similar or the given caption is not informative, additional rounds of dialog are necessary to identify the target image.

Language Generation We observe a marginal increase of perplexity from SL-Q to RL-Q in Table 2, thus demonstrating that there is a bottleneck when applying RL to improve the language generation. By decoupling the policy learning from the language generation and alternatively optimizing the dialog policy and language model, our RL-Q-IG avoids language deviation while still achieving an optimal dialog policy for the image retrieval task. To further evaluate the contribution from the RL and alternative training curriculum, we conduct two ablation studies. RL-Q-IG-NA is fine-tuned with a policy learning objective that excludes alternatively applying the language model loss. While RL-Q-IG-NA only achieves an in-

Model	Win	Fluency	Relevance	Comprehension	Diversity
RL-Q	59.6	4.19	3.22	2.60	2.50
SL-Q-IG	62.7	4.18	3.96	3.18	3.22
RL-Q-IG	67.5	4.40	4.02	3.50	3.25

Table 3: Evaluation results on the human-AI image guessing game initialized with ground truth captions

Model	Win	Fluency	Relevance	Comprehension	Diversity
RL-Q	29.2	4.04	2.88	2.71	2.29
SL-Q-IG	40.6	4.16	3.19	2.75	2.69
RL-Q-IG	67.6	4.23	3.74	3.32	3.06

Table 4: Evaluation results on the human-AI image guessing game initialized with human generated captions

cremental improvement over the full framework RL-Q-IG in terms of the PMR rate with less than 0.1%, it suffers from a dramatic increase of perplexity from 61.42 to 363.88, thus suggesting that alternatively applying the supervised learning objective can prevent the language model from deviating from the human language distribution. We additionally apply policy learning on the question decoder of SL-Q-IG and follow the RL fine-tuning process in (Das et al., 2017b) to train the agent, RL-Q-IG-W. While applying word-level RL enables RL-Q to achieve a moderate improvement over SL-Q in terms of PMR, we did not observe, the same degree of advantage in RL-Q-IG-W over SL-Q-IG. Additionally, RL-Q-IG-W is affected by a marginal increase in perplexity in comparison to the SL pre-trained agent, which approves the drawbacks of applying RL on a large action space in language generation.

5.2 Results on Human-AI Image Guess Game

The performance of a dialog agent evaluated with a user simulator does not necessarily reflect its performance on real human (de Vries et al., 2016). We conduct human evaluation on different dialog agents. From the results summarized in Table 3 and Table 4, we observe a consistent optimal performance of our method from conversations with AI agent to conversations with real human. Our RL-Q-IG significantly outperforms the baseline RL agent in all criteria for both settings. RL-Q-IG’s advantage over SL-Q-IG is not significant in the game when agents are primed with ground truth image caption. This observation correlates with the result in the Human-AI game, as both RL-Q-IG and SL-Q-IG achieve superior PMR over 96% when presented with the ground truth caption. However, if a human gen-

erated caption is given, the performance of the SL pre-trained agent suffers a big drop in all metrics except fluency while our RL agent maintains similar performance. Applying RL to fine-tune the dialog system enables the agent to generate more consistent dialogs in unseen scenarios. We also notice a degradation of the baseline RL agent from its performance with the user simulator, which suggests deviation from natural language is due to the sub-optimal RL training on a large action space.

We conduct a qualitative analysis on the generated dialogs from the three models with human players. Besides a marginal improvement over the RL baseline model and SL pretrained agent in terms of decreased repetition and grammar mistakes, there is a distinct superiority in regards to the relevance to the image caption in the questions generated from our RL agent. For example, in Table 9, we demonstrate the three dialogs generated by RL-Q-IG, SL-Q-IG and RL-Q on one game. Given the image caption *bunches of bananas hang on a wall and arranged for sale.*, RL-Q and SL-Q-IG ask very general questions that are not related to the caption such as “planes”, “zoo” and “animals”. In comparison, our agent asks high-quality questions regarding the caption that covers “bananas” and “fruits”. These questions help our RL agent obtain useful information to guess the target image. This advantage is also evident from the results of comparative evaluation on the degree of relevance of the questions in Table 5. We credit the positive result to the dialog policy, which explores multiple paths to conduct the conversation. The optimal path will involve a set of questions that obtains the maximum information of the target image such that it can construct the best estimation of the target image.

