Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Varun Gangal, Sarah Mallepalle, Rhea Jain

7 November 2017

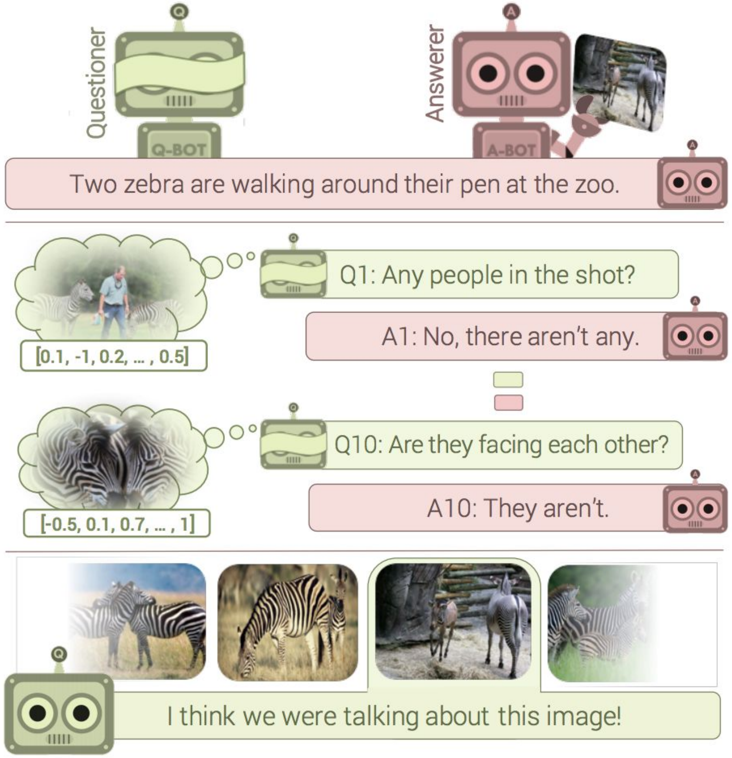
Introduction

- Focus: Visually-grounded conversational artificial intelligence
- Goal:
  - Develop agents that can “see”: understanding the contents of an image
  - Develop agents that can “communicate” what they see: hold a dialog using natural language involving questions and answers about an image.

Introduction

- This is the first-goal driven training for visual question answering and dialog agents.
- Presented as an “image guessing” game between two agents: a question bot (Q-Bot) and an answer bot (A-bot)
  - Communicate in natural language dialog.
  - Q-bot selects an unseen image from a lineup depending on the information it receives from A-bot
  - Method of deep reinforcement learning.

A Visual


Questioner
Answerer
Q-BOT
A-BOT
Two zebra are walking around their pen at the zoo.
Q1: Any people in the shot?
[0.1, -1, 0.2, ... , 0.5]
A1: No, there aren't any.
Q10: Are they facing each other?
[-0.5, 0.1, 0.7, ... , 1]
A10: They aren't.
I think we were talking about this image!

Recent Works

- Recent works studying visually grounded dialog treat dialog as a "static" supervised learning problem, rather than an interactive agent learning problem.

Recent Works

1. Use a dataset of a human sequence of Q-A pairs about an image: $(q_1, a_1) \ldots (q_T, a_T)$
2. A machine (deep NN) is provided with an image I, human dialog up to round t-1, and a follow-up question q_t. Use supervised learning to generate human response, a_t.
3. The machine's answer $\hat{a}_t$ is thrown away, because at the next round t+1, the machine is provided with the ground-truth human dialog including the human response at.
   - The machine is never allowed to steer the conversation, because its answer would not be exactly in the dataset, making it not viable for supervised learning.

Our Setup: Interactive Guessing Game

- Q-bot is shown a 1-sentence description/caption of an unseen image and communicates in natural language with A-bot, who sees the image.

- Objective: Q-bot needs to build a mental model of the unseen image purely from the natural language dialog, and retrieve image from a lineup.

