Investigating Visual Dialog Models for Goal-Driven Self-Talk

Prithvijit Chattopadhyay
School of Interactive Computing
Georgia Tech
prithvijit3@gatech.edu

Devi Parikh
School of Interactive Computing
Georgia Tech
parikh@gatech.edu

Abstract

*Prior work by Das & Kottur et al. [3] on visual dialog has proposed training goal-driven visual dialog agents in a cooperative image guessing game – a questioner (*Q-BOT*) and answerer (*A-BOT*) talking to each other to help* Q-BOT *predict an unknown image – and has shown that training such agents with reinforcement learning improves performance over supervised learning counterparts in the game. We propose an approach to improve the quality of dialogs (or conversations) generated by such goal-driven visual dialog agents. Specifically, we introduce a ranking based answer evaluator that learns to rank "human-generated" answers above "machine-generated" answers. Once learned, the answer evaluator is used as a black-box to provide a score of humanness to the responses generated by* A-BOT *during* Q-BOT-A-BOT *"self-talk", thereby incentivzing* A-BOT *to generate more human-like answers. We show that the dialogs generated by our proposed approach are comparable in terms of "relevance" to the image being talked about and are more "diverse" compared to the ones generated by Das & Kottur et al. [3].*

1. Introduction

Visual-Dialog is a high-level AI task introduced by Das *et al*. [2] where the intent is to design an AI agent capable of conversing with a human about an image. More specifically, given an image, a dialog history, and a follow-up question about the image, the agent has to

- *ground the question in the image* – e.g., reason about whether nouns, pronouns in the question refer to specific objects in the image
- *appropriately infer relevant context from history* – e.g., realize that 'her' is being used to refer to the woman referred to in the previous exchange
- *accurately answer the question*

The task is positioned suitably as a middle-ground between specific downstream 'chatbot'-based applications and as a benchmark to evaluate machine intelligence. Das *et al*. [2] also released a large-scale 'VisDial' dataset consisting of 10 question-answer pairs (comprising a dialog) on $\sim$120k images from human-human interactions to train large-scale models. Ever since it's introduction, the task has gained immense popularity in the AI community and has resulted significant progress along the lines of image-based dialog.

Das *et al*. [2] frame this task as a supervised learning problem. Posing visual dialog as a supervised learning problem is unnatural. This is because the AI agent answering questions conditioned on the image and dialog history never gets to encounter this notion of *compounding errors* – by construction, it doesn't have to account for wrongful answer predictions made at an earlier round of dialog. This is because it always gets fed the ground-truth question and dialog history and not what the agent might have said earlier. This assumption is hardly true practice. While interacting with a human, the previous (correct or incorrect) answers should a form a part of the dialog history tracked by a 'chatbot' deployed in the real-world.

To make such dialog agents responsible for the incorrect predictions made at an earlier round, Das & Kottur *et al*. [3] modify the visual dialog task to a goal-driven setting. Specifically, [3] designs a cooperative image-guessing between a questioning (Q-BOT) and an answering (A-BOT) agent. Q-BOT is shown a one-line description of an unseen image that A-BOT has access to and Q-BOT is allowed to ask questions (in natural language) to A-BOT for a fixed number of rounds and simultaneously make predictions of the unseen image. Das & Kottur *et al*. [3] cast this asymmetric information game in a reinforcement learning (RL) framework – both the agents are rewarded for Q-BOT's image-guessing performance. Thus, there is incentive for Q-BOT to ask questions informative of the "hidden" image, and for A-BOT to provide meaningful answers to the same. This setting yields several advantages – (1) both Q-BOT and A-BOT are now responsible for the things that they've uttered in the past and (2) more importantly, having two bots conversing with each other about an image makes the data collection process much cheaper – we can just allow two 'perfect' agents to talk to each other about an image and utilize those conversations as future training data instead of asking more humans to do the same.

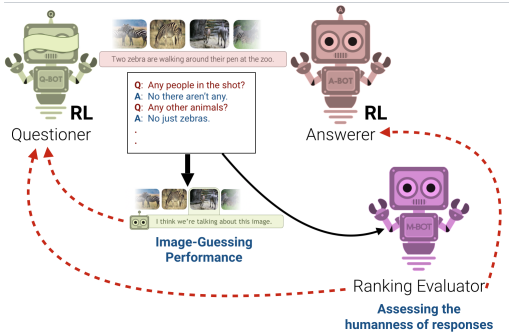

Figure 1. **RL-Bots-Rank-Eval-Gumbel-ST**: An overview of our proposed approach. We first replace the gradient estimator used to update language models of Q-BOT and A-BOT [3] from REINFORCE → Gumbel-ST. In addition to the vanilla image-guessing pipeline adopted in [3], we also add a ranking evaluator which incentivizes the A-BOT to produce for human-like answers.

