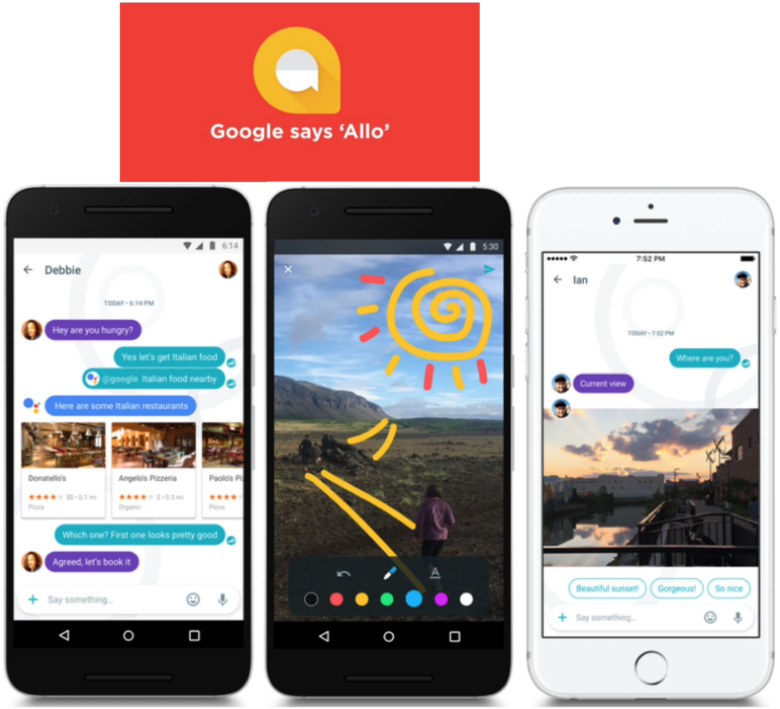
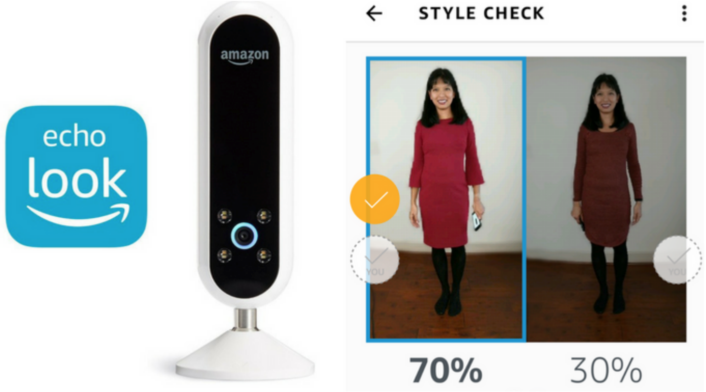
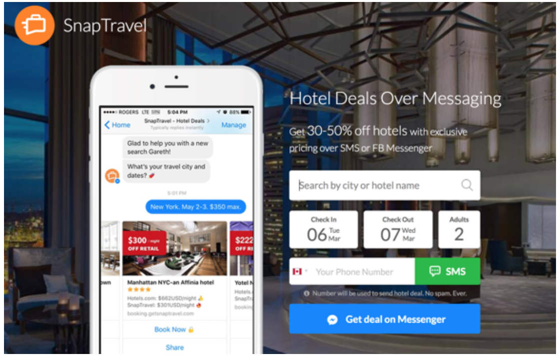
Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Maria Lubeznov
Weining Hu

Popularity on Virtual Assistant

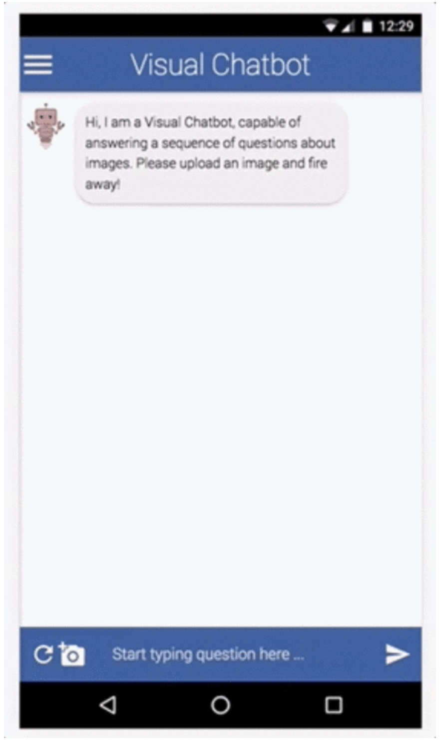






Motivation – Trend in Combining Vision and Language

1. Aiding visually impaired users in understanding their surroundings or social media content
2. Allowing medical personnel to better interpret medical scans
3. Helping AR/VR applications where a user could chat in natural language and work with a virtual companion who is seeing what they are seeing based on a visual common ground