Our Setup: Interactive Guessing Game

- Process:
  - At every round of dialog, Q-bot listens to A-bot's answer, updates its beliefs, and makes a prediction y^ about the visual representation of the unseen image.
    - This description can be in many forms, from image embeddings to textual descriptions to pixel-level image generations.
  - Q-bot receives a reward from the environment based on how close Q-bot's prediction is to the true fc7 vector representation of I.

Experimental Results

Experiment 1: A “sanity check” of RL where perception is perfect in a synthetic world.

- Images containing a single object defined by shape, color, and style.
- Q-not must identify an image by learning about these three attributes.
- Communicate via ungrounded vocabulary: symbols with no pre-specified human interpretable meanings (X, Y, 1, 2, etc.)
- Result: Automatic emergence of grounded language and communication between visual dialog agents.

Experimental Results

Experiment 2: Large-scale real-image experiments on VisDial dataset.

- Imperfect perception on real images
- Discovering a human-interpretable language and communication strategy from scratch.
- Pretrain with SL with dialog data in VisDial before "fine-tuning" with RL.
- Result: RL bots significantly outperform SL bots
- Main Difference: SL Q-bot attempts to mimic how humans ask questions, while the RL trained Q-bot shifts strategies and asks questions the A-bot is better at answering, so dialog is more informative.

REINFORCEMENT LEARNING FOR DIALOG AGENTS

ACTION SPACE

- Q-BOT , A-BOT : All possible output sequences under a token vocabulary V - Discrete
- Q-BOT : Visual representation of the unseen image - Continous

STATE

- Q-BOT : $s_t^Q = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}]$
- A-BOT: $s_t^A = [I, c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t]$

