What's to know? Uncertainty as a Guide to Asking Goal-oriented Questions

Ehsan Abbasnejad, Qi Wu, Javen Shi, Anton van den Hengel
{ehsan.abbasnejad,qi.wu01,javen.shi,anton.vandenhengel}@adelaide.edu.au
Australian Institute of Machine Learning & The University of Adelaide, Australia

Abstract

One of the core challenges in Visual Dialogue problems is asking the question that will provide the most useful information towards achieving the required objective. Encouraging an agent to ask the right questions is difficult because we don't know a-priori what information the agent will need to achieve its task, and we don't have an explicit model of what it knows already. We propose a solution to this problem based on a Bayesian model of the uncertainty in the implicit model maintained by the visual dialogue agent, and in the function used to select an appropriate output. By selecting the question that minimises the predicted regret with respect to this implicit model the agent actively reduces ambiguity. The Bayesian model of uncertainty also enables a principled method for identifying when enough information has been acquired, and an action should be selected. We evaluate our approach on two goal-oriented dialogue datasets, one for visual-based collaboration task and the other for a negotiation-based task. Our uncertainty-aware information-seeking model outperforms its counterparts in these two challenging problems.

1. Introduction

One of the fundamental problems in any challenge that requires actively seeking the information required to carry out a task is that of identifying the information that will best enable the agent to achieve its objective. Identifying the information needed, and how to get it, is inherently complex, not least because the space of all possibly useful information is so large. We propose a solution to this problem here that is applicable to reinforcement learning in general, and that we demonstrate on the challenging problem of goal-oriented visual dialogue.

Goal-oriented visual dialogue requires the participants to engage in a natural language conversation towards a specified objective. The objectives of the two participants might be collaborative, such as communicating the identity of a specific object in an image [14], or they may be adversarial (see Sec. 4.2).

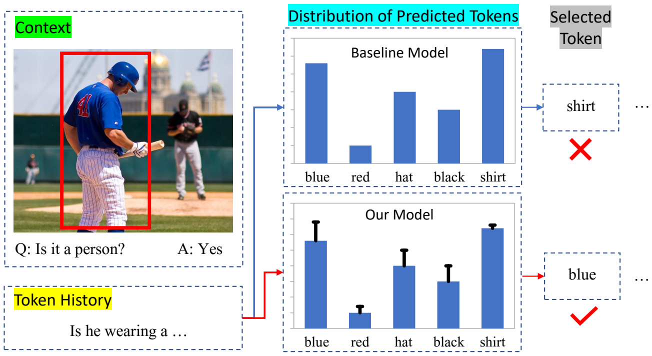

Figure 1. In GuessWhat [14] one player knows the correct object (here shown in a red box), and the other must ask questions to identify it. Traditionally an agent would generate questions by sequentially selecting words with the highest conditional probability, even though knowing the answer might be uninformative (in this case 'shirt' in the Baseline histogram). Our solution, however, selects 'blue', which corresponds to the highest sum of the probability and standard deviation (likely to be the most 'informative').

The primary technological challenge in goal-oriented visual dialogue is to devise natural language interactions that are directed towards achieving the required objective. This is in contrast to the more traditional approach that aims only to keep the other participant talking for as long as possible [22, 27, 37]. Note that the performance criteria in these two approaches are opposite, as in goal-directed visual dialogue success is indicated by achieving the shortest possible conversation.

Inspired by the success of deep learning in both computer vision and natural language processing (NLP), most recent goal-oriented dialogue studies rely on sequence-to-sequence (seq2seq) deep learning models [34, 37]. Obtaining the large datasets this approach requires is challenging, however. As a partial solution, a combination of seq2seq and deep reinforcement learning [35] are are commonly used to train a model (*i.e*. agent) with unlimited self-generated data in a self-play environment. Even if this was achieved ideally, however, it is unlikely that it would lead to an agent capable of carrying out the complex reasoning needed to devise the next interaction that will recover exactly the information required to achieve an as yet unspecified objective.

Ideally, rather than learning to generate questions solely by reinforcement learning, the method should calculate the question that, when answered, will provide the most useful information for achieving the agent's objective. The direct approach would require enumerating everything the agent might ever need to know, and the value of each such piece of information towards achieving its as yet unspecified objective. This would allow the identification of the missing piece of information that is most critical to achieving the agent's objective, and the formulation of a corresponding question.

This direct approach is infeasible because the agent has the capacity to store all of the information it might need to hold about the task, the intention of its counterpart, the image, and so on. Additionally, in the current state of the art approaches, this information is stored implicitly in the weights of a neural network. Defining the scope of such an information store is impossible, which makes measuring its information content infeasible. Explicitly relating the information stored to the agent's objective is similarly infeasible. This makes it impossible to directly identify the question that will provide the most useful information towards achieving the agent's objective.

Visual dialogue models trained using reinforcement learning already learn to estimate the value of a particular question as a step towards achieving their objective. This is represented in the model's value function. All that is required is a method for identifying the gaps in the model's internal information. We could then combine these information gaps with the learned policy to identify the most useful question.

Given that the models in question represent their internal information implicitly, a good approximation of the model's information gaps is available in the uncertainty of its internal state. By propagating the model's internal uncertainty through the question generation process we can thus identify questions that best reflect the model's ambiguity in achieving its objective. This is as compared to the current process that selects the question the model is most certain about (see Fig. 1).

We thus propose an *information-seeking decoder* (see Fig. 2) that chooses each word in a question based on its uncertainty about the environment and conditioned on the history of the conversations. We prove this leads to the minimum *expected regret*. An additional benefit of having an accessible estimate of a model's uncertainty is that it allows a more systematic identification of the point at which enough information has been gathered to make the required decision.