Model	Prefered (%)
RL-Q	8.93
SL-Q-ImGuess	39.90
RL-Q-IG	51.20

Table 5: Results on comparative evaluation of relevance on the human-AI image guessing dialogs

6 Conclusion and Future Work

We present a novel framework for building a task-oriented visual dialog system. We model the agent to simultaneously optimize two actions: guessing the image and generating effective questions. We achieve this simultaneous optimization through alternatively applying reinforcement learning to obtain an effective image guessing policy, whilst also applying supervised learning to enhance the quality of generated questions. By decoupling the policy learning from language generation, we overcome language degeneration in the word-level reinforcement learning framework. Both analytical and human evaluation suggests our proposed framework leads to a higher task completion rate and an improved dialog quality.

In the future, we plan to collect a fashion retrieval visual dialog dataset which simulates a realistic application for multi-modal dialog systems. To address the limitation of a high image retrieval rate with just the use of captions from the VisDial dataset, we plan to format a challenging candidate image pool in which images are visually similar to each other. This will incentivize the dialog system to conduct multiple rounds of dialog in order to retrieve the target image successfully. Furthermore, we will explore additional task-oriented settings where we can decouple task accomplishment from language generation to evaluate the extent our framework can generalize to other conversational tasks.

References

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. [Evaluating visual conversational agents via cooperative human-ai games](#). *CoRR*, abs/1708.05122.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, and Rogério Schmidt Feris. 2018. [Dialog-based interactive image retrieval](#). *CoRR*, abs/1805.00145.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. [Deal or no deal? end-to-end learning for negotiation dialogues](#). *CoRR*, abs/1706.05125.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. [Deep reinforcement learning for dialogue generation](#). *CoRR*, abs/1606.01541.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Lets go public! taking a spoken dialog system to the real world. In *in Proc. of Interspeech 2005*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. [Hierarchical neural network generative models for movie dialogues](#). *CoRR*, abs/1507.04808.

Weiyang Shi and Zhou Yu. 2018. [Sentiment adaptive end-to-end dialog systems](#). *CoRR*, abs/1804.10731.

K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Richard S. Sutton and Andrew G. Barto. 1998. *Introduction to Reinforcement Learning*, 1st edition. MIT Press, Cambridge, MA, USA.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2016. [Guesswhat?! visual object discovery through multi-modal dialogue](#). *CoRR*, abs/1611.08481.

Jason D. Williams and Steve Young. 2007. [Partially observable markov decision processes for spoken dialog systems](#). *Comput. Speech Lang.*, 21(2):393–422.

Jiaping Zhang, Tiancheng Zhao, and Zhou Yu. 2018. [Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog](#). *CoRR*, abs/1805.03257.

Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. 2017. [Asking the difficult questions: Goal-oriented visual question generation via intermediate rewards](#). *CoRR*, abs/1711.07614.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. [Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models](#). *CoRR*, abs/1902.08858.

A Human Evaluation Interface

We demonstrate our web interface for our Human-AI guessing game in Fig 3. At the beginning of every game the human player is presented with a random sampled target image. For the trained Q-bot, 20 images are sampled as the candidate pool with the objective of retrieving the target image. At the end of the conversation, the Q-Bot’s guess is revealed to the human player followed by a re-

quest to evaluate the quality of the conversation in the four aspects introduced in section 4.4 on a scale from 1 to 5.

Our human relevancy evaluation interface is demonstrated in Fig 4. We present side-by-side questions generated in the same game from each of the three different models and ask the human evaluators to pick the conversation that is most semantically relevant to the target image caption.