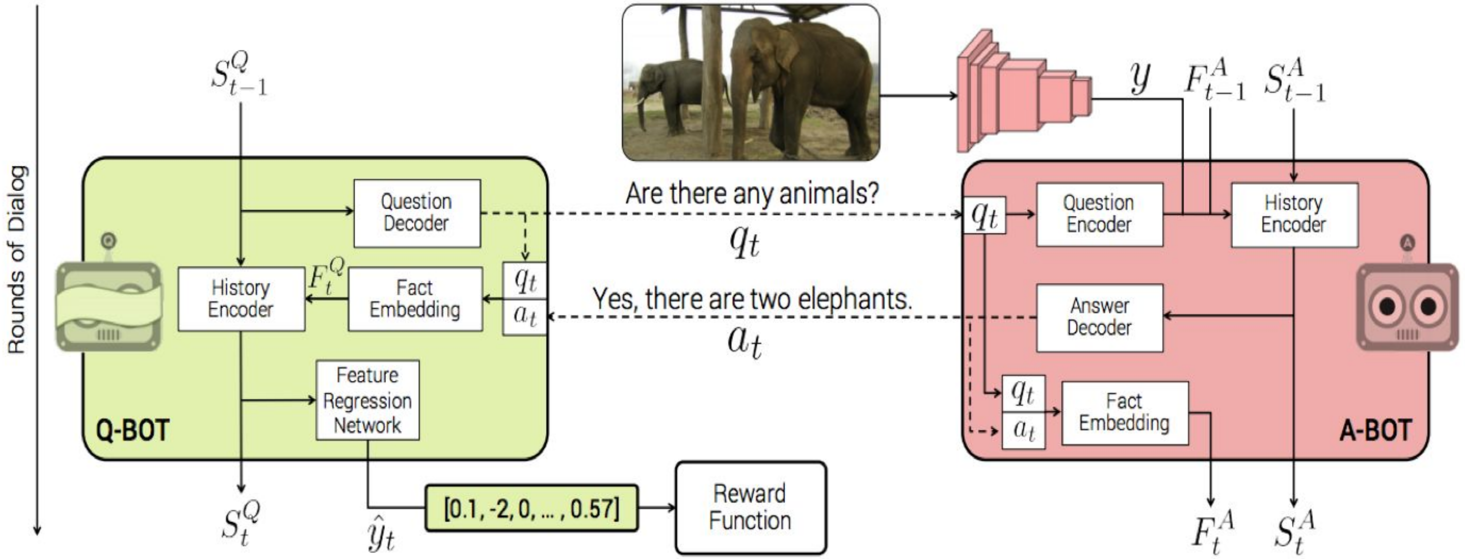
POLICY

- Q-BOT : $\pi_Q(q_t|s_t^Q;\theta_Q)$
- A-BOT : $\pi_A(q_t|s_t^A;\theta_A)$

REWARD

$$\sum_{t=1}^{T} r_t(s_t^Q,(q_t,a_t,y_t)) = l(\hat{y_0},y_{gt}) - l(\hat{y_T},y_{gt})$$

Policy Networks for Q-Bot and A-Bot



Joint Training with Policy Gradients

Maximize the Reward Function over agent's policies

REINFORCE ALGORITHM

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

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

Scalar Reward at the end of every round

Informative - Increases Probabilities

Emergence of Grounded Dialog

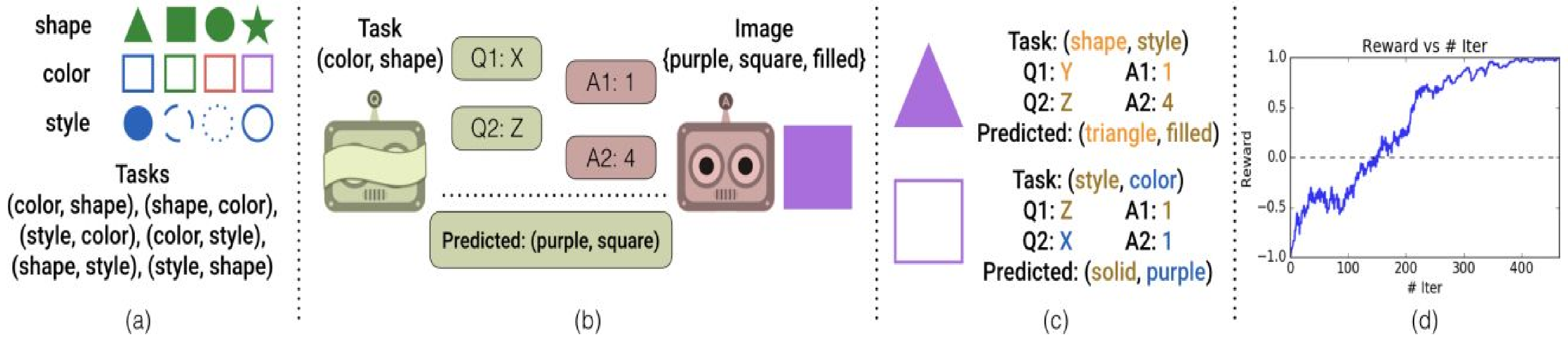


Figure 3: Emergence of grounded dialog: (a) Each 'image' has three attributes, and there are six tasks for Q-BOT (ordered pairs of attributes). (b) Both agents interact for two rounds followed by attribute pair prediction by Q-BOT. (c) Example 2-round dialog where grounding emerges: *color*, *shape*,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

Why Do We Need Supervised Pre-Training?

- Large action space ($|V_q| * |V_a|$)
- Need a good starting point
- No point learning a new language from-scratch (when we have English)
- Loss function: Max-likelihood $P(y|x)$

How does RL-training improve things?