While this Q-BOT-A-BOT "self-talk" offers exciting prospects, the current state of such agents is far from enabling us to achieve these goals – (1) there is lack of clarity about how to judge the "humanness" of such dialogs, (2) bots trained in this paradigm are unable to significantly facilitate human-AI team performance [1] and (3) there's a clear lack of diversity in the things (QA exchanges) being talked about at different rounds of dialog [5]. In this project, we restrict ourselves to (1) – we ground notions of humanness of generated dialogs in terms of "relevance" and "diversity" of QA exchanges at different rounds and devise an approach to improve on said measures.

Relevance – we consider Q-BOT's ability to predict the image from the generated exchanges between Q-BOT-A-BOT as a measure of relevance. **Diversity** – we consider the ability to talk about different "facts" (QA-exchanges) at different rounds as a measure of diversity. Therefore, given the base set of models introduced in [3], the goal in this project is to devise an approach get better versions of Q-BOT and A-BOT that perform well on the aforementioned measures of relevance and diversity. We also report results on the standard evaluation metrics introduced in [2].

2. Approach

We introduce two major changes to the training pipeline of Das & Kottur *et al*. [3]. First, we change the way in which sentences are generated by the language specific components of Q-BOT and A-BOT. Following this, we introduce an answer evaluator which has learned to rank human-like answers above answers generated by A-BOT and use this evaluator as a black-box unit to incentivize A-BOT to generate more human-like answers during "self-talk".

2.1. Modifying the Language Generation Scheme

To provide some context, we briefly describe how natural-language sentences are generated by Q-BOT and A-BOT. As is standard with deep-learning based recurrent language models, given a large-vocabulary of tokens or words ('a', 'what', 'man', etc.) the language components of these bots model a probability distribution over possible sentences ('what is the man doing?') – as a sequence of tokens from the vocabulary – given appropriate dialog context (dialog history and image).

Generating a sentence from these modeled probability distributions essentially boils down to sampling a sentence. However, since sampling in itself is not *exactly* differentiable, standard methods to update Q-BOT and A-BOT based on gradients of the former's image-guessing performance cannot be used. Casting this in a RL framework, Das & Kottur [3] make use of REINFORCE [6], a standard gradient estimator used in RL problems. However, gradients estimated from REINFORCE suffer from high-variance, which makes the overall learning process slower and unstable. We replace REINFORCE with the Gumbel Straight-Through Estimator (Gumbel-ST) [4] which – (1) makes the sampling process *loosely* differentiable and (2) results in gradients with lower variance, thereby making the whole learning process faster and much more stable.

2.2. Ranking-based Answer Evaluator

On top of the setting used by Das & Kottur *et al*. [3], we add a learned answer evaluator. Specifically, given the dialog context (image and history), the evaluator learns to rank answers provided by humans higher than the ones generated by an A-BOT. Once we've learnt a sufficiently good ranking based evaluator, we use it as black-box and use the ranking score of the answers generated during the Q-BOT-A-BOT exchanges as an additional source of reward to incentivize A-BOT to generate more human-like answers.

Pre-training the Evaluator. Specifically, we use the supervised learning answering (A-BOT) model from [2] as the source of "machine" or generated answers ($\mathbf{A}_{gen}$) and answers from the VisDial dataset [2] as the source for "human" generated answers ($\mathbf{A}_h$). Given the dialog context $\mathbf{c}$ from the VisDial dataset, consisting of the image and dialog history, in pre-training, the evaluator is asked to minimize the objective:

$$\mathcal{L}_{pre} = \log(1+\exp\left(f(\mathbf{A}_{gen})^T g(\mathbf{c}) - f(\mathbf{A}_h)^T g(\mathbf{c})\right)) \quad (1)$$

where $f(\cdot)$ and $g(\cdot)$ are functions that transform the answers and the dialog context to vectors of the same dimension and $f(\cdot)^T g(\cdot)$ denotes the inner products of these vectors. Eq. 2 encourages the evaluator to rank the "human" responses higher than the "machine" generated responses.

Using Evaluator Score as Reward. When Q-BOT and A-BOT are interacting with each other in the image guessing game, we basically treat the learnt evaluator as a "frozen" black-box and consider minimizing the following

$$\mathcal{L}_{sft} = \log(1 + \exp\left(-f(\mathbf{A}_{gen}^{sft})^T g(\mathbf{c}^{sft})\right)) \quad (2)$$

as an incentive to encourage more human-like answers from A-BOT in addition to providing more informative answers

Model	NDCG ↑	MRR ↑	MR ↓
SL-A-BOT [3]*	54.33	0.4586	19.77
RL-A-BOT-REINFORCE [3]	**55.11**	0.4637	19.58
RL-A-BOT-Gumbel-ST	54.47	0.4625	19.45
RL-A-BOT-Rank-Eval-Gumbel-ST	54.92	**0.4645**	**19.42**

Table 1. Performance of the answerer (A-BOT) on the VisDial [2] metrics on v1.0-val when trained via "self-talk" with a questioner (Q-BOT) using different techniques. ↑ indicates higher is better. ↓ indicates lower is better. *The baseline SL-A-BOT performance is included here for completeness.