Motivation – Trend in Combining Vision and Language

1. Aiding visually impaired users in understanding their surroundings or social media content
2. Allowing medical personnel to better interpret medical scans
3. Helping AR/VR applications where a user could chat in natural language and work with a virtual companion who is seeing what they are seeing based on a visual common ground

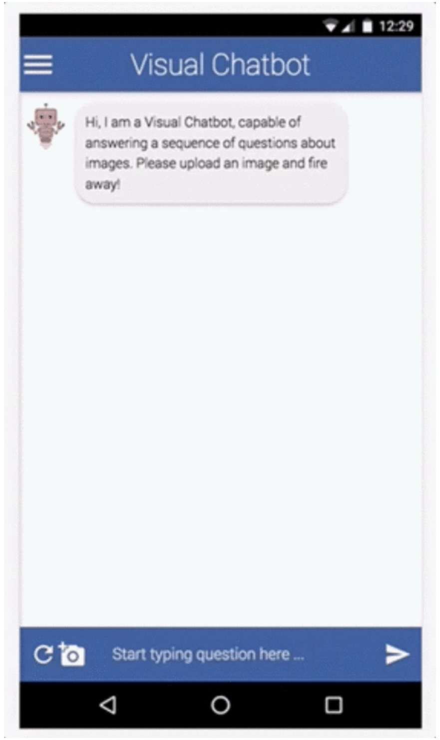


Image captioning vs VQA vs Visual dialog

Inputs

- Dialog history
- Image + caption
- Question

Output

- Answer

Captioning
Two people are in a wheelchair and one is holding a racket.

VQA
Q: How many people on wheelchairs ?
A: Two

Q: How many wheelchairs ?
A: One

Visual Dialog
Q: How many people are on wheelchairs ?
A: Two
Q: What are their genders ?
A: One male and one female
Q: Which one is holding a racket ?
A: The woman

Visual Dialog
Q: What is the gender of the one in the white shirt ?
A: She is a woman
Q: What is she doing ?
A: Playing a Wii game
Q: Is that a man to her right
A: No, it's a woman

Motivation2 - They are not perfect

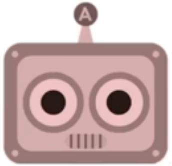
Vision+language recipe -> **DATASETS (Collected from Amazon mechanical turk)**

Treat dialog as static supervised learning -> **Training (Ground truth answer provided)**

These lead to two bad consequence!

Visual Dialog: Problem #1

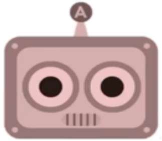
Q: How many people on wheelchairs?
A: Two
Q: What gender are the people in the wheelchairs?
A: One is female, one is male
Q: Which one is holding the racket?
A: The female



Model **can't steer conversation** and **doesn't get to see the future consequences of its utterances during training**

Visual Dialog: Problem #2

Q: How's the weather?



Ground truth: **Sunny**

Clear
Can't see the sky
Looks warm
It's not raining
I can't see the sky, but I see shadows, so probably sunny
…