- Problem 1: The Common Response Problem
- Problem 2: Forgetting/Repetition
- Hypotheses validated by qualitative examples (next)
- Loss function:

$$r_t(\cdot) = \|y^{gt} - \hat{y}_{t-1}\|_2^2 - \|y^{gt} - \hat{y}_t\|_2^2$$

Common Response Problem

- Diverse informative responses; few perfunctory responses
- Q: What's the man wearing?
- Perfunctory: "I can't tell", "A shirt", "Clothes"
- Informative: "A red spotted T-shirt", "a black robe", "a wizard's cloak"
- A known problem even for chatbots (general purpose dialog)

Image + Caption	Human-Human dialog [4]		SL-pretrained Q-BOT-A-BOT dialog		RL-full-QAf Q-BOT-A-BOT dialog	
	What are their genders?	1 man 1 woman	How many people?	2	Are people male or female or both?	I think 1 is male, can't see other 1 ...
	Are they both adults?	Yes	Male or female?	I can't tell	What are they wearing?	They are all wearing snow pants and jackets
	Do they wear goggles?	Looks like sunglasses	What are they wearing?	I can't tell, they are far away	Is it snowing?	It does not appear to be snowing at moment
	Do they have hats on?	Man does	What color are skis?	I can't tell	Can you see any trees?	Yes, there are lot of trees in background
	Are there any other people?	No	Are they wearing goggles?	Yes	Can you see any buildings in background?	No, I can not
	What color is man's hat?	Black	Are they wearing goggles?	Yes	Can you see any trees?	Yes, there are lot of trees in background
	Is it snowing now?	No	Are they wearing goggles?	Yes	Can you see any buildings in background?	No, I do not see any buildings
	What is woman wearing?	Blue jacket and black pants	Are they wearing goggles?	Yes	Can you see any trees?	Yes, there are lot of trees in background
	Are they smiling?	Yes	Are they wearing hats?	Yes	Can you see any buildings in background?	No , I do not see any buildings
	Do you see trees?	Yes	Are they wearing goggles?	Yes	Does it look like they are in park?	It does not appear to be
A couple of people in the snow on skis.						

	Is giraffe adult?	I cannot tell, I only see face	Is there only 1 giraffe?	Yes	Is there only 1 giraffe?	Yes
	Is grass at higher altitude than its feet?	No	Are there any other animals?	No	Are there any other animals?	No
	Is rock big?	Yes	Are there any people?	No	Are there any people?	No
	How is weather?	Sunny	Are there any other animals?	No	Does this look like zoo?	Can't tell
	What color is rock?	Tan color	Is it day or night?	Day	Is there fence?	No
	Is grass green or dry?	Green but looks little dry	Is it sunny?	Yes	Is it day or night?	Day
	Does giraffe look healthy?	Yes	What time of day it is?	I can't tell	Is it sunny?	Yes
	Does giraffe look happy?	It looks tired	Are there any other animals?	No	What time of day is it?	Morning
	Is giraffe laying down?	Yes	Are there any trees?	No	Are there any trees?	No
	Is giraffe's mouth closed?	Yes	Are there any other animals?	No	Can you see sky?	No
A giraffe is resting its head in the grass near a rock.						