We evaluate our model primarily on the well-known collaborative goal oriented visual dialogue problem GuessWhat [14]. To demonstrate that it is equally applicable to (non-collaborative) negotiation tasks we also relate its performance on Deal or No Deal [21]. GuessWhat is a visual dialogue game between two agents in which they cooperate to identify one of many objects in an image. Deal or No Deal challenges two players to partition a collection of items such that each is assigned to one only player. In contrast to GuessWhat, this game is semi-cooperative in that one player can win more than their counterpart. Our approach significantly outperforms the baseline on both tasks. Our framework is summarised in Fig. 2.

Overall, our contributions are fourfold:

- We propose a Bayesian Deep Learning method for quantifying the uncertainty in the internal representation of a Reinforcement Learning model. This is significant as it provides a theoretically sound method for propagating uncertainty to the output space of the model.
- We describe an uncertainty-aware information-seeking decoder for goal-oriented conversation that actively formulates questions that will provide the information the agent needs to achieve its objective.
- We devise a method that exploits the confidence of the predictor as a measure to indicate if the model has enough information to produce an accurate output. We show this approach is effective and leads to fewer rounds of conversation for a goal to be achieved.
- We show that in both visual and textual dialogue challenges, whether cooperation or adversarial behaviour is desired, our approach outperforms the baselines. To the best of our knowledge, this is the first approach that works well across domains and tasks.

2. Related Work

Goal-oriented dialogue Dialogue generation [22, 23, 29, 3] has been studied for many years in the NLP literature, and has many applications. Dialogue generation is typically viewed as a Seq2Seq problem, or formulated as a statistical machine translation problem [26, 29, 2]. Recently, dialogue systems have been extended to the visual domain. For example, Das *et al*. [11] proposed a visual dialogue task that allows a machine to chat with a human about the content of a given image. Goal-oriented dialogue requires the agent understand a user request and complete a related task with a clear goal within a limited number of turns. Early goal-oriented dialogue systems [38, 41] model conversation as partially observable Markov Decision Processes (MDP) with many hand-crafted features for the state and action space representations, which restrict their usage to narrow domains. Bordes *et al*. [8] propose a goal-oriented dialogue

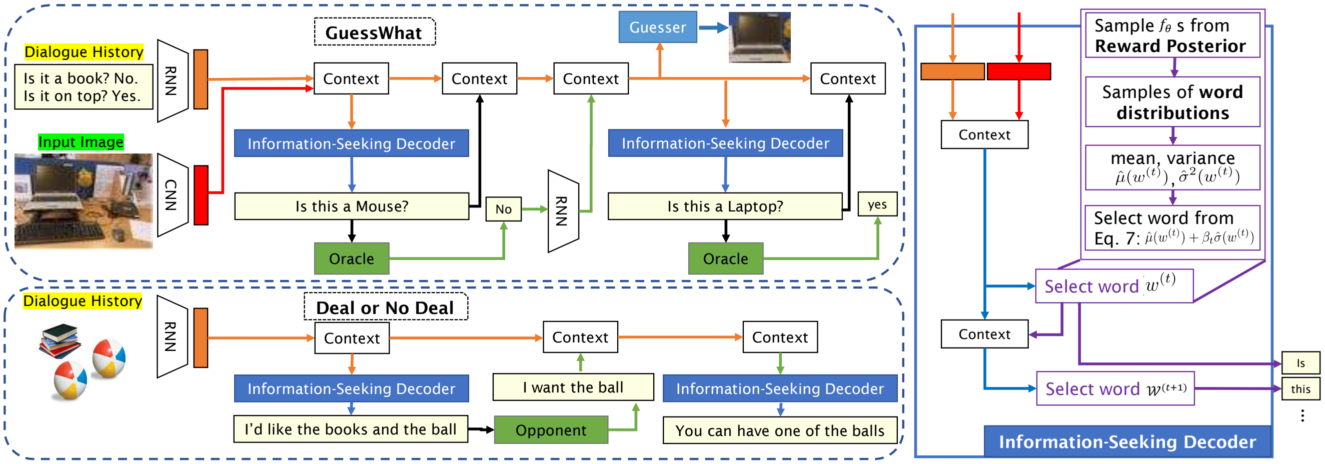

Figure 2. The framework in two applications in this paper: we develop a generic information-seeking decoder for dialogue systems. Our decoder selects each word optimistically with an upper bound on the reward to maximise the information obtained (details in Sec. 3.2). Samples from the reward posterior is taken by applying dropout to the context variable of the RNN which provably performs variational inference (see Eq. 1 and the Supplements for details).

test-bed that requires a user chat with a bot to book a table at a restaurant. In visual goal-oriented dialogue De Vries *et al*. [14] propose a guess-what game style dataset, where one person asks questions about an image to guess which object has been selected, and the second person answers questions as yes/no/NA.

RL in dialogue generation Reinforcement learning (RL) has been applied in many dialogue settings. Li *et al*. [22] simulate two virtual agents and hand-craft three rewards to train the response generation model. Recently, some works [6, 32] make an effort to integrate the Seq2Seq model and RL. RL has also been widely used to improve dialogue managers, which manage transitions between dialogue states [25, 28]. In visual dialogue, Das *et al*. [11] use reinforcement learning to improve cooperative bot-bot dialogues, and Wu *et al*. [40] combine reinforcement learning and generative adversarial networks (GANs) to generate more human-like visual dialogues. In [12], Das *et al*. introduce a reinforcement learning mechanism for visual dialogue generation. They establish two RL agents corresponding to question and answer generation respectively, to finally locate an unseen image from a set of images. The question agent predicts the feature representation of the image and the reward function is given by measuring how close the representation is compared to the true feature.

Uncertainty There are typically two sources of uncertainty to be considered: *Aleatoric* and *Epistemic* [19]. The former addresses the noise inherent in the observation while the later captures our ignorance about which model generated our data. Both sources can be captured with *Bayesian deep learning* approaches, where a prior distribution over the model weights is considered. However, performing Bayesian inference on a deep neural network is challenging and computationally expensive. One simple technique that has recently gained attention is to use Monte Carlo [1] dropout sampling which places a Bernoulli distribution over network weights [16, 17, 5, 4].

