Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, Dhruv Batra

IEEE ICCV 2017

Presented By:
Nalin Chhibber
nalin.chhibber@uwaterloo.ca

CS 885: Reinforcement Learning
Pascal Poupart

UNIVERSITY OF WATERLOO

Outline

- Introduction
- Paper overview
- Contribution and key takeaways
- Critique
- Class discussion

Introduction

Problem Space: Intersection of **Vision** and **Language**

- Image Captioning
  - Predict one sentence description of an image.
- Visual Question Answering
  - Predict a natural language answer given an image and a question.
- Visual Dialog
  - Predict a free-form NL answer given an image, a dialog history, and a follow-up question.

UNIVERSITY OF WATERLOO

Paper Overview

Focused on creating a visually-grounded conversational artificial intelligence (AI)

Develop AI agents that can

- See (understand contents of an image)
- Communicate (understand and hold a dialog in natural language)

Applications:

- Help visually impaired users understand their surroundings
- Enable analysts to sift through large quantities of surveillance data

UNIVERSITY OF WATERLOO

Paper Overview

Most of the previous work treat this as a static supervised learning problem

Problem-1