Evaluation infeasible for utterances outside the dataset

Problem formulation

A-BOT

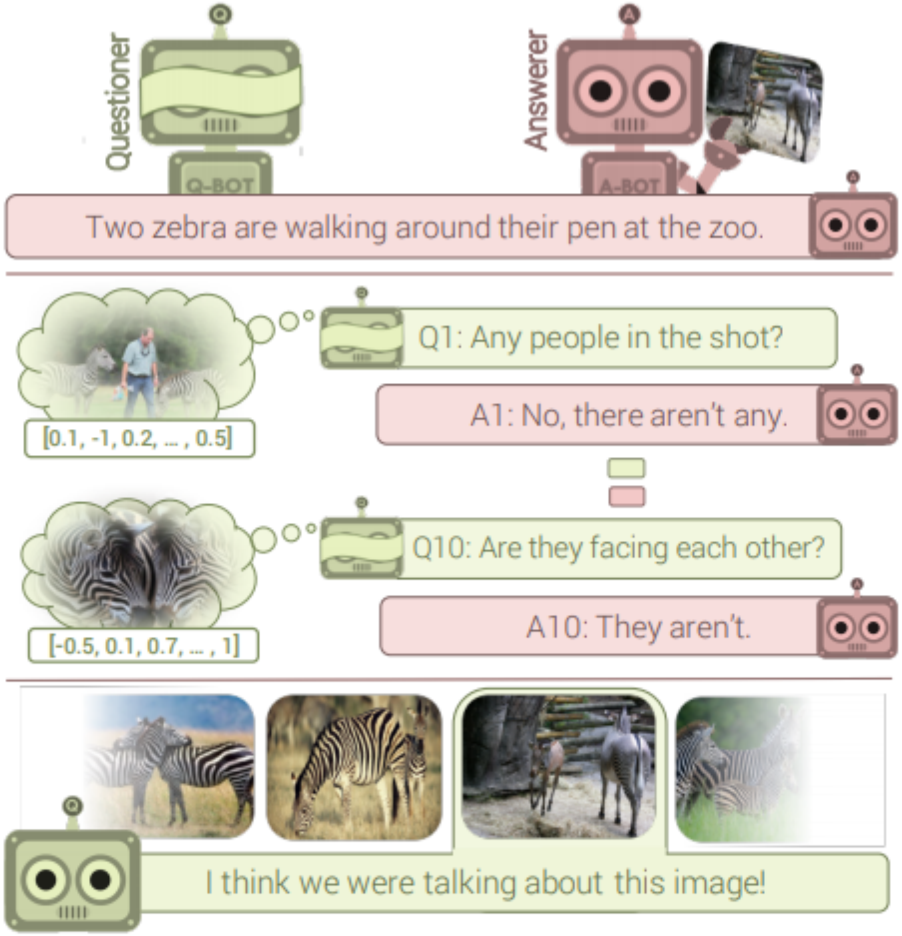
- Picks an image from dataset
- Answers question about the image

Q-BOT

- Asks questions
- Update its mental model of unseen image
- Makes prediction for image feature vector