Most recently, Lipton *et al*. [24] proposed a Bayes-by-Backprop Q-network (BBQ-network) to approximate the Q-function and the uncertainty in its approximation. It encourages a dialogue agent to explore state-action regions in which the agent is relatively uncertain in its action selection. However, the BBQ-network only uses Thompson sampling to model the distribution of rewards for the words in a Bayesian manner and ignores the uncertainty in the estimators. This can lead to very uncertain decisions about confident actions or vice versa. Our method, on the other hand, models both the uncertainty in the actions (i.e. word choices), and the estimators, by directly incorporating the variance in the sampling procedure. We also provide a theoretical justification for the selection which is guaranteed to minimise regret.

3. Goal-oriented Dialogue Systems

We ground our goal-oriented dialogue problem as an interactive game between two agents for a collection of items. The items are either 1) multiple objects in an image for one agent to identify by asking the other questions, or 2) objects for the agents to split by negotiation. Conditioned on this game, once enough information is gathered, a *Guesser* takes the dialogue history and predicts the goal. The game is a success when the goal is achieved. The game between these two agents effectively simulates real natural language based conversation to achieve a particular goal, e.g. uncovering an unknown object, or an agreed split.

Each game is defined as a tuple (I, D, O, o^*), where I is the observed collection, D is the dialogue with T_{dialogue} rounds of conversation pairs $(\mathcal{W}_j, \mathcal{W}'_j)_{j=1}^{T_{\text{dialogue}}}$ and $\mathcal{W}_j =$

$(w^{(t)})_{t=1}^{M_j}$ is a sequence of M_j tokens $w^{(t)}$ with a a predefined vocabulary V, and $\mathcal{W}'_j$ is the response. $O = (o_n)_{n=1}^{N_o}$ is the list of objects, where N_o is the number of candidate objects in the collection. o^* is the target or a list of targets. In the GuessWhat game [14], o^* is a target object that the dialogue refers to. In the Deal or No Deal [21], it is a list of target objects that the negotiator agent is interested in.

Given an input collection I, an initial statement $\mathcal{W}_1$ is generated by sampling from the model until the stop token is encountered. Then the counterpart agent receives the statement $\mathcal{W}_1$, and generates the answer $\mathcal{W}'_1$, the pair $(\mathcal{W}_1, \mathcal{W}'_1)$ is appended to the dialogue history. We repeat this loop until the end of dialogue token is sampled, or the number of questions reaches the maximum. Finally, the *Guesser* takes the whole dialogue D and the object list O as inputs to predict the goal. We consider the goal reached if o^* is selected.

3.1. RL for Dialogue Generation

We model dialogue generation as a Markov Decision Process (MDP) to be solved by using a reinforcement learning (RL) agent [35]. The agent interacts with the environment over a sequence of discrete steps in which we have the dialogue generated based on the collection I at time step t in round T, the state of agent with the history of conversation pairs and the tokens of current question generated so far: $S_t = \big(I, (\mathcal{W}_j, \mathcal{W}'_j)_{j=1}^{T-1}, (w_T^{(t)})_{t=1}^{m}\big)$, where $t = \sum_{k=1}^{T-1} M_k + m$. The action of agent is to choose the subsequent token $w_T^{(t+1)}$ from the vocabulary V (we drop T for brevity). Depending on the action the agent takes, the transition between two states falls into one of the following:

1) w^{t+1} = end of statement: The current statement is finished, it is the other agent's turn.

2) w^{t+1} = end of dialogue: The dialogue is finished, the *Guesser* selects the output from list O.

3) Otherwise, the newly generated token w^{t+1} is appended to the current statement, the next state $S_{t+1} = \big(I, (\mathcal{W}_j, \mathcal{W}'_j)_{j=1}^{T-1}, (w_T^{(t)})_{T=1}^{m+1}\big)$.

The maximum length of a statement $\mathcal{W}_j$ is M_{max}, and the maximum number of rounds in a dialogue is T_{dialogue}. Therefore, the number of time steps t of any dialogue are $t \leq M_{max} * T_{\text{dialogue}}$. We use the stochastic policy $\pi_{\boldsymbol{\theta}}(w|S)$, where $\boldsymbol{\theta}$ represents the parameters of the deep neural network that produces the probability distributions for each state. The goal of the policy learning is to estimate the parameter $\boldsymbol{\theta}$. At the end of the dialogue, a decision about the unknown goal is made for which a reward is given by the environment. RL seeks to maximise the expected reward.

After a complete dialogue is generated, we update the RL agent's parameters based on the outcome of the dialogue. Let $r_{w^{(t)}}$ be the reward for achieving the goal after completing the dialogue, γ be a discount factor, and b be a bias function estimating the running average of the completed dialogue rewards so far[1]. Let future reward R for an action $w^{(t)}$ be $R(w^{(t)}) = \mathbb{E}\big[\sum_{i=0}^{\infty} \gamma^i (r_{w^{t+i}} - b(w^{t+i}))\big]$ where expectation is with respect to the policy π. The parameters of this model comprising of the policy and the bias function are then optimised using gradient policy theorem [36] and REINFORCE [39]. The policy determines how a statement is made in a dialogue system. Note that at each step there is an estimation of the reward (which is never directly observed) for each word in the RL and the observable reward is only given to the complete dialogue. Upon receiving the reward for the complete dialogue the parameters are accordingly updated. Utilising this estimation of the reward at each stage and a particular choice of the word strategy, a sequence of words is generated.

