

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Abhishek Das^{1*}, Satwik Kottur^{2*}, José M.F. Moura², Stefan Lee³, Dhruv Batra^{1,4}

¹Georgia Institute of Technology, ²Carnegie Mellon University, ³Virginia Tech, ⁴Facebook AI Research

visualdialog.org

Abstract

We introduce the first goal-driven training for visual question answering and dialog agents. Specifically, we pose a cooperative ‘image guessing’ game between two agents – Q-BOT and A-BOT– who communicate in natural language dialog so that Q-BOT can select an unseen image from a lineup of images. We use deep reinforcement learning (RL) to learn the policies of these agents end-to-end – from pixels to multi-agent multi-round dialog to game reward.

We demonstrate two experimental results.

First, as a ‘sanity check’ demonstration of pure RL (from scratch), we show results on a synthetic world, where the agents communicate in ungrounded vocabularies, i.e., symbols with no pre-specified meanings (X, Y, Z). We find that two bots invent their own communication protocol and start using certain symbols to ask/answer about certain visual attributes (shape/color/style). Thus, we demonstrate the emergence of grounded language and communication among ‘visual’ dialog agents with no human supervision.

Second, we conduct large-scale real-image experiments on the VisDial dataset [5], where we perform supervised pretraining with human-dialog data and show that the RL fine-tuned agents significantly outperform their supervised counterparts. Interestingly, the RL Q-BOT learns to ask questions that A-BOT is good at, ultimately resulting in more informative dialog and a better team. Further, pretraining with human-dialog data (in English) ensures human-interpretability and scope for pairing these agents with humans.

1. Introduction

The focus of this paper is visually-grounded conversational artificial intelligence (AI). Specifically, we would like to develop agents that can ‘see’ (i.e., understand the contents of an image) and ‘communicate’ that understanding in natural language (i.e., hold a dialog involving questions and answers about that image). We believe the next generation of intelligent systems will need to possess this ability to hold a dialog about visual content for a variety of applications: e.g., helping visually impaired users understand their sur-

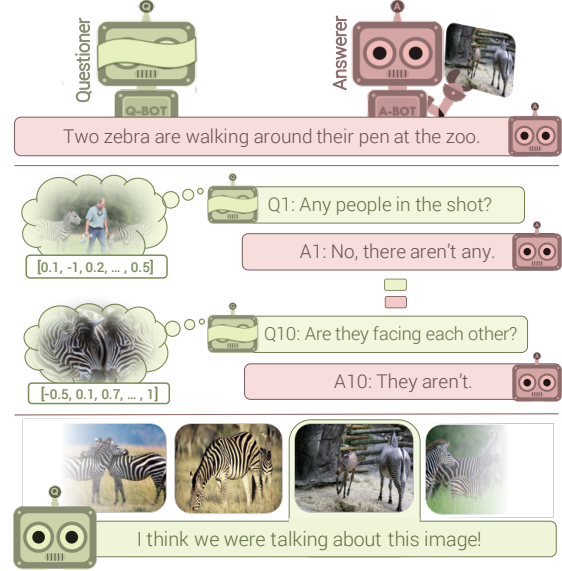


Figure 1: We propose a cooperative image guessing game between two agents – Q-BOT and A-BOT– who communicate through a natural language dialog so that Q-BOT can select a particular unseen image from a lineup. We model these agents as deep neural networks and train them end-to-end with reinforcement learning.

roundings [3] or social media content [40] (‘Who is in the photo? Dave. What is he doing?’), enabling analysts to sift through large quantities of surveillance data (‘Did anyone enter the vault in the last month? Yes, there are 103 recorded instances. Did any of them pick something up?’), and enabling users to interact naturally with intelligent assistants (either embodied as a robot or not) (‘Did I leave my phone on my desk? Yes, it’s here. Did I miss any calls?’).

Despite rapid progress at the intersection of vision and language, in particular, in image/video captioning [4, 14, 36–38, 41] and question answering [2, 25, 28, 34, 35], it is clear we are far from this grand goal of a visual dialog agent.

Two recent works [5, 6] have proposed studying this task of visually-grounded dialog. Perhaps somewhat counter-intuitively, both these works treat dialog as a *static* supervised learning problem, rather than the *interactive* agent learning problem that it naturally is. Specifically, both works [5, 6] first collect a dataset of human-human dialog, i.e., a sequence of question-answer pairs about an image $(q_1, a_1), \dots, (q_T, a_T)$. Next, a machine (a deep neu-

*The first two authors (AD, SK) contributed equally.

ral network) is provided with the image I , the human dialog recorded till round $t - 1$, $(q_1, a_1), \dots, (q_{t-1}, a_{t-1})$, the follow-up question q_t , and is supervised to generate the human response a_t . Essentially, at each round t , the machine is artificially ‘injected’ into the conversation between two humans and asked to answer the question q_t ; but the machine’s answer \hat{a}_t is thrown away, because at the next round $t + 1$, the machine is again provided with the ‘ground-truth’ human-human dialog that includes the human response a_t and not the machine response \hat{a}_t . Thus, the machine is *never allowed to steer the conversation* because that would take the dialog out of the dataset, making it non-evaluable.

In this paper, we generalize the task of Visual Dialog beyond the necessary first stage of supervised learning – by posing it as a cooperative ‘image guessing’ game between two dialog agents. We use deep reinforcement learning (RL) to learn the policies of these agents end-to-end – from pixels to multi-agent multi-round dialog to the game reward.

Our setup is illustrated in Fig. 1. We formulate a game between a questioner bot (Q-BOT) and an answerer bot (A-BOT). Q-BOT is shown a 1-sentence description (a caption) of an unseen image, and is allowed to communicate in natural language (discrete symbols) with the answering bot (A-BOT), who is shown the image. The objective of this fully-cooperative game is for Q-BOT to build a mental model of the unseen image purely from the natural language dialog, and then retrieve that image from a lineup of images.

Notice that this is a challenging game. Q-BOT must ground the words mentioned in the provided caption (*‘Two zebra are walking around their pen at the zoo.’*), estimate which images from the provided pool contain this content (there will typically be many such images since captions describe only the salient entities), and ask follow-up questions (*‘Any people in the shot? Are there clouds in the sky? Are they facing each other?’*) that help it identify the correct image.