RL reward for both agents

- how close the prediction to true image



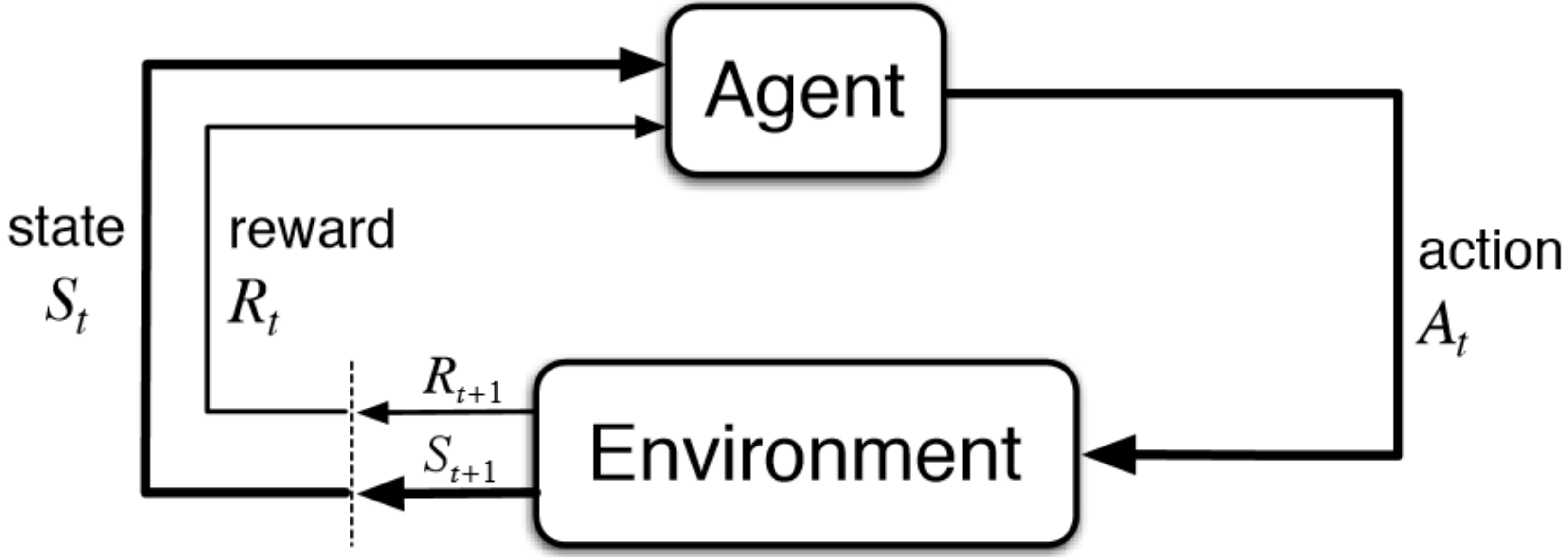
Visual dialog dataset - VisDial

VisDial

Training set (235M)
82,783 images
Validation set (108M)
40,504 images

```
'questions': [
        'does it have a doorknob',
        'do you see a fence around the bear',
        ...
      ],
      'answers': [
        'no, there is just green field in foreground',
        'countryside house',
        ...
      ],
      'dialogs': [
        {
          'image_id': <COCO image id>,
          'caption': <image caption from COCO>,
          'dialog': [
            {
              'question': <index of question in `data.questions` list>,
              'answer': <index of answer in `data.answers` list>,
              'answer_options': <100 candidate answer indices from data.answers`>,
              'gt_index': <index of `answer` in `answer_options`>
            },
            ... (10 rounds of dialog)
```

Reinforcement Learning for Dialog Agents



Reinforcement Learning for Dialog Agents

Agent: Cooperative A-Bot and Q-bot

State:

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

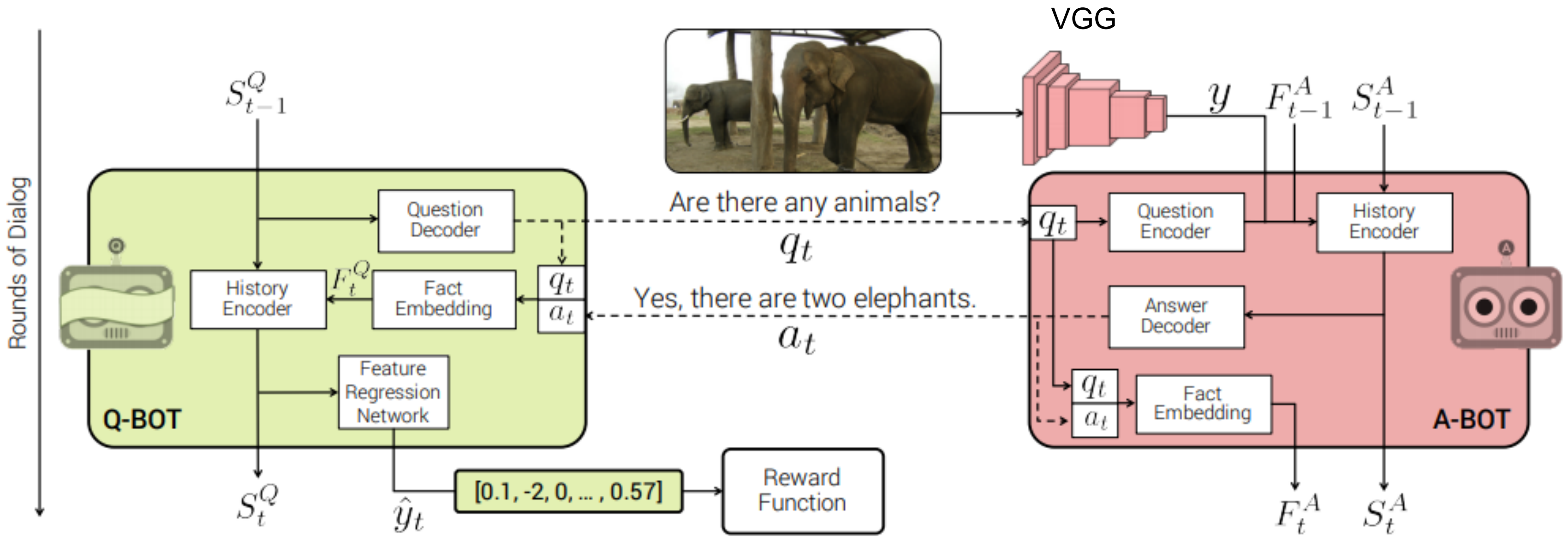
Action: question and answer pair

$$(q_t, a_t).$$

Reward:

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

Policy networks for Q –Bot and A-Bot



Policy network - Q-BOT

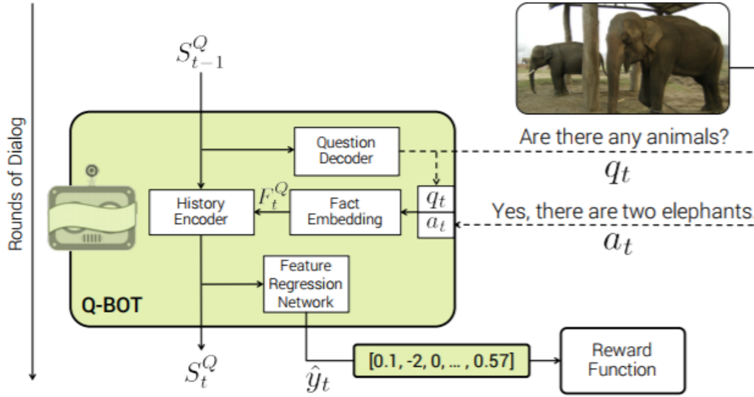
Fact encoder - LSTM $(q_t, a_t) \to F_t^Q \in \mathbb{R}^{512}$

State encoder - LSTM $(F_1^Q, \ldots, F_t^Q) \to S_t^Q$

Question decoder - LSTM which uses S_{t-1}^Q as initial hidden state and generates q_t by sequentially sampling words.

Feature regression network - 1 FC layer $\hat{y}_t = f(S_t^Q)$

Parameters - θ_Q (LSTMs) and θ_f (FC layer)



Policy network - A-BOT

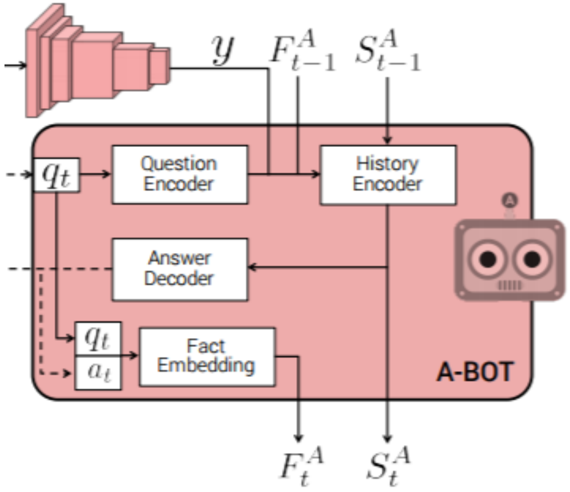
Question encoder - LSTM $q_t \rightarrow Q_t^A \in \mathbb{R}^{512}$

Fact encoder - same as in Qbot.

State encoder - LSTM $\left((y, Q_1^A, F_0^A), \ldots, (y, Q_t^A, F_{t-1}^A)\right) \rightarrow S_t^A$

Answer decoder - LSTM which uses S_t^A as initial hidden state and generates a_t by sequentially sampling words

Parameters - θ_A (LSTMs)



Training

1. Supervised pre training on VisDial
    a. Faster RL convergence
    b. To prevent the bots from inventing their own uninterpretable language
2. Fine tuning with RL - parameters are updated in response to experienced rewards.

Supervised training

BOTs are trained separately.

Q-BOT

- Is trained to generate the follow-up question by the questionnaire, given the caption and the QA history.

A-BOT

- Is trained to generate the response by the answerer, given the image, caption and the QA history.

CNN for image feature extraction - pre trained on ImageNet

Reinforcement fine tuning

Objective: maximize the expected reward, under agents' policies, for each dialog round (time t).