In the subsequent section we discuss a particular strategy that utilises the uncertainty in the policy (model) and seeks to provide a better approach for exploration of the space of possible dialogues. Moreover, since REINFORCE is a Monte Carlo estimate that is known to have a high variance, there is an additional source of uncertainty in evaluation of the expected rewards. As such, it is essential to consider uncertainty in policies manifesting in word choices.

3.2. Information-seeking Decoder

The decoder's objective is, given the dialogue thus far, to choose the subsequent word such that the resulting response is most "informative". To that end, we assume there is an underlying reward for each word $r_{w^{(t)}}$ at step t that we seek to uncover by exploring the space of actions (tokens in the vocabulary). A common practice is to model this value as the output of a *deterministic* function $f_{\boldsymbol{\theta}}(w^{(t)}) : V \rightarrow \mathbb{R}$ parameterised by $\boldsymbol{\theta}$ such as a neural network for sequential problems (e.g. LSTMs [18] or GRUs [9]). To select the subsequent action using this function one can greedily select the action with highest value or sample from a softmax (categorical distribution) built from this function.

However, this approach does not account for the uncertainty in the prediction of the reward $r_{w^{(t)}}$. This uncertainty has two main sources, (1) model uncertainty which is due to the imperfections in the parameters and (2) prediction uncertainty which is due to the lack of information about each action and its consequence. We choose a prior for the parameters and update them with the likelihood of the dialogue observations to obtain the posterior distribution in a Bayesian manner. The posterior at round T is:

$$p(\boldsymbol{\theta}|\mathcal{R}_t, D_t, I) = \frac{1}{Z}\prod_t p(w^{(t)}|(\mathcal{W}_j, \mathcal{W}'_j)_{j=1}^{T-1}, \mathcal{R}_t, I, f_{\boldsymbol{\theta}})p(\boldsymbol{\theta}) \tag{1}$$

where Z is the normaliser and $p(\boldsymbol{\theta})$ is a prior for the parameters. Here, $\mathcal{R}_t = r_{w^{(1)}}, \ldots, r_{w^{(t)}}$ is the set of rewards collected up to step t in the dialogue. This formulation has a

[1] This bias function reduces the variance of the estimator.

self-regularising behaviour that, unlike likelihood maximisation, is less susceptible to a local optima and performs better in practice. The predictive distribution of rewards from which each word is chosen becomes[2]:

$$
\begin{aligned}
&p(r_{w^{(t+1)}}|w^{(t+1)},\mathcal{R}_t,D_t,I)\\
&\quad= \int p(r_{w^{(t+1)}}|w^{(t+1)},f_{\boldsymbol{\theta}})p(\boldsymbol{\theta}|\mathcal{R}_t,D_t,I)d\boldsymbol{\theta} \qquad (2)\\
&\quad\approx \frac{1}{N}\sum_{i=1}^{N} p(r_{w^{(t+1)}}|w^{(t+1)},f^{(i)}_{\boldsymbol{\theta}}),\ f^{(i)}_{\boldsymbol{\theta}} \sim p(\boldsymbol{\theta}|\mathcal{R}_t,D_t,I)
\end{aligned}
$$

where N is the number of samples for the Monte Carlo estimation of the integral and

$$p(r_{w^{(t+1)}}|w^{(t+1)},f_{\boldsymbol{\theta}}) = \text{softmax}(f_{\boldsymbol{\theta}}(w_1^{(t+1)}),\ldots,w_{|V|}^{(t+1)}) \qquad (3)$$

where $|V|$ is the size of the dictionary. However, the posterior $p(\boldsymbol{\theta}|\mathcal{R}_t,D_t,I)$ in Eq.1 does not have a closed-form solution. Thus, we resort to variational inference [15], the details of which is provided in the Supplements. In a nutshell, inspired by [16, 17] we show that the posterior is approximated by a particular mixture model which is equivalent to performing typical MAP with dropout regularisation for dialogue generation. Further, the Monte Carlo estimate in Eq.2 is efficiently computed by applying dropout N times in the RNN network (note we take N context variables in Fig. 2). Hence, the mean and variance of the rewards computed from the posterior are:

$$\hat{\mu}(w^{(t)}) = \frac{1}{N}\sum_{i=1}^{N} f^{(i)}_{\boldsymbol{\theta}}(w^{(t)}) \qquad (4)$$

$$\hat{\sigma}^2(w^{(t)}) = \frac{1}{N}\sum_{i=1}^{N}\left(f^{(i)}_{\boldsymbol{\theta}}(w^{(t)}) - \mu(w^{(t)})\right)^2 + \tau^{-1} \qquad (5)$$

where τ is the precision parameter. Using Chebyshev's inequality, we have:

$$p\left(\left|f_{\boldsymbol{\theta}}(w^{(t)}) - \hat{\mu}(w^{(t)})\right| < \beta_t\hat{\sigma}(w^{(t)})\right) \quad \geq 1 - \frac{1}{\beta_t^2} \qquad (6)$$

which means for $\beta_t > 0$, with high probability we have $\left|f_{\boldsymbol{\theta}}(w^{(t)}) - \hat{\mu}(w^{(t)})\right| < \beta_t\hat{\sigma}(w^{(t)})$ for a random function $f_{\boldsymbol{\theta}}(w^{(t)})$. Hence we have an *upper bound* on the random function $f_{\boldsymbol{\theta}}$ with high probability:

$$f_{\boldsymbol{\theta}}(w^{(t)}) \quad < \quad \hat{\mu}(w^{(t)}) + \beta_t\hat{\sigma}(w^{(t)}) \qquad (7)$$

Selecting an action (word) with this upper bound both accounts for the estimation of the high–reward values by $\hat{\mu}(w^{(t)})$ and the uncertainty in this estimation for the given word $\hat{\sigma}(w^{(t)})$. In this bound, β_t controls how much the uncertainty is taken into account for selecting a word. Furthermore, it is clear that with $\beta_t \rightarrow 0$ this upper bound approaches greedy selection. In the reinforcement learning context, this approach mediates the exploration-exploitation dilemma by changing β_t.