Analogously, A-BOT must build a mental model of what Q-BOT understands, and answer questions (*‘No, there aren’t any. I can’t see the sky. They aren’t.’*) in a precise enough way to allow discrimination between similar images from a pool (that A-BOT does not have access to) while being concise enough to not confuse the imperfect Q-BOT.

At every round of dialog, Q-BOT listens to the answer provided by A-BOT, updates its beliefs, and makes a prediction about the visual representation of the unseen image (specifically, the fc7 vector of I), and receives a reward from the environment based on how close Q-BOT’s prediction is to the true fc7 representation of I . The goal of Q-BOT and A-BOT is to communicate to maximize this reward. One critical issue is that both the agents are imperfect and noisy – both ‘forget’ things in the past, sometimes repeat themselves, may not stay consistent in their responses, A-BOT does not have access to an external knowledge-base so it cannot answer all questions, *etc.* Thus, to succeed at the

task, they must learn to play to each other’s strengths.

An important question to ask is – why force the two agents to communicate in discrete symbols (English words) as opposed to continuous vectors? The reason is twofold. First, discrete symbols and natural language are interpretable. By forcing the two agents to communicate and understand natural language, we ensure that humans can not only inspect the conversation logs between two agents, but more importantly, communicate with them. After the two bots are trained, we can pair a human questioner with A-BOT to accomplish the goals of visual dialog (aiding visually/situationally impaired users), and pair a human answerer with Q-BOT to play a visual 20-questions game. The second reason to communicate in discrete symbols is to prevent cheating – if Q-BOT and A-BOT are allowed to exchange continuous vectors, then the trivial solution is for A-BOT to ignore Q-BOT’s question and directly convey the fc7 vector for I , allowing Q-BOT to make a perfect prediction. In essence, discrete natural language is an interpretable low-dimensional “bottleneck” layer between these two agents.

Contributions. We introduce a novel goal-driven training paradigm for visual question answering and dialog agents. Despite significant popular interest in VQA (>200 works citing [2] since 2015), all previous approaches have been based on supervised learning, making this the first instance of *goal-driven* training for VQA / visual dialog.

We demonstrate two experimental results.

First, as a ‘sanity check’ demonstration of pure RL (from scratch), we show results on a diagnostic task where perception is perfect – a synthetic world with ‘images’ containing a single object defined by three attributes (shape/color/style). In this synthetic world, for Q-BOT to identify an image, it must learn about these attributes. The two bots communicate via an ungrounded vocabulary, *i.e.*, symbols with no pre-specified human-interpretable meanings (‘X’, ‘Y’, ‘1’, ‘2’). When trained end-to-end with RL on this task, we find that the two bots *invent their own communication protocol* – Q-BOT starts using certain symbols to query for specific attributes (‘X’ for color), and A-BOT starts responding with specific symbols indicating the value of that attribute (‘1’ for red). Essentially, we demonstrate the *automatic emergence of grounded language and communication* among ‘visual’ dialog agents with no human supervision!

Second, we conduct large-scale real-image experiments on the VisDial dataset [5]. Imperfect perception on real images makes the discovery of human-interpretable language and communication strategy from scratch both difficult and an unnecessary re-invention of English. Thus, we pretrain with supervised learning (SL) on VisDial before fine-tuning with RL; this alleviates challenges in making RL converge to something meaningful. We show that these RL fine-tuned bots significantly outperform the supervised bots. Most interestingly, while the supervised Q-BOT attempts to mimic

how humans ask questions, the RL trained Q-BOT *shifts strategies* and asks questions that the A-BOT is better at answering, ultimately resulting in more informative dialog and a better team.

2. Related Work

Vision and Language. A number of problems at the intersection of vision and language have recently gained prominence, *e.g.*, image captioning [7, 9, 15, 38], and visual question answering (VQA) [2, 11, 24, 25, 28]. Most related to this paper are two recent works on visually-grounded dialog [5, 6]. Das *et al.* [5] proposed the task of Visual Dialog, collected the VisDial dataset by pairing two subjects on Amazon Mechanical Turk to chat about an image (with assigned roles of ‘Questioner’ and ‘Answerer’), and trained neural visual dialog answering models. Note that the task assigned to subjects in [5] – “ask questions to imagine the hidden image better” – is similar to our game’s goal. De Vries *et al.* [6] extended the Referit game [16] to a ‘Guess-What’ game, where one person asks questions about an image to guess which object has been ‘selected’, and the second person answers questions in ‘yes’/‘no’/NA (natural language answers are disallowed). One disadvantage of Guess-What is that it requires bounding box annotations for objects; our image guessing game does not need any such annotations and thus an unlimited number of game plays may be simulated. Moreover, as described in Sec. 1, both these works unnaturally treat dialog as a static supervised learning problem. Although both datasets contain thousands of human dialogs, they still only represent an incredibly sparse sample of the vast space of visually-grounded questions and answers. Training robust, visually-grounded dialog agents via supervised techniques is still a challenging task.

In our work, we take inspiration from the AlphaGo [31] approach of supervision from human-expert games and reinforcement learning from self-play. Similarly, we perform supervised pretraining on human dialog data and fine-tune in an end-to-end goal-driven manner with deep RL.

20 Questions and Lewis Signaling Game. Our proposed image-guessing game is naturally the visual analog of the popular 20-questions game. More formally, it is a generalization of the Lewis Signaling (LS) [20] game, widely studied in economics and game theory. LS is a cooperative game between two players – a *sender* and a *receiver*. In the classical setting, the world can be in a number of finite discrete states $\{1, 2, \dots, N\}$, which is known to the sender but not the receiver. The sender can send one of N discrete symbols/signals to the receiver, who upon receiving the signal must take one of N discrete actions. The game is perfectly cooperative, and one simple (though not unique) Nash Equilibrium is the ‘identity mapping’, where the sender encodes each world state with a bijective signal, and similarly the receiver has a bijective mapping from a signal to an action. Our proposed ‘image guessing’ game is a generalization of

LS with Q-BOT being the receiver and A-BOT the sender. However, in our proposed game, the receiver (Q-BOT) is not passive. It actively solicits information by asking questions. Moreover, the signaling process is not ‘single shot’, but proceeds over multiple rounds of conversation.