Transitioning from SL to RL

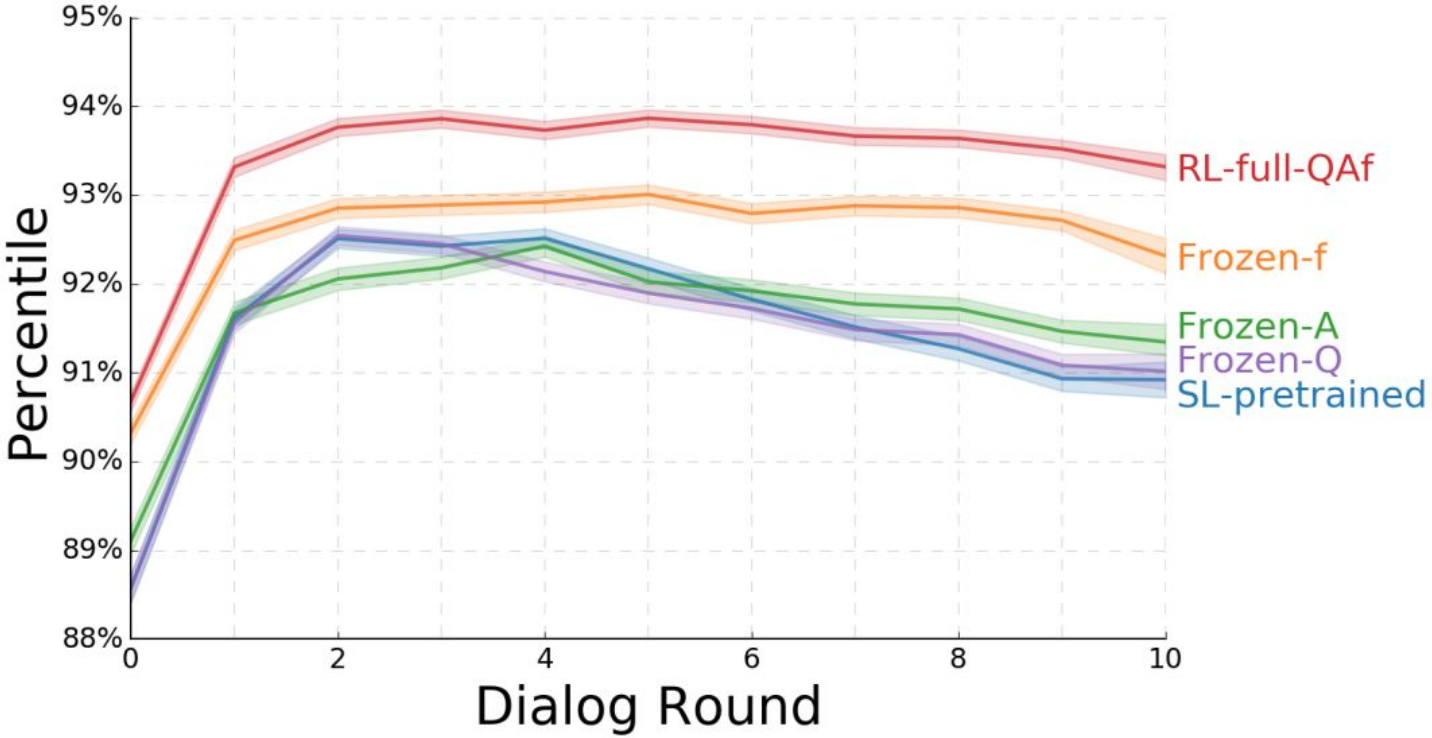
- For each dialog, SL for first k rounds, RL for remaining 10-K rounds
- Decrease K by 1 each epoch
- After 9 epochs, fully RL-based training

Ablations

- SL-pretrain - Only supervised pre-training
- Frozen-Q: Q-bot frozen, A-bot and f RL trained
- Frozen-f: f Frozen, Q-bot and A-bot RL trained
- RL-full-QAF: Everything RL trained
- Frozen-Q-multi: Q-bot frozen, A-bot trained with multi-task reward (likelihood+success)

The Guessing Game

- Slightly unconventional test setup
- Give a test source image and caption as the starting point
- Regressed feature vector predicted should be closer to the feature vector of source image than other test images.
- **Metric**: What's the percentile of source image?



All agents forget, RL somewhat less

Model	MRR	R@5	R@10	Mean Rank
SL-pretrain	0.436	53.41	60.09	21.83
Frozen-Q	0.428	53.12	60.19	21.52
Frozen-f	0.432	53.28	60.11	21.54
RL-full-QAf	0.428	53.08	60.22	21.54
Frozen-Q-multi	**0.437**	**53.67**	**60.48**	**21.13**

Experiments on VizDial - how likely are the relevant responses

Human Study

- Show humans the dialog, ask them to guess the right image from a pool.
- Mean Rank; After RL training: 2.73
- Mean Rank; Before RL training: 3.70
- RL-trained agent's dialog are more indicative of image as per humans