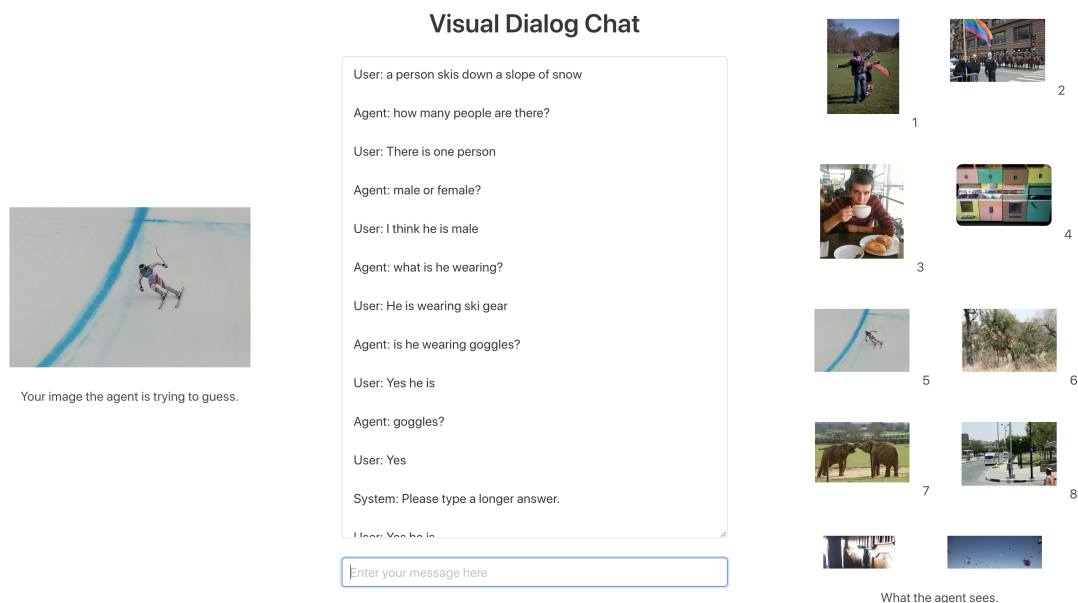


Figure 3: The web interface for human-AI guessing game. The left image is a target image randomly sampled from (Das et al., 2017a). The center section is a chat platform for human to communicate with a trained Q-Bot. On the right hand side are the 20 candidate images sampled for the Q-Bot to retrieve the target image.

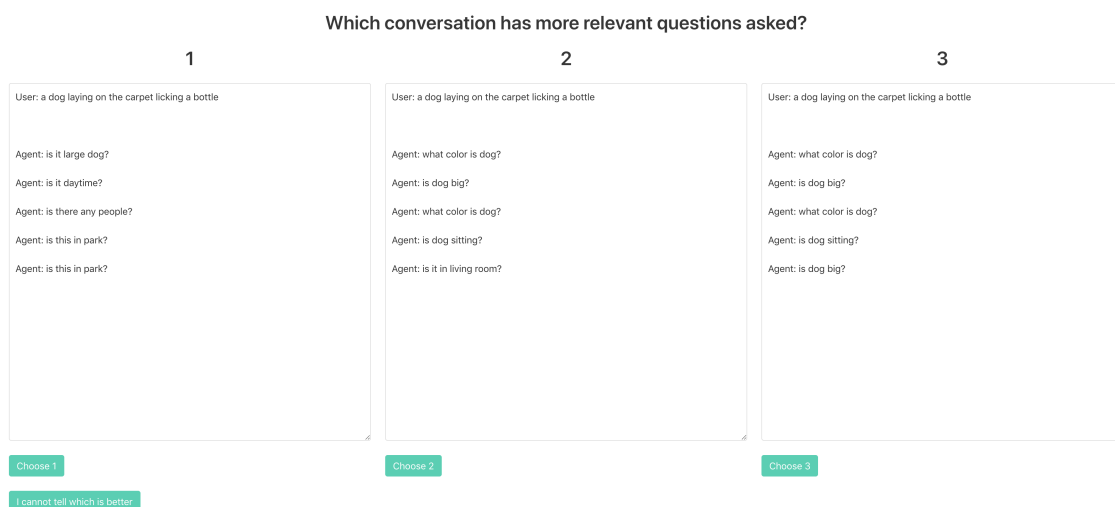


Figure 4: The user interface for human relevancy experiment. The conversation from the three different models on the same game setting are presented in the three boxes. On top of every conversation is the caption for the target image. Followed by that are the five questions asked by the Q-Bot in that game.