Text-only or Classical Dialog. Li *et al.* [21] have proposed using RL for training dialog systems. However, they hand-define what a ‘good’ utterance/dialog looks like (non-repetition, coherence, continuity, *etc.*). In contrast, taking a cue from adversarial learning [12, 22], we set up a cooperative game between two agents, such that we do not need to hand-define what a ‘good’ dialog looks like – a ‘good’ dialog is one that leads to a successful image-guessing play.

Emergence of Language. There is a long history of work on language emergence in multi-agent systems [27]. The more recent resurgence has focused on deep RL [1, 8, 10, 13, 18, 19, 23, 26]. The high-level ideas of these concurrent works are similar to our synthetic experiments. For our large-scale real-image results, we do not want our bots to invent their own uninterpretable language and use pretraining on VisDial [5] to achieve ‘alignment’ with English.

3. Cooperative Image Guessing Game:

In Full Generality and a Specific Instantiation

Players and Roles. The game involves two collaborative agents – a questioner bot (Q-BOT) and an answerer bot (A-BOT) – with an information asymmetry. A-BOT sees an image I , Q-BOT does not. Q-BOT is primed with a 1-sentence description c of the unseen image and asks ‘questions’ (sequence of discrete symbols over a vocabulary V), which A-BOT answers with another sequence of symbols. The communication occurs for a fixed number of rounds.

Game Objective in General. At each round, in addition to communicating, Q-BOT must provide a ‘description’ \hat{y} of the unknown image I based only on the dialog history and both players receive a reward from the environment inversely proportional to the error in this description under some metric $\ell(\hat{y}, y^{gt})$. We note that this is a general setting where the ‘description’ \hat{y} can take on varying levels of specificity – from image embeddings (*i.e.*, fc7 vectors of I) to textual descriptions to pixel-level image generations.

Specific Instantiation. In our experiments, we focus on the setting where Q-BOT is tasked with estimating a vector embedding of the image I . Given some feature extractor (*i.e.*, a pretrained CNN model, say VGG-16), no human annotation is required to produce the target ‘description’ \hat{y}^{gt} (simply forward-prop the image through the CNN). Reward/error can be measured by simple Euclidean distance, and any image may be used as the visual grounding for a dialog. Thus, an unlimited number of ‘game plays’ may be simulated.

4. Reinforcement Learning for Dialog Agents

In this section, we formalize the training of two visual dialog agents (Q-BOT and A-BOT) with Reinforcement Learn-

ing (RL) – describing formally the *action, state, environment, reward, policy*, and training procedure. We begin by noting that although there are two agents (Q-BOT, A-BOT), since the game is perfectly cooperative, we can without loss of generality view this as a single-agent setup where the single “meta-agent” is comprised of two “constituent agents” communicating via a natural language bottleneck layer.

Action. Both agents share a common action space consisting of all possible output sequences under a token vocabulary V . This action space is discrete and in principle, infinitely-large since arbitrary length sequences q_t, a_t may be produced and the dialog may go on forever. In our synthetic experiment, the two agents are given different vocabularies to coax a certain behavior to emerge (details in Sec. 5). In our VisDial experiments, the two agents share a common vocabulary of English tokens. In addition, at each round of the dialog t , Q-BOT also predicts \hat{y}_t , its current guess about the visual representation of the unseen image. This component of Q-BOT’s action space is continuous.

State. Since there is information asymmetry (A-BOT can see the image I , Q-BOT cannot), each agent has its own observed state. For a dialog grounded in image I with caption c , the state of Q-BOT at round t is the caption and dialog history so far $s_t^Q = [c, q_1, a_1, \dots, q_{t-1}, a_{t-1}]$, while the state of A-BOT also includes the image $s_t^A = [I, c, q_1, a_1, \dots, q_{t-1}, a_{t-1}, q_t]$.

Policy. We model Q-BOT and A-BOT as operating under stochastic policies $\pi_Q(q_t | s_t^Q; \theta_Q)$ and $\pi_A(a_t | s_t^A; \theta_A)$, such that questions and answers may be sampled from these policies conditioned on the dialog/state history. These policies will be learned by two separate deep neural networks parameterized by θ_Q and θ_A . In addition, Q-BOT includes a feature regression network $f(\cdot)$ that produces an image representation prediction *after listening to the answer at round t* , i.e., $\hat{y}_t = f(s_t^Q, q_t, a_t; \theta_f) = f(s_{t+1}^Q; \theta_f)$. Thus, the goal of policy learning is to estimate the parameters $\theta_Q, \theta_A, \theta_f$.

Environment and Reward. The environment consists of the other agent and the image I upon which the dialog is grounded. Since this is a purely cooperative setting, both agents receive the same reward. Let $\ell(\cdot, \cdot)$ be a distance metric on image representations (Euclidean distance in our experiments). At each round t , we define the reward as:

$$r_t \left(\underbrace{s_t^Q}_{\text{state}}, \underbrace{(q_t, a_t, \hat{y}_t)}_{\text{action}} \right) = \underbrace{\ell(\hat{y}_{t-1}, y^{gt})}_{\text{distance at } t-1} - \underbrace{\ell(\hat{y}_t, y^{gt})}_{\text{distance at } t} \quad (1)$$

i.e., the *change in distance* to the true representation before and after a round of dialog. In this way, we consider a question-answer pair to be low quality (i.e., have a negative reward) if it leads the questioner to make a *worse* estimate of the target image representation than if the dialog had ended.

Note that the total reward summed over all time steps of a dialog is a function of only the initial and final states due to the cancellation of intermediate terms, i.e.,

$$\sum_{t=1}^T r_t \left(s_t^Q, (q_t, a_t, \hat{y}_t) \right) = \underbrace{\ell(\hat{y}_0, y^{gt}) - \ell(\hat{y}_T, y^{gt})}_{\text{overall improvement due to dialog}} \quad (2)$$

This is again intuitive – ‘How much do the feature predictions of Q-BOT improve due to the dialog?’ The details of policy learning are given in Sec. 4.2, before which we describe the inner working of the two agents.

4.1. Policy Networks for Q-BOT and A-BOT

Fig. 2 shows an overview of our policy networks for Q-BOT and A-BOT and their interaction within a single round of dialog. Both the agent policies are modeled via Hierarchical Recurrent Encoder-Decoder neural networks, which have recently been proposed for dialog modeling [5, 29, 30].