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

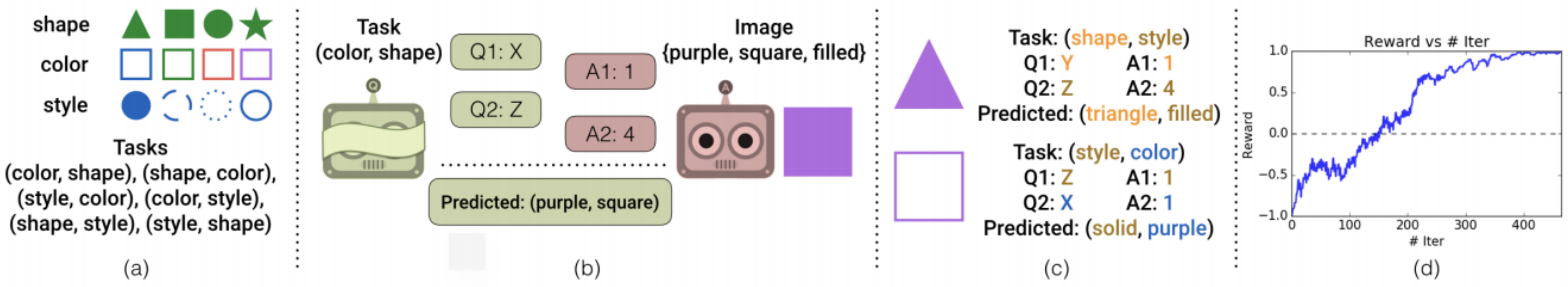
REINFORCE algorithm is applied to calculate the $\nabla_{\theta_Q} J$ and $\nabla_{\theta_A} J$.

Intuition:

- Informative QA round → positive reward → positive gradient(probability)
- Poor exchange → negative reward → negative gradient(probability)

θ_f is updated in a "supervised" way.

Synthetic example



64 images.

$V_A = \{1, 2, 3, 4\} \quad V_Q = \{X, Y, Z\}$ = {shape, color, style}

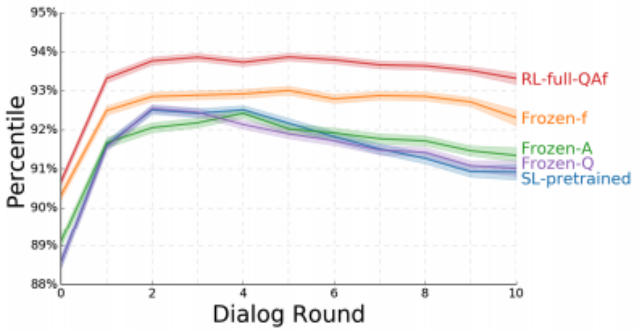
RL iteration = 10K episodes

Results - supervised vs RL dialogues

Image + Caption	Human-Human dialog [4]		SL-pretrained Q-BOT-A-BOT dialog		RL-full-QAf Q-BOT-A-BOT dialog	
 A couple of people in the snow on skis.	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 cracker with shredded cheese and leafy greens on a white plate.	Are there any people in picture?	No, there are no people	Are there any people?	No	Is plate large or small?	Large
	Is cheese white?	Yes, cheese is white	What color is plate?	White	Is it on table?	I can't tell
	Are crackers white?	No, cracker is brownish	Is plate on table?	I can't tell	What kind of donut?	Plain with sugar on it
	Are there lots of greens?	No, only 1 sprig	What color is plate?	White	What kind of chips?	Potato chips
	Was picture taken in kitchen?	I cannot tell, it is too zoomed in	Is plate on table?	I can't tell	Any people?	No
	How many crackers are there?	There is 1	Is there anything else on table?	No	Is this inside or outside?	I can't tell, it's close up
	Is there design on plate?	No, plate is plane white	Is there anything else on table?	No	What color is table?	White
	Is there good lighting?	Yes, lighting is decent	Is there anything else on table?	No	Any chairs?	No
	Is plate on table?	I cannot tell I can only see plate as background	Is there anything else on table?	No	What color is table?	I can't tell
	Does cracker look delicious?	Absolutely	Is there anything else on table?	No	Is it large or small?	I can't see plate

Evaluation

Dataset: test split for VisDial v0.5 (~9.5k images)
Input : Image + caption
Rounds: 10
Output: Sort the entire test set in ascending distance to this prediction