B Qualitative Examples


Image	RL-Q-ImGuess	SL-Q-ImGuess	RL-Q
	<p>Q: what color is vase? A: it is transparent</p> <p>Q: is it on table? A: yes it is</p> <p>Q: is it in kitchen? A: I do not know</p> <p>Q: any people? A: no, there is no people</p> <p>Q: what color is the table? A: yes there is</p>	<p>Q: what color is vase? A: it is brown</p> <p>Q: is it in vase? yes it is</p> <p>Q: is it in vase? A: yes it is</p> <p>Q: is it in vase? A: yes it is</p> <p>Q: is it in vase? A: yes it is</p> <p>Q: is it in vase? A: yes it is</p>	<p>Q: what color is table? A: it is orange</p> <p>Q: is there anything else on table? A: yes there are flowers</p> <p>Q: what color is table? A: it is orange</p> <p>Q: is there anything else on table? A: yes there are</p> <p>Q: is there anything else on table? A: yes there is</p>

Table 6: A dialog example with the human generated caption: **there is a plant in a vase and cookies**. Blue highlights diverse questions and orange indicates poor diversity.


Image	RL-Q-ImGuess	SL-Q-ImGuess	RL-Q
	<p>Q: are men old? A: No they are not</p> <p>Q: are they in uniform? A: I'm not sure</p> <p>Q: is it sunny? A: yes it is</p> <p>Q: are they on road? A: yes they are</p> <p>Q: are they in parking lot? A: No they're not</p>	<p>Q: how old are men? A: 30 years old</p> <p>Q: are they in city? A: yes they are</p> <p>Q: is it sunny? A: yes it is</p> <p>Q: any other people? A: no other people</p> <p>Q: animals? A: no other animals</p>	<p>Q: what color is hat? A: there is no hat</p> <p>Q: is it sunny? A: it seems yes</p> <p>Q: is this in city? A: it is in city</p> <p>Q: are there any people in picture? A: there are two people</p> <p>Q: is this in home or restaurant? A: it is outside</p>

Table 7: A dialog example with the human generated caption: **two men in formal wear standing next to a monster truck**. Blue highlights ideal relevant questions and orange indicates less relevant questions.


Image	RL-Q-ImGuess	SL-Q-ImGuess	RL-Q
	<p>Q: how many horses are there? A: there are no horses</p> <p>Q: are they in wild? A: no, they are in village</p> <p>Q: is there fence? A: yes, there is fence</p> <p>Q: any people? A: yes, a person</p> <p>Q: animals? A: yes, several bulls</p>	<p>Q: how many horses are there? A: There are no horses, only bulls</p> <p>Q: what color are horses? A: they are white and black</p> <p>Q: are they in field? A: Yes they are</p> <p>Q: is it sunny? A: Yes it is</p> <p>Q: is it sunny? A: Yes it is</p>	<p>Q: is this in park? A: maybe it is in park</p> <p>Q: are there any people in picture? A: yes there is one picture</p> <p>Q: is this in home or restaurant? A: it is in neither of those</p> <p>Q: are there any people in picture? A: yes there is one person</p> <p>Q: is this in restaurant? A: no it is not</p>

Table 8: A dialog example with the human generated caption: **an image of running with the bulls outside**. Blue indicates ideal relevant questions and orange highlights irrelevant/repeating questions.


Image	RL-Q-ImGuess	SL-Q-ImGuess	RL-Q
	<p>Q: what color is kite? A: It is orange</p> <p>Q: is it sunny? A: Yes it is</p> <p>Q: any other people? A: Yes there are two people</p> <p>Q: any animals? A: No there is not</p> <p>Q: any trees? A: Yes, there are several trees</p>	<p>Q: what color is horse? A: there is no horse</p> <p>Q: how old is man? A: he is young</p> <p>Q: is he wearing hat? A: no, he is not wearing hat</p> <p>Q: what color is horse? A: there is no horse</p> <p>Q: is it sunny? A: yes it is</p>	<p>Q: is there only 1 person? A: no there is not</p> <p>Q: is this in park? A: yes it is</p> <p>Q: are there any people in picture? A: yes there are</p> <p>Q: is there any other food on table? A: no food on the table</p> <p>Q: is there anything else on table? A: nothing . .</p>

Table 9: A dialog example with the human generated caption: **a man holding a kite while a girl tries to fly it**. Blue indicates ideal relevant questions and orange indicates poor relevance.