Q-BOT consists of the following four components:

- **Fact Encoder:** Q-BOT asks a question q_t : ‘Are there any animals?’ and receives an answer a_t : ‘Yes, there are two elephants.’. Q-BOT treats this concatenated (q_t, a_t) -pair as a ‘fact’ it now knows about the unseen image. The fact encoder is an LSTM whose final hidden state $F_t^Q \in \mathbb{R}^{512}$ is used as an embedding of (q_t, a_t) .
- **State/History Encoder** is an LSTM that takes the encoded fact F_t^Q at each time step to produce an encoding of the prior dialog including time t as $S_t^Q \in \mathbb{R}^{512}$. Notice that this results in a two-level hierarchical encoding of the dialog $(q_t, a_t) \rightarrow F_t^Q$ and $(F_1^Q, \dots, F_t^Q) \rightarrow S_t^Q$.
- **Question Decoder** is an LSTM that takes the state/history encoding from the previous round S_{t-1}^Q and generates question q_t by sequentially sampling words.
- **Feature Regression Network** $f(\cdot)$ is a single fully-connected layer that produces an image representation prediction \hat{y}_t from the current encoded state $\hat{y}_t = f(S_t^Q)$.

Each of these components and their relation to each other are shown on the left side of Fig. 2. We collectively refer to the parameters of the three LSTM models as θ_Q and those of the feature regression network as θ_f .

A-BOT has a similar structure to Q-BOT with slight differences since it also models the image I via a CNN:

- **Question Encoder:** A-BOT receives a question q_t from Q-BOT and encodes it via an LSTM $Q_t^A \in \mathbb{R}^{512}$.
- **Fact Encoder:** Similar to Q-BOT, A-BOT also encodes the (q_t, a_t) -pairs via an LSTM to get $F_t^A \in \mathbb{R}^{512}$. The purpose of this encoder is for A-BOT to remember what it has already told Q-BOT and be able to understand references to entities already mentioned.
- **State/History Encoder** is an LSTM that takes as input at each round t – the encoded question Q_t^A , the image features from VGG [32] y , and the previous fact encoding F_{t-1}^A – to produce a state encoding, i.e. $((y, Q_1^A, F_0^A), \dots, (y, Q_t^A, F_{t-1}^A)) \rightarrow S_t^A$. This allows the model to contextualize the current question w.r.t. the history while looking at the image to seek an answer.

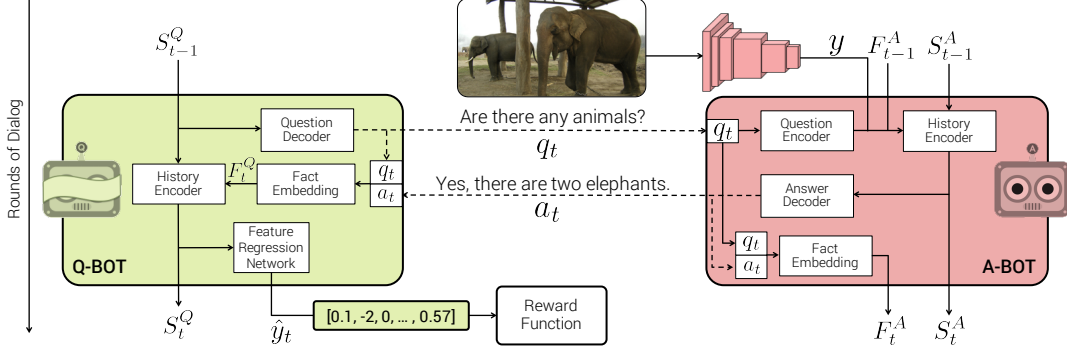


Figure 2: Policy networks for Q-BOT and A-BOT. At each round t of dialog, (1) Q-BOT generates a question q_t from its question decoder conditioned on its state encoding S_{t-1}^Q , (2) A-BOT encodes q_t , updates its state encoding S_t^A , and generates an answer a_t , (3) both encode the completed exchange as F_t^Q and F_t^A , and (4) Q-BOT updates its state to S_t^Q , predicts an image representation \hat{y}_t , and receives a reward.

- **Answer Decoder** is an LSTM that takes the state encoding S_t^A and generates a_t by sequentially sampling words.

Our code and models are publicly available at github.com/batra-mlp-lab/visdial-rl.

To recap, a dialog round at time t consists of 1) Q-BOT generating a question q_t conditioned on its state encoding S_{t-1}^Q , 2) A-BOT encoding q_t , updating its state encoding S_t^A , and generating an answer a_t , 3) Q-BOT and A-BOT both encoding the completed exchange as F_t^Q and F_t^A , and 4) Q-BOT updating its state to S_t^Q based on F_t^Q and making an image representation prediction \hat{y}_t for the unseen image.

4.2. Joint Training with Policy Gradients

In order to train these agents, we use the REINFORCE [39] algorithm that updates policy parameters $(\theta_Q, \theta_A, \theta_f)$ in response to experienced rewards. In this section, we derive the expressions for the parameter gradients for our setup.

Recall that our agents take actions – communication (q_t, a_t) and feature prediction \hat{y}_t – and our objective is to maximize the expected reward under the agents’ policies, summed over the entire dialog:

$$\max_{\theta_Q, \theta_A, \theta_f} J(\theta_Q, \theta_A, \theta_f) \quad \text{where,} \quad (3a)$$

$$J(\theta_Q, \theta_A, \theta_f) = \mathbb{E}_{\pi_Q, \pi_A} \left[\sum_{t=1}^T r_t(s_t^Q, (q_t, a_t, y_t)) \right] \quad (3b)$$

While the above is a natural objective, we find that considering the entire dialog as a single RL *episode* does not differentiate between individual good or bad exchanges within it. Thus, we update our model based on per-round rewards,

$$J(\theta_Q, \theta_A, \theta_f) = \mathbb{E}_{\pi_Q, \pi_A} \left[r_t(s_t^Q, (q_t, a_t, y_t)) \right] \quad (4)$$