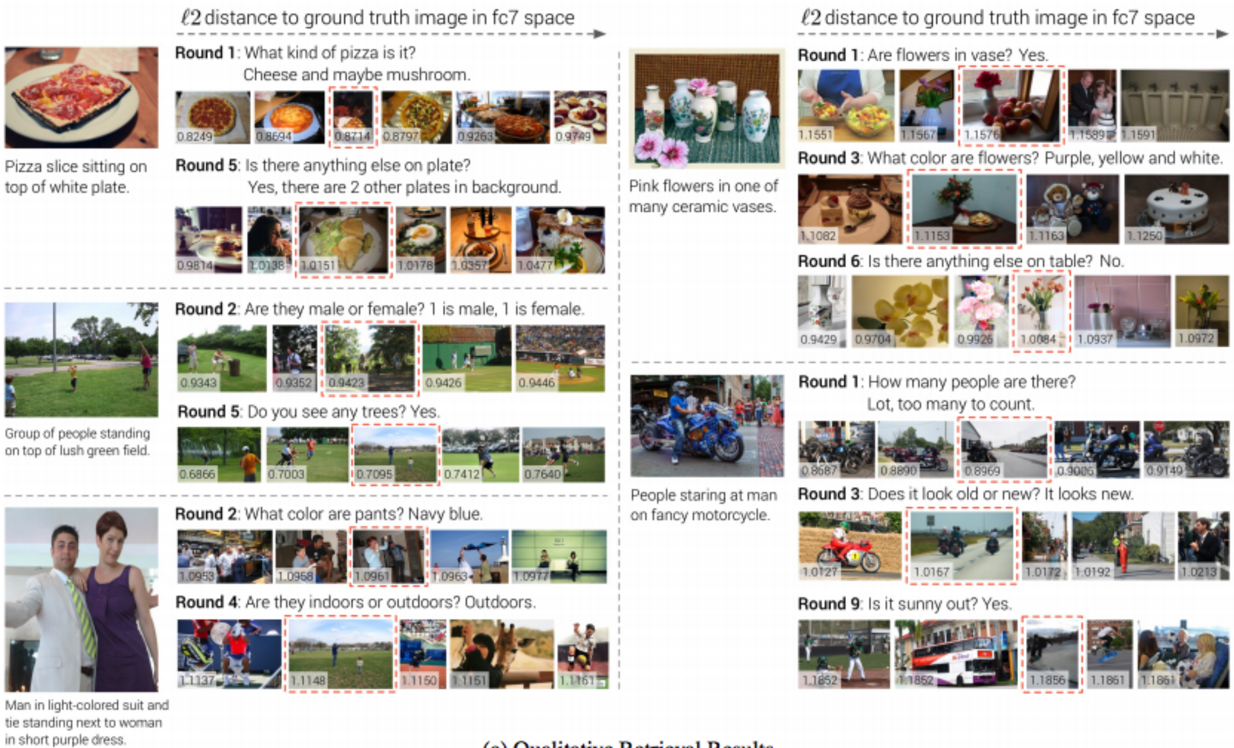


(a) Guessing Game Evaluation.

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**

(b) Visual Dialog Answerer Evaluation.

Evaluation



(c) Qualitative Retrieval Results.

Weakness of the evaluation

1. No evaluation on the response time
2. No evaluation on how the initial caption quality affect the later dialog

Strengths and weaknesses of the approach

Strengths

- Self talk → unlimited data. No need to collect very large datasets
- Evaluation is image guessing → no need to evaluate natural language generation.
- Method is agent driven - learning to deal with actions consequences

Weaknesses

- No explanation of why the agent becomes less sure after 2 dialog rounds.
- Authors mention that A-BOT responses are not enough “human like”.

Potential extensions

Our thoughts

- Address repetitive questions
- It may be a good idea to combine this method with some sort of CGAN from paper presented last week, to better emulate human dialogs.
- Evaluate the impact of different image captioning quality on final reward achievement

Follow up papers

- Evaluate if and how this method can be applied on human-machine interaction ([1]).
- Investigate how similar method can be applied on other tasks that require negotiation ([2])

References

[1] Evaluating Visual Conversational Agents via Cooperative Human-AI Games

[2] Deal or No Deal? End-to-End Learning for Negotiation Dialogues

[3] Demo

[3] https://research.fb.com/visual-reasoning-and-dialog-towards-natural-language-conversations-about-visual-data/

Appendix - technical details

- Supervised training for the first K rounds of dialog and transition to RL for the remaining $10 - K$ rounds
- In each epoch, K gradually annealed to 0.
- 15 epochs
- Adam optimizer with learning rate 0.001
- Gradients are clamped to [-5,5] to avoid explosion.