Model	Percentile ↑	Unique Ques. ↑	Dist-3 ↑	Ent-3 ↑
SL-Baseline [3]*	93.44	0.7556	0.6841	6.14
RL-Bots-REINFORCE [3]	**96.30**	0.7016	0.6476	6.13
RL-Bots-Gumbel-ST	96.21	0.6530	0.6254	6.12
RL-Bots-Rank-Eval-Gumbel-ST	96.25	**0.7309**	**0.6653**	**6.16**

Table 2. Performance of Q-BOT-A-BOT teams in terms of "relevance" (indicated by percentile) and "diversity" (indicated by all other metrics) of generated dialogs on images from VisDial v1.0-val [2]. ↑ indicates higher is better. ↓ indicates lower is better. *The baseline SL-Bots performance is included here for completeness.

to Q-BOT to maximize the latter's image-guessing performance in the cooperative game.

3. Experiments

3.1. Evaluation Metrics

Recall that we measure the performance of Q-BOT-A-BOT exchanges over 10 rounds of dialog in terms of *relevance* and *diversity* (see Sec. 1). In Table. 1, we report performances of various A-BOT's on the standard metrics introduced in [2] – namely, (1) **NDCG** - a score indicating the quality of ranking of a set of ground-truth answers based on A-BOT's belief normalized by how relevant the humans consider each of the answers to be, (2) **Mean Rank (MR)** and **Mean Reciprocal Rank (MRR)** – based on the ranking of the ground-truth answer for a fixed dialog context under the model's belief. In Table. 2, we report Q-BOT-A-BOT "self-talk" performances in terms of `Relevance` – **Percentile** in which the image guessed by Q-BOT based on the generated dialog lies and `Diversity` – (1) **Unique questions.** the number of unique questions per-dialog instance, (2) **Dist-n and Ent-n.** the number and entropy of unique trigrams in the generated QA-exchanges normalized by the total number of tokens.

3.2. Results

We summarize our key observations below:

- **A-BOT on VisDial**: We observe that (see Table. 1) REINFORCE → Gumbel-ST results in a minor drop in performance in terms of NDCG, MR and MRR. Adding the ranking evaluator during self-talk improves both MRR and MR over REINFORCE and Gumbel-ST. However, these improvements (and drops in performances) are not statistically significant.
- **Relevance**: We observe that (see Percentile in Table. 2) REINFORCE → Gumbel-ST results in a drop in performance in terms of relevance. Adding the ranking-evaluator during self-talk improves performance over Gumbel-ST but still falls short of REINFORCE by a small margin. Note that the marginal drop in relevance is not significant – the dialogs generated by our approach are comparable to the ones generated by REINFORCE in terms of relevance.
- **Diversity**: We observe that (see Unique Questions, Dist-3 and Ent-3 in Table. 2) while REINFORCE → Gumbel-ST results in a drop in performance in terms of diversity metrics, adding the ranking-evaluator during self-talk consistently improves the diversity of the generated dialogs. Therefore, compared to the training paramdigm proposed in [3], RL-Bots-Rank-Eval-Gumbel-ST generates Q-BOT-A-BOT exchanges which are compratively *relevant* but more *diverse*.
- **Comparison to SL-A-BOT**: Note that in terms of the VisDial evaluation metrics, while self-talk generally improves the A-BOT's in terms of NDCG, this improvement comes at a cost of reduction in MR and MRR (see Table. 1). From Table. 2, we find that while self-talk improves *relevance* over the baseline SL versions of Q-BOT-A-BOT from [2, 3], it results in a drop in *diversity* of the generated dialogs.

Conclusion. We find that our proposed approach RL-Bots-Rank-Eval-Gumbel-ST generates dialog which is *relevant* enough but more *diverse* compared to the vanilla REINFORCE training pipeline proposed in [3]. More generally, we do note that the generated dialogs from Q-BOT-A-BOT exchanges are from enabling us to achieve the goals highlighted in Sec. 1 as they stil fall behind supervised versions of these models [2, 3] in terms of diversity. Therefore, there's still a long way to go in terms of using Q-BOT-A-BOT self-talk as a method to generate better chatbots or as a source of dialog data. This work just scratches the surface of such approaches, and numerous other avenues of exploration remain. Introducing a similar evaluator for the questions generated by Q-BOT is one natural extension.

References

[1] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. Evaluating visual conversational agents via cooperative human-ai games. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, 2017.

[2] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *CVPR*, 2017.

[3] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In *ICCV*, 2017.

[4] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[5] Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. Improving generative visual dialog by answering diverse questions. *arXiv preprint arXiv:1909.10470*, 2019.

[6] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.