Following the REINFORCE algorithm, we can write the gradient of this expectation as an expectation of a quantity related to the gradient. For θ_Q , we derive this explicitly:

$$\begin{aligned} \nabla_{\theta_Q} J &= \nabla_{\theta_Q} \left[\mathbb{E}_{\pi_Q, \pi_A} [r_t(\cdot)] \right] \quad (r_t \text{ inputs hidden to avoid clutter}) \\ &= \nabla_{\theta_Q} \left[\sum_{q_t, a_t} \pi_Q(q_t | s_{t-1}^Q) \pi_A(a_t | s_t^A) r_t(\cdot) \right] \\ &= \sum_{q_t, a_t} \pi_Q(q_t | s_{t-1}^Q) \nabla_{\theta_Q} \log \pi_Q(q_t | s_{t-1}^Q) \pi_A(a_t | s_t^A) r_t(\cdot) \\ &= \mathbb{E}_{\pi_Q, \pi_A} \left[r_t(\cdot) \nabla_{\theta_Q} \log \pi_Q(q_t | s_{t-1}^Q) \right] \end{aligned} \quad (5)$$

Similarly, gradient w.r.t. θ_A , i.e., $\nabla_{\theta_A} J$ can be derived as

$$\nabla_{\theta_A} J = \mathbb{E}_{\pi_Q, \pi_A} \left[r_t(\cdot) \nabla_{\theta_A} \log \pi_A(a_t | s_t^A) \right]. \quad (6)$$

As is standard practice, we estimate these expectations with sample averages. Specifically, we sample a question from Q-BOT (by sequentially sampling words from the question decoder LSTM till a stop token is produced), sample its answer from A-BOT, compute the scalar reward for this round, multiply that scalar reward by the gradient of the log-probability of this exchange, and propagate backward to compute gradients with respect to all parameters θ_Q, θ_A . This update has an intuitive interpretation – if a particular (q_t, a_t) is *informative* (i.e., leads to positive reward), its probabilities will be pushed up (positive gradient). Conversely, a poor exchange leading to negative reward will be pushed down (negative gradient).

Finally, since the feature regression network $f(\cdot)$ forms a deterministic policy, its parameters θ_f receive ‘supervised’ gradient updates for differentiable $\ell(\cdot, \cdot)$.

5. Emergence of Grounded Dialog

To succeed at our image guessing game, Q-BOT and A-BOT need to accomplish a number of challenging sub-tasks – they must learn a common language (*do you understand what I mean when I say ‘person’?*) and develop mappings between symbols and image representations (*what does ‘person’ look like?*), i.e., A-BOT must learn to ground language in visual perception to answer questions and Q-BOT must learn to predict plausible image representations

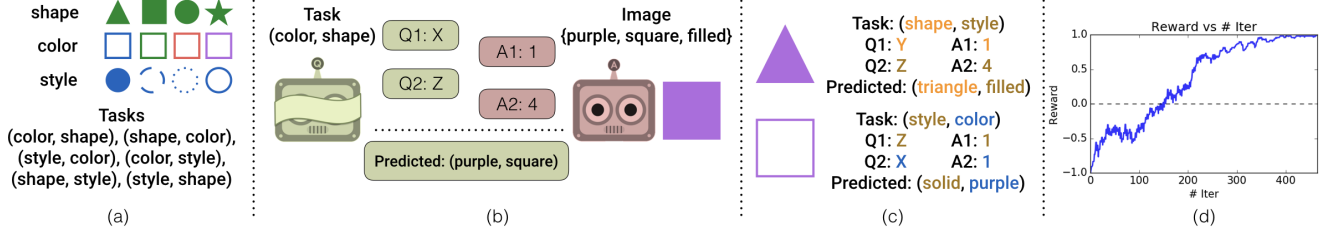


Figure 3: Emergence of grounded dialog: (a) Each ‘image’ has three attributes, and there are six tasks for Q-BOT (ordered pairs of attributes). (b) Both agents interact for two rounds followed by attribute pair prediction by Q-BOT. (c) Example 2-round dialog where grounding emerges: *color, shape, style* have been encoded as *X, Y, Z* respectively. (d) Improvement in reward while policy learning.

– all in an end-to-end manner from a distant reward function. Before diving in to the full task on real images, we conduct a ‘sanity check’ on a synthetic dataset with perfect perception to ask – *is this even possible?*

Setup. As shown in Fig. 3, we consider a synthetic world with ‘images’ represented as a triplet of attributes – 4 *shapes*, 4 *colors*, 4 *styles* – for a total of 64 unique images. A-BOT has perfect perception and is given direct access to this representation for an image. Q-BOT is tasked with deducing two attributes of the image in a particular order – e.g., if the task is (*shape, color*), Q-BOT would need to output (*square, purple*) for a (*purple, square, filled*) image seen by A-BOT (see Fig. 3b). We form all 6 such tasks per image.

Vocabulary. We conducted a series of pilot experiments and found the choice of the vocabulary size to be crucial for coaxing non-trivial ‘non-cheating’ behavior to emerge. For instance, we found that if the A-BOT vocabulary V_A is large enough, say $|V_A| \geq 64$ (#images), the optimal policy learnt simply ignores what Q-BOT asks and A-BOT conveys the entire image in a single token (e.g. token 1 \equiv (*red, square, filled*)). As with human communication, an impoverished vocabulary that cannot possibly encode the richness of the visual sensor is necessary for non-trivial dialog to emerge. To ensure at least 2 rounds of dialog, we restrict each agent to only produce a single symbol utterance per round from ‘minimal’ vocabularies $V_A = \{1, 2, 3, 4\}$ for A-BOT and $V_Q = \{X, Y, Z\}$ for Q-BOT. Since $|V_A|^{\text{\#rounds}} < \text{\#images}$, a non-trivial dialog is necessary to succeed at the task.

Policy Learning. Since the action space is discrete and small, we instantiate Q-BOT and A-BOT as fully specified tables of Q-values (state, action, future reward estimate) and apply tabular Q-learning with Monte Carlo estimation over 10k episodes to learn the policies. Updates are done alternately where one bot is frozen while the other is updated. During training, we use ϵ -greedy policies [33], ensuring an action probability of 0.6 for the greedy action and split the remaining probability uniformly across other actions. At test time, we default to greedy, deterministic policy obtained from these ϵ -greedy policies. The task requires outputting the correct attribute value pair based on the task and image. Since there are a total of $4 + 4 + 4 = 12$ unique values across the 3 attributes, Q-BOT’s final action selects one of $12 \times 12 = 144$ attribute-pairs. Further, task information is

excluded from Q-BOT’s state for this final action. We use +1 and −1 as rewards for right and wrong predictions.