This upper-bound is inspired by the Upper Confidence Bound (UCB) which is popular in multi-armed bandit problems [10]. A similar upper bound for Gaussian processes was proposed in [30]. However, this bound for neural networks, in particular for dialogues systems, is novel.

Expected Regret and Information For a dialogue agent, a metric for evaluating performance is cumulative regret, that is the loss due to not knowing the best word to choose at a given time. Suppose the best action at round t is $w_*^{(t)}$ for our choice $w^{(t)}$, we incur instantaneous expected regret, $\rho_t = \mathbb{E}_{f_{\boldsymbol{\theta}}}\left[f_{\boldsymbol{\theta}}(w_*^{(t)}) - f_{\boldsymbol{\theta}}(w^{(t)})\right]$. The cumulative regret ρ'_T after T rounds is the sum of instantaneous expected regrets: $\rho'_T = \sum_{t=1}^{T}\rho_t$. Note that neither ρ_t nor ρ'_T are ever revealed during dialogues generation. Our expected regret at each round is bounded as

$$\rho_t < \mathbb{E}_{f_{\boldsymbol{\theta}}}\left[\hat{\mu}(w^{(t)}) + \beta_t\hat{\sigma}(w^{(t)}) - f_{\boldsymbol{\theta}}(w^{(t)})\right] < 2\beta_t\hat{\sigma}(w^{(t)}) \qquad (8)$$

where the first inequality is due to $f_{\boldsymbol{\theta}}(w_*^{(t)}) < \hat{\mu}(w^{(t)}) + \beta_t\hat{\sigma}(w^{(t)})$ and the second one is because $\left|f_{\boldsymbol{\theta}}(w^{(t)}) - \hat{\mu}(w^{(t)})\right| < \beta_t\hat{\sigma}(w^{(t)})$, then $-f_{\boldsymbol{\theta}}(w^{(t)}) < -\hat{\mu}(w^{(t)}) + \beta_t\hat{\sigma}(w^{(t)})$. Therefore, we have

$$\rho'_T \quad < \quad 2\sum_t \beta_t\hat{\sigma}(w^{(t)}) \qquad (9)$$

As such, the expected regret for each word selected is bounded by the standard deviation of the predicted reward. When we choose words with high standard deviation, we actively seek to gain more information about the uncertain words to effectively reduce our expected regret.

Further, let's assume the predictive distribution is near Gaussian with mean and variance $\hat{\mu}(w^{(t)}), \hat{\sigma}^2(w^{(t)})$ (which considering the central limit theorem is natural). The entropy is then $\frac{1}{2}\log(2\pi e\hat{\sigma}^2(w^{(t)}))$. Hence, selecting actions with higher uncertainty is also justified from an information-theoretic perspective as means of selecting informative words. In a dialogue system, when uttering a sentence with length T the information we can obtain is at most (using the union bound) $\frac{1}{2}\sum_t^T\log(2\pi e\hat{\sigma}^2(w^{(t)}))$. An alternative to using the approach in Eq. 7 is to choose the words with highest entropy (the most informative words). However, that is an extreme case that will lead the RL algorithm to continuously explore the dialogue space.

3.3. Stopping Dialogue

One of the key challenges in a goal-oriented dialogue system is to identify the point at which the agent has sufficient information to make the required decision. We specified above that the probability of the unknown goal given the dialogue thus far is $p(o_{t+1}|D_t)$. The uncertainty in

[2] An alternative view is that we model $f_{\boldsymbol{\theta}}$ as a stochastic function and choose words accounting for their uncertainty. $f_{\boldsymbol{\theta}}$ is fully realised by its parameters $\boldsymbol{\theta}$, hence we use the uncertainty in the functional and the parameters interchangeably.

Algorithm 1 Training information-seeking dialogues

```
for Each update do
    # Generate trajectories
    for $k = 1$ to $K$ do                                   ▷ Select $K$ of the objects/items
        Pick target objects $o_k^* \in O_k$
        Set $D_t$ to initial input collection
        for $j = 1$ to $T_{\text{dialogue}}$ do             ▷ Generate $(\mathcal{W}_j, \mathcal{W}_j')$ pairs
            while $w^{(t+1)}$ not <stop> do
                Sample $f_{\boldsymbol{\theta}}^{(n)} \sim p(f_{\boldsymbol{\theta}}|\mathcal{W}_j, \mathcal{R}_t), n = 1, \ldots, N$
                Set $\hat{\mu}(w^{(t)}), \hat{\sigma}(w^{(t)})$ from $f_{\boldsymbol{\theta}}^{(n)}$
                $w^{(t+1)} = \arg\max_w \ \hat{\mu}(w) + \beta_t \hat{\sigma}^2(w)$
            end while
            $\mathcal{W}_j' = \text{SuperviseAgent}(\mathcal{W}_j, D_t)$
            if $<stop> \in \mathcal{W}_j$ or $H(o_{t+1}|D_t) \leq \eta$ then    ▷ Sec. 3.3
                delete $(\mathcal{W}_j, \mathcal{W}_j')$ and break;
            else
                append $w^{(t+1)}$ to $\mathcal{W}_j$
            end if
            append $(\mathcal{W}_j, \mathcal{W}_j')$ to $D_t$
        end for
        $o_k = \text{argmax}_o \ p(o|D_t)$                  ▷ Predict the goal
        reward $= \begin{cases} 1 & \text{If } o_k = o_k^* \\ 0 & \text{Otherwise} \end{cases}$
    end for
    Evaluate policy and update parameters $\boldsymbol{\theta}$
end for
```

this measure reflects the agent's confidence in its prediction, and thus provides a natural measure for the stopping criteria of the conversation. Intuitively, the agent stops when it feels confident in its prediction of the goal. Hence, we have $H(o_{t+1}|D_t) \leq \eta$ where H is the entropy and η is an appropriately chosen hyper-parameter for the confidence. When η is larger, we allow for less confident predictions leading to shorter dialogues. See Alg. 1 for the full algorithm.

4. Experiments

To evaluate the performance of the proposed approach we conducted experiments on two different goal-oriented dialogue tasks: GuessWhat [14] and Deal or No Deal [21]. Our approach outperforms the baseline in both cases. In both experiments we pre-train the networks using the supervised model and refine using reinforcement learning. To that end, we employ a two stage algorithm in which we learn to imitate the human dialogue behaviour in a supervised learning task and subsequently fine-tune for better generalisation and goal discovery using reinforcement learning. In both experiments, our decoder takes the history of the dialogue in addition to input collection (e.g. an image) and, guided by the uncertainty of each word, produces a question. Similar two-stage approaches are taken in [13, 14, 21]. Without using supervised learning first, the dialogue model may diverge from human language.

4.1. GuessWhat

In GuessWhat [14] a visually rich image with several objects is shown to two players. One player selects an object from the image. The task of the other player, the questioner, is to locate the unknown object by asking a series of yes/no questions. After enough information is gathered by the questioner, it then guesses what the selected object was. If the questioner guesses the correct object the game is successfully concluded. It is desirable for the questioner to guess the correct answer in as few rounds of questioning as possible. The dataset includes $155,281$ dialogues of $821,955$ pairs of question/answers with vocabulary size $11,465$ on $66,537$ unique images and $134,074$ objects.

	Sampling	Greedy	Beam Search	Avg. Ques
New Object				
Supervised [14]	41.6	43.5	47.1	5
RL [31]	58.5	60.3	60.2	5
TPG [43]	62.6	-	-	5
Ours	**61.4**	**62.1**	**63.6**	5
Ours ($\eta = 0.05$)	58.5	59.5	59.6	**4.2**
Ours ($\eta = 0.01$)	59.8	59.3	60.4	4.5
Ours+MN	**68.3**	**69.2**	-	5
New Image				
Supervised [14]	39.2	40.8	44.6	5
RL [31]	56.5	58.4	58.4	5
Ours	**59.0**	**59.82**	**60.6**	5
Ours ($\eta = 0.05$)	56.7	56.5	57.3	**4.3**
Ours ($\eta = 0.01$)	58.0	57.5	58.5	4.5
Ours+MN	**66.3**	**67.1**	-	5

Table 1. Accuracy in identifying the goal object in the GuessWhat dataset (higher is better). The numbers in parentheses is the threshold used in questions for guessing the object in the image. Average number of questions is shown at the last column (lower is better).

Implementation Details We follow the same experimental setup as [14] in which three main components are built: a yes/no answering agent, a guesser and a questioner. The questioner is a recurrent neural network (RNN) that produces a sequence of state vectors for a given input sequence by applying long-short term memory (LSTM) as a transition function. The output of this LSTM network is the internal estimate of the reward with size 1024. To obtain a distribution over tokens, a softmax is applied to this output.

The samples of the reward estimate in the questioner are taken utilising dropout with parameter 0.5. Subsequently, the upper bound in Eq. 7 is calculated to choose words. For this experiment we set $\beta_t = 1$[3].

Once the questioner is trained using our information seeking decoder in RL, we take three approaches to evaluating the performance of the questioner: (1) *sampling* where the subsequent word is sampled from the multinomial distribution in the vocabulary, (2) *greedy* where the word with maximum probability is selected and (3) *beam search* keeping the K-most promising candidate sequences at each time step (we choose $K = 20$ in all experiments). During training the baseline uses the greedy approach to select the se-

[3] We observed marginal performance improvement by using a larger β on Guesswhat, despite the additional training overhead.

is it a person?	Yes
is he wearing a brown coat?	No
is he wearing a white shirt?	No
is he wearing a blue shirt?	No
are they sitting down?	No
is it the guy in the orange shirt to the left?	Yes

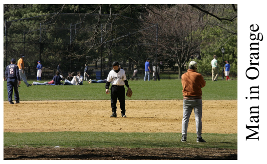

Figure 3. Sample dialogue from the GuessWhat dataset. The agent asks about a brown coat and then changes it to orange in anticipation of wrong identification.

quence of words as in [14].

Overall Results We compare two cases, labelled *New Object* and *New Image*. In the former the object sought is new, but the image has been seen previously. In the latter the image is also previously unseen. We report the prediction accuracy for the guessed objects. It is clear that the accuracies are generally higher for the new objects as they are obtained from the already seen images.

The results are summarised in Tab. 1. As shown, simply applying REINFORCE improves the output of the system significantly, in particular in the new image case where the generalisation is tested. This improvement is because the question generator has the chance to better explore possible questions. Additionally, the greedy approach outperforms others in the RL baseline in [31]. This illustrates that the distribution of the words obtained from the softmax in the question generator is not very peaked and the difference between the best and second best word is often small. This indicates that the prediction at test time is very uncertain and supports our approach.

Since our approach seeks uncertain words, those words are exploited at training time, which leads to lower variance (a more peaked distribution) and better performance of the greedy selection. Beam search significantly increases performance when we carry out 5 rounds (as in [14, 31]) of question-answering. This is because the most informative words are selected by our approach which, combined with the beam-search's mechanism for forward exploration, leads to better performance.

Note that our approach is generic enough that can be used in combination with other architectures (e.g. [20, 42]). For instance, in Tab. 1 "Ours+MN" uses the Memory Network [33] and Attention mechanism [7] in the Guesser (similar to that of [43]) which leads to better question generation. Fig. 3 shows one example produced by our dialogue generator. More examples can be found in the supplements.

Ablation Study on Early Stopping In goal-oriented dialogue systems, it is desirable to make a decision as soon as possible. In this experiment, we control the dialogue length by changing the threshold η (see Sec. 3.3 for more details). When η is larger, we accept less confident predictions leading to shorter dialogues. As shown in the Tab. 1, our models achieve a comparable performance to the baseline even using shorter rounds of question answering. The Fig. 4 shows the proportion of dialogues successful in identifying the goal object at each round. Our model achieves higher accuracy even in the earlier rounds, e.g. at the round three.

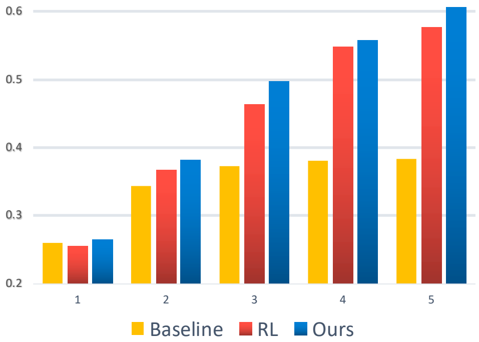

Figure 4. The proportion of dialogues successful in identifying the goal object at each round in GuessWhat.

Human Study To evaluate how well humans can guess the target object based on the questions generated by our models, we conduct a human study. Following [42], we show human subjects 50 images with generated question-answer pairs from our model, and let them guess the objects. We ask three human subjects to play on the same split and the game is recognised as successful if at least two of them give the right answer. In our experiment, the average performance of humans was 79% compared to 52% and 70% for the supervised [14] and RL [31] models. We are even better than a model proposed in [42] (76%), which has three complex hand-crafted rewards. These results indicate that our agent can provide more useful information that can benefit a human in achieving the final goal.

4.2. Deal or No Deal

Here two agents receive a collection of items, and are instructed to divide them so that each item is assigned to one agent. This problem is, unlike the GuessWhat game, semi-cooperative game in that the goals are adversarial. Each agent's goal is to maximise its own rewards which may be in direct contradiction with its opponents goals.

Each item has a different random non-negative value for each agent. These random values are constrained so that:

		Score	% Agreed	% Selection
Baseline	Supervised [21]	5.4 vs. 5.5	87.9	50.78 vs 49.23
	RL [21]	7.1 vs. 4.2	89.9	55.81 vs 44.19
	RL+Rollouts [21]	8.3 vs. 4.2	94.4	60.02 vs 39.98
Ours	$\beta_t = 1$	8.09 vs 4.08	92.02	77.13 vs 22.87
	$\beta_t = 10$	8.27 vs 4.23	94.79	88.56 vs 11.44
	$\beta_t = 1000$	8.21 vs 4.33	94.65	87.05 vs 12.95
	$\beta_t =$ 10+Rollouts	**8.58** vs 4.13	**95.75**	**93.62** vs 6.38

Table 2. Prioritising words with greater uncertainty leads to better performance in negotiations.'% Selection' represents the percentage of trials in which the final decision is made by each agent.

(1) the sum of values for all items for each agent is 10; (2) each item has a non-zero value for at least one agent; and (3) there are items with non-zero value for both agents. These constraints are to ensure both agents cannot receive a maximum score, and that no item is worthless to both agents. After 10 turns, agents are given the option to complete the negotiation with no agreement, which is worth 0 points to each. There are 3 item types (*books*, *hats*, *balls*) in the dataset and between 5 and 7 total items in the collection.

Implementation Details The supervised learning model comprises 4 recurrent neural networks implemented as GRUs. The agent's input goal is encoded as the hidden state of a GRU with size 64. The tokens are generated by sampling from the distribution of tokens. Simple maximum likelihood often leads to accepting an offer because it is more often than proposing a counter offer. To remedy this problem, similar to the previous GuessWhat experiment, we perform goal-oriented reinforcement learning to fine-tune the model. In addition, following [21] we experimented with rollouts. That is, considering the future expected reward in the subsequent dialogue, which is similar to the beam search in the previous experiment.

Results & Ablation Analysis Results are shown in Tab. 2. We report the average reward for each agent and the percentage of agreed upon negotiations. We see that our approach significantly outperforms the baseline RL. This is due to the information-seeking behaviour of our approach that leads to the agent learning to perform better negotiations and achieve agreements when the deals are acceptable.

We also evaluate the influence of β_t (in Eq. 7), which controls how much the uncertainty is taken into account in selecting a word, in turn controling the extent of exploration in dialogue generation. Increasing β_t leads to more exploration and more confidence in the actions at the expense of later convergence. From Tab. 2, we can see that a larger β_t leads to better performance. We also observed that if β_t is too high, say 1000, it diminishes performance as the agent continues exploring (by uttering risky statements that may lead to better understanding of the agent's counterpart at the cost of losing the deal) rather than exploiting its knowledge about the best word choices at each step of the negotiation.

Alice : book=(2, 0) hat=(2, 5) ball=(1, 0) Bob : book=(2, 2) hat=(2, 2) ball=(1, 2)	
Our Approach vs Baseline	Baseline vs Baseline
Alice: i would like the hats and the books. **Bob**: i need the hats and the books. **Alice**: you can have the ball if i can have the rest **Bob**: ok deal	**Alice**: i'd like the hats and the ball. **Bob**: you can have the ball , but i need the hats and the books . **Alice**: i need the hats and a book . **Bob**: no deal . i can give you the ball and both books
Alice: **10** , Bob: **2** points	Alice: **0** , Bob: **4** points
Alice : book=(1, 0) hat=(1, 7) ball=(3, 1) Bob : book=(1, 9) hat=(1, 1) ball=(3, 0)	
Bob: i would like the book and the hat . **Alice** : you can have the book if i can have the rest **Bob**: ok , deal	**Bob**: i want the book and 2 balls **Alice**: i need the hat and the balls **Bob**: i need the book and one ball **Alice**: how about i take the hat and 1 ball? **Bob**: sorry i cant make a deal without the book **Alice**: then we will need the hat and the book
Alice: **10**, Bob: **9** points	Alice: **0** (7)*, Bob: **0** points

Table 3. Samples from the negotiation experiments: Our approach is Alice and Bob is the baseline. * is the potential reward.

Tab. 3 shows examples of the negotiations generated using our model. The baseline model sometimes refuses an option that could lead to a desirable deal. She learns to be forceful in some cases, and consistent. This is because our model uncovers that by taking this risk, the counterpart may change his strategy. This is in part due to the fact that the supervised case the agent is willing to compromise quickly and our approach exploits that. This is achieved by repeating the same proposition by our model. Furthermore, our model learns to give her counterpart an option to give him a false sense of control over the negotiation, thus deceiving him. While she seems to have given-up in favour of the other's benefit, she enforces her choice and is consistent.

5. Conclusion

One of the primary limitations of current goal-directed dialogue systems is their limited ability to identify the information required to achieve their goal, and the steps required to obtain it. This limitation inherent in any reinforcement learning-based system that needs to learn to acquire the information required to achieve a goal. We have described a simple extension to reinforcement learning that overcomes this limitation, and enables an agent to select the action that is most likely to provide the information required to meet their objective. The selection process is simple, and controllable, and minimises the expected regret. It also enables a principled approach to identifying the appropriate point at which to stop seeking more information, and act.

The approach we propose is based on a principled Bayesian formulation of the uncertainty in both the internal state of the model, and the process used to select actions using this state information. We have demonstrated the performance of the approach when applied to generating goal-oriented dialogue, which is one of the more complex problems in its class due to the generality of the actions involved (natural language), and the need to adapt to the unknown intentions of the other participant. The proposed approach none the less outperforms the comparable benchmarks.

Acknowledgment: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research

References

[1] E. Abbasnejad, J. Domke, and S. Sanner. Loss-calibrated monte carlo action selection. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3447–3453. AAAI Press, 2015. 3

[2] E. Abbasnejad, S. Sanner, E. V. Bonilla, and P. Poupart. Learning community-based preferences via dirichlet process mixtures of gaussian processes. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 1213–1219, 2013. 2

[3] E. Abbasnejad, Q. Wu, I. Abbasnejad, J. Shi, and A. van den Hengel. An active information seeking model for goal-oriented vision-and-language tasks. *arXiv preprint arXiv:1812.06398*, 2018. 2

[4] M. E. Abbasnejad, A. Dick, Q. Sh i, and A. van den Hengel. Active learning from noisy tagged images. 2018. 3

[5] M. E. Abbasnejad, Q. Shi, I. Abbasnejad, A. van den Hengel, and A. R. Dick. Bayesian conditional generative adverserial networks. *CoRR*, abs/1706.05477, 2017. 3

[6] 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

[7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. 7

[8] A. Bordes, Y.-L. Boureau, and J. Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016. 2

[9] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*, 2014. 4

[10] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. 2008. 5

[11] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. 2017. 2, 3

[12] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*, 2017. 3

[13] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2970–2979, 2017. 6

[14] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 4, 6, 7

[15] C. W. Fox and S. J. Roberts. A tutorial on variational bayesian inference. *Artificial Intelligence Review*, 38(2):85–95, Aug 2012. 5

[16] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33th International Conference on International Conference on Machine Learning*, 2016. 3, 5

[17] Y. Gal and Z. Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, USA, 2016. Curran Associates Inc. 3, 5

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. 9:1735–80, 12 1997. 4

[19] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5574–5584. Curran Associates, Inc., 2017. 3

[20] S. Lee, Y. Heo, and B. Zhang. Answerer in questioner's mind for goal-oriented visual dialogue. *CoRR*, abs/1802.03881, 2018. 7

[21] M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. *ArXiv e-prints*, 2017. 2, 4, 6, 7, 8

[22] 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. 1, 2, 3

[23] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017. 2

[24] Z. Lipton, X. Li, J. Gao, L. Li, F. Ahmed, and L. Deng. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. *arXiv preprint arXiv:1711.05715*, 2017. 3

[25] O. Pietquin, M. Geist, S. Chandramohan, and H. Frezza-Buet. Sample-efficient batch reinforcement learning for dialogue management optimization.

ACM Transactions on Speech and Language Processing (TSLP), 7(3):7, 2011. 3

[26] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. pages 583–593. Association for Computational Linguistics, 2011. 2

[27] 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*, volume 16, pages 3776–3784, 2016. 1

[28] S. Singh, D. Litman, M. Kearns, and M. Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research*, 16:105–133, 2002. 3

[29] 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. 2

[30] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 2010. 5

[31] F. Strub, H. De Vries, J. Mary, B. Piot, A. Courville, and O. Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*, 2017. 6, 7

[32] 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

[33] S. Sukhbaatar, a. szlam, J. Weston, and R. Fergus. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc., 2015. 7

[34] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA, 2014. MIT Press. 1

[35] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 16:285–286, 1998. 1, 4

[36] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, Cambridge, MA, USA, 1999. MIT Press. 4

[37] O. Vinyals and Q. Le. A neural conversational model. 06 2015. 1

[38] Z. Wang and O. Lemon. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 423–432, 2013. 2

[39] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8, May 1992. 4

[40] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. *arXiv preprint arXiv:1711.07613*, 2017. 3

[41] S. Young, M. Gašić, B. Thomson, and J. D. Williams. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013. 2

[42] J. Zhang, Q. Wu, C. Shen, J. Zhang, J. Lu, and A. Van Den Hengel. Goal-oriented visual question generation via intermediate rewards. In *European Conference on Computer Vision*, pages 189–204. Springer, 2018. 7

[43] R. Zhao and V. Tresp. Improving goal-oriented visual dialog agents via advanced recurrent nets with tempered policy gradient. In *IJCAI*, 2018. 6, 7