Results. Fig. 3d shows the reward achieved by the agents’ policies vs. number of RL iterations (each with 10k episodes/dialogs). We can see that the two quickly learn the optimal policy. Fig. 3b,c show some example exchanges between the trained bots. We find that the two invent their own communication protocol – Q-BOT consistently uses specific symbols to query for specific attributes: $X \rightarrow \text{color}$, $Y \rightarrow \text{shape}$, $Z \rightarrow \text{style}$. And A-BOT consistently responds with specific symbols to indicate the inquired attribute, e.g., if Q-BOT emits X (asks for *color*), A-BOT responds with: $1 \rightarrow \text{purple}$, $2 \rightarrow \text{green}$, $3 \rightarrow \text{blue}$, $4 \rightarrow \text{red}$. Similar mappings exist for responses to other attributes. Essentially, we find the *automatic emergence of grounded language and a communication protocol* among ‘visual’ dialog agents without any human supervision! Refer to [18] for detailed analysis of ‘cheating’ strategies that emerge in a similar toy setup.

6. Experiments

Our synthetic experiments in the previous section establish that when faced with a cooperative task where information must be exchanged, two agents with perfect perception are capable of developing a complex communication protocol.

In general, with imperfect perception on real images, discovering human-interpretable language and communication strategy from scratch is both tremendously difficult and an unnecessary re-invention of English. We leverage the recently introduced VisDial v0.5 dataset [5] that contains human dialogs (10 rounds of question-answer pairs) on 68k images from the COCO dataset, for a total of 680k QA-pairs (some examples shown in Tab. 1).

Image Feature Regression. We consider an instantiation of the image guessing game described in Sec. 3 – specifically at each round t , Q-BOT regresses to the vector embedding \hat{y}_t of image I corresponding to the fc7 output from VGG-16 [32]. The distance metric used in the reward computation is ℓ_2 , i.e. $r_t(\cdot) = \|y^{gt} - \hat{y}_{t-1}\|_2^2 - \|y^{gt} - \hat{y}_t\|_2^2$.

Training Strategies. We found two training strategies to be crucial to ensure/improve the convergence of the RL framework described in Sec. 4, to produce any meaningful dialog exchanges, and to ground the agents in natural language.



Image + Caption	Human Dialog	SL-pretrained	RL-full-QAf
	What are their genders? 1 man 1 woman Are they both adults? Yes Do they wear goggles? Looks like sunglasses Do they have hats on? Man does Are there any other people? No What color is man's hat? Black Is it snowing now? No What is woman wearing? Blue jacket and black pants Are they smiling? Yes Do you see trees? Yes	How many people? 2 Male or female? I can't tell What are they wearing? I can't tell, they are far away What color are skis? I can't tell Are they wearing goggles? Yes Are they wearing goggles? Yes Are they wearing goggles? Yes Are they wearing goggles? Yes Are they wearing goggles? Yes Are they wearing goggles? Yes	Are people male or female or both? I think 1 is male, can't see other 1 ... What are they wearing? They are all wearing snow pants and jackets Is it snowing? It does not appear to be snowing at moment Can you see any trees? Yes, there are lot of trees in background Can you see any buildings in background? No, I can not Can you see any trees? Yes, there are lot of trees in background Can you see any buildings in background? No, I do not see any buildings Can you see any trees? Yes, there are lot of trees in background Can you see any buildings in background? No, I do not see any buildings Does it look like they are in park? It does not appear to be
	How old is man? In his teens Is he wearing helmet? No Is he wearing any protective gear? No What color is his hair? Dirty blonde or dark brown What color is his shirt? Black with red letters Is he wearing shorts? Yes What color are his shorts? Denim Are there other people visible? No What color are his shoes? I can't see them Is this at skate park? Yes	How old is man? I can't see his face, I don't know What is he wearing? shorts and t shirt What color is his shirt? white What color are his pants? black What color is his shirt? blue What color are his shoes? Black What color are his shoes? Black What color are his shoes? Black What color are his shoes? Black What color are his shoes? Black	What color is skateboard? It is hard to tell but I think it's black Is he wearing helmet? Yes Is he wearing knee pads? No, no knee pads Is it day or night? It is daytime Is it sunny? It is bright out, but I see no shadows, so I can not Can you see any other people? There are 2 other people in background Is it day or night? It is daytime Is it sunny? Yes, it is sunny What color is skateboard? It is hard to tell but I think it's black Is he wearing helmet? Yes, he is wearing helmet

Table 1: Selected examples of Q-BOT-A-BOT interactions for SL-pretrained and RL-full-QAf. RL-full-QAf interactions are diverse, less prone to repetitive and safe exchanges (“can’t tell”, “don’t know” *etc.*), and more image-discriminative.

1) **Supervised Pretraining.** We train both agents in a supervised manner on the `train` split of VisDial [5] v0.5 under an MLE objective. Thus, conditioned on human dialog history, Q-BOT is trained to generate the follow-up question by human1, A-BOT is trained to generate the response by human2, and the feature network $f(\cdot)$ is optimized to regress to y . The VGG-16 CNN in A-BOT is pretrained on ImageNet. This pretraining ensures that the agents can generally recognize some objects/scenes and emit English.

2) **Curriculum Learning.** After supervised pretraining, we ‘smoothly’ transition the agents to RL training according to a curriculum. Specifically, we continue supervised training for the first K rounds of dialog and transition to policy-gradient updates for the remaining $10 - K$ rounds. We start at $K = 9$ and gradually anneal to 0. This curriculum ensures that the agent team does not suddenly diverge off-policy, if one incorrect q or a is generated.

Models are pretrained for 15 epochs on VisDial, then transitioned to policy-gradient training by annealing K down by 1 every epoch. All LSTMs are 2-layered with 512-d hidden states. We use Adam [17] with a learning rate of 10^{-3} , and clamp gradients to $[-5, 5]$ to avoid explosion. There is no explicit state-dependent baseline in our training as we initialize from supervised pretraining and have zero-centered rewards, which ensures a good proportion of random samples are both positively and negatively reinforced.

Model Ablations. We compare to a few natural ablations of our full model, denoted RL-full-QAf. First, we evaluate the purely supervised agents (denoted SL-pretrained), *i.e.*, trained only on VisDial data (no RL). Comparison to these agents establishes how much RL helps over supervised learning. Second, we fix one of Q-BOT or A-BOT to the supervised pretrained initialization and train the other agent (and the regression network f) with RL; we label these as Frozen-Q or Frozen-A respectively. Comparing to these partially frozen agents tell us the importance of coordinated communication. Finally, we freeze the regression network f to the supervised pretrained initialization while training Q-BOT and A-BOT with RL. This measures

improvements from language adaptation alone.

We quantify performance along two axes – how well agents perform at image guessing and how closely they emulate human dialogs (*i.e.* performance on the VisDial dataset [5]).

Evaluation: Guessing Game. To assess how well the agents have learned to cooperate at the image guessing task, we setup an image retrieval experiment based on the `test` split of VisDial v0.5 ($\sim 9.5k$ images), which were never seen by the agents in RL training. We present each image + an automatically generated caption [15] to the agents, and allow them to communicate over 10 rounds of dialog. After each round, Q-BOT predicts a feature representation \hat{y}_t . We sort the entire `test` set in ascending distance to this prediction and compute the rank of the source image.

Fig. 4a shows the mean percentile rank of the source image for our method and the baselines across the rounds (shaded region indicates standard error). A percentile rank of 95% means that the source image is closer to the prediction than 95% of the images in the set. Tab. 1 shows example exchanges between two humans (from VisDial), the SL-pretrained and the RL-full-QAf agents.

- **RL improves image identification.** We see that RL-full-QAf outperforms SL-pretrained and all other ablations (*e.g.*, at round 10, improving percentile rank by over 3%), indicating that our training framework is indeed effective at training these agents for image guessing.

- **All agents ‘forget’; RL agents forget less.** One interesting trend we note in Fig. 4a is that all methods significantly improve from round 0 (caption-based retrieval) to rounds 2 or 3, but beyond that all methods with the exception of RL-full-QAf get *worse*, even though they have strictly more information. As shown in Tab. 1, agents will often get stuck in infinite repeating loops but this is much rarer for RL agents. Moreover, even when RL agents repeat themselves, it is after longer gaps (2-5 rounds). We conjecture that this is because errors cascade in SL agents, while RL agents can drive conversations away from these poor paths.

- **RL leads to more informative dialog.** SL A-BOT tends

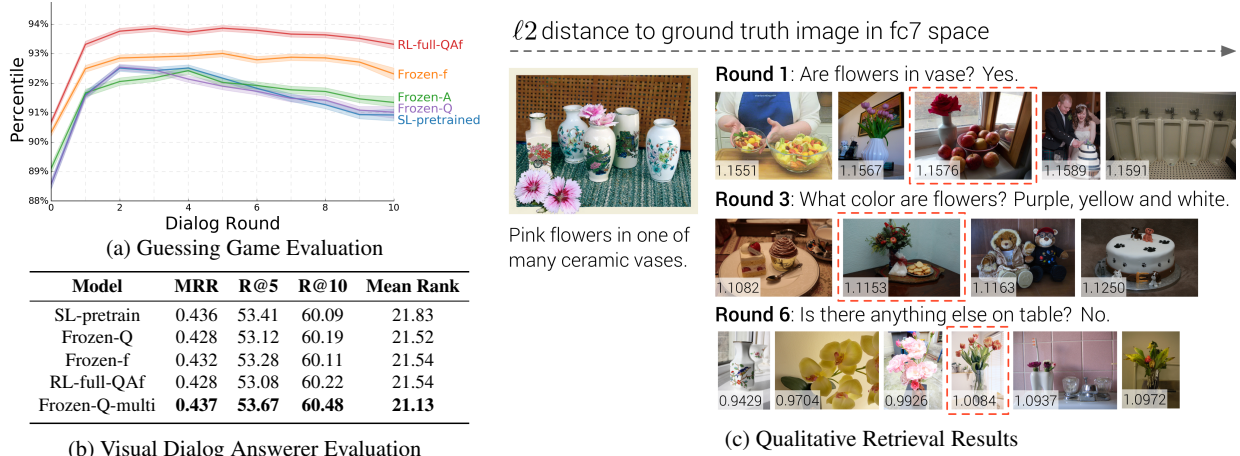


Figure 4: **a) Guessing Game Evaluation.** Percentile of GT image (higher is better) based on image retrieval using fc7 predictions vs. round of dialog. Round 0 corresponds to image guess based on caption. RL-full-QAf outperforms SL-pretrained and other ablations. Error bars show standard error of means. **(c)** shows qualitative results on image retrieval. Left column shows true image and caption, right column shows dialog exchange and image guess (highlighted in red) alongside closest images from VisDial test in sorted order of distance to ground truth. See supplementary for more qualitative results. **b) VisDial Evaluation.** Performance of A-BOT on VisDial v0.5 test, by mean reciprocal rank (MRR), recall@ k for $k = \{5, 10\}$ and mean rank. Higher is better for MRR and recall@ k , while lower is better for mean rank. Frozen-Q-multi outperforms other models on VisDial metrics by 3% relative gain.

to produce ‘safe’ generic responses (*‘I don’t know’, ‘I can’t see’*) but RL A-BOT responses are much more detailed (*‘It is hard to tell but I think it’s black’*). These observations are consistent with work in text-only dialog [21]. Our hypothesis is that human responses are diverse and SL trained agents tend to ‘hedge their bets’ and achieve a reasonable log-likelihood by being non-committal. In contrast, such ‘safe’ responses do not help Q-BOT in picking the correct image, thus encouraging a more informative RL A-BOT.

Evaluation: Emulating Human Dialogs. To quantify how well the agents emulate human dialog, we evaluate A-BOT on the retrieval metrics proposed by Das *et al.* [5]. Specifically, every question in VisDial is accompanied by 100 candidate responses. We use the log-likelihood assigned by the A-BOT answer decoder to sort these candidates and report the results in Tab. 4b. We find that despite the RL A-BOT’s answer being more informative, the improvements on VisDial metrics are minor. We believe this is because while the answers are correct, they may not necessarily mimic human responses (which is what the answer retrieval metrics check for). In order to dig deeper, we train a variant of Frozen-Q with a multi-task objective – simultaneous (1) ground truth answer supervision and (2) image guessing reward, to keep A-BOT close to human-like responses. We use a weight of 1.0 for the SL loss and 10.0 for RL. This model, denoted Frozen-Q-multi, performs better than all other approaches on VisDial answering metrics, improving the best reported result on VisDial by 0.7 mean rank (relative improvement of 3%). Note that this gain is ‘for free’ since no additional annotations were required for RL.

Human Study. We conducted a human interpretability study to measure (1) whether humans can easily un-

derstand the Q-BOT-A-BOT dialog, and (2) how image-discriminative the interactions are. We show human subjects a pool of 16 images, the agent dialog (10 rounds), and ask humans to pick their top-5 guesses for the image the two agents are talking about. We find that mean rank of the ground-truth image for SL-pretrained agent dialog is 3.70 vs. 2.73 for RL-full-QAf dialog. In terms of MRR, the comparison is 0.518 vs. 0.622 respectively. Thus, under both metrics, humans find it easier to guess the unseen image based on RL-full-QAf dialog exchanges, which shows that agents trained within our framework (1) successfully develop image-discriminative language, and (2) this language is interpretable; they do not deviate off English.

7. Conclusions

To summarize, we introduce a novel training framework for visually-grounded dialog agents by posing a cooperative ‘image-guessing’ game between two agents. We use deep reinforcement learning to end-to-end learn the policies of these agents – from pixels to multi-agent multi-round dialog to game reward. We demonstrate the power of this framework in a completely ungrounded synthetic world, where the agents communicate via symbols with no pre-specified meanings (X, Y, Z). We find that two bots invent their own communication protocol without any human supervision. We go on to instantiate this game on the VisDial [5] dataset, where we pretrain with supervised dialog data to align the communication between agents to ‘English’, ensuring human-interpretability and scope for pairing these agents with humans. We find that the RL fine-tuned agents not only significantly outperform SL agents, but learn to play to each other’s strengths, all the while remaining interpretable to outside humans observers.

Acknowledgements. We thank Devi Parikh for helpful discussions. This work was funded in part by the following awards to DB – NSF CAREER award, ONR YIP award, ONR Grant N00014-14-1-0679, ARO YIP award, ICTAS Junior Faculty award, Google Faculty Research Award, Amazon Academic Research Award, AWS Cloud Credits for Research, and NVIDIA GPU donations. SK was supported by ONR Grant N00014-12-1-0903, and SL was partially supported by the Bradley Postdoctoral Fellowship. Views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

- [1] J. Andreas, A. Dragan, and D. Klein. Translating neuralese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 232–242, Vancouver, Canada, July 2017. Association for Computational Linguistics. 3
- [2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1, 2, 3
- [3] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. In *UIST*, 2010. 1
- [4] X. Chen and C. L. Zitnick. Mind’s Eye: A Recurrent Visual Representation for Image Caption Generation. In *CVPR*, 2015. 1
- [5] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual Dialog. In *CVPR*, 2017. 1, 2, 3, 4, 6, 7, 8
- [6] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. GuessWhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017. 1, 3
- [7] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*, 2015. 3
- [8] K. Evtimova, A. Drozdov, D. Kiela, and K. Cho. Emergent language in a multi-modal, multi-step referential game. *arXiv preprint arXiv:1705.10369*, 2017. 3
- [9] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From Captions to Visual Concepts and Back. In *CVPR*, 2015. 3
- [10] J. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, 2016. 3
- [11] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In *NIPS*, 2015. 3
- [12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *NIPS*, 2014. 3
- [13] S. Havrylov and I. Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *ICLR Workshop*, 2017. 3
- [14] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *CVPR*, 2016. 1
- [15] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 3, 7
- [16] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014. 3
- [17] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 7
- [18] S. Kottur, J. M. Moura, S. Lee, and D. Batra. Natural language does not emerge ‘naturally’ in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. 3, 6
- [19] A. Lazaridou, A. Peysakhovich, and M. Baroni. Multi-agent cooperation and the emergence of (natural) language. In *ICLR*, 2017. 3
- [20] D. Lewis. *Convention: A philosophical study*. John Wiley & Sons, 2008. 3
- [21] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*, 2016. 3, 8
- [22] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017. 3
- [23] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017. 3
- [24] M. Malinowski and M. Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *NIPS*, 2014. 3
- [25] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 1, 3
- [26] I. Mordatch and P. Abbeel. Emergence of grounded compositional language in multi-agent populations. *arXiv preprint arXiv:1703.04908*, 2017. 3
- [27] S. Nolfi and M. Mirolli. *Evolution of Communication and Language in Embodied Agents*. Springer Publishing Company, Incorporated, 1st edition, 2009. 3
- [28] M. Ren, R. Kiros, and R. Zemel. Exploring Models and Data for Image Question Answering. In *NIPS*, 2015. 1, 3
- [29] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*, 2016. 4
- [30] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. *arXiv preprint arXiv:1605.06069*, 2016. 4
- [31] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016. 3

- [32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4, 6
- [33] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998. 6
- [34] M. Tapaswi, Y. Zhu, R. Stiefelhaven, A. Torralba, R. Ur-
tasun, and S. Fidler. MovieQA: Understanding Stories in
Movies through Question-Answering. In *CVPR*, 2016. 1
- [35] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S. C. Zhu.
Joint Video and Text Parsing for Understanding Events and
Answering Queries. *IEEE MultiMedia*, 2014. 1
- [36] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney,
T. Darrell, and K. Saenko. Sequence to Sequence - Video to
Text. In *ICCV*, 2015. 1
- [37] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J.
Mooney, and K. Saenko. Translating Videos to Natural Lan-
guage Using Deep Recurrent Neural Networks. In *NAACL*
HLT, 2015. 1
- [38] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and
tell: A neural image caption generator. In *CVPR*, 2015. 1, 3
- [39] R. J. Williams. Simple statistical gradient-following algo-
rithms for connectionist reinforcement learning. *Machine*
learning, 8(3-4):229–256, 1992. 5
- [40] S. Wu, H. Pique, and J. Wieland. Using artifi-
cial intelligence to help blind people ‘see’ facebook.
[http://newsroom.fb.com/news/2016/04/using-artificial-
intelligence-to-help-blind-people-see-facebook/](http://newsroom.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/), 2016.
1
- [41] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhut-
dinov, R. S. Zemel, and Y. Bengio. Show, Attend and Tell:
Neural Image Caption Generation with Visual Attention. In
ICML, 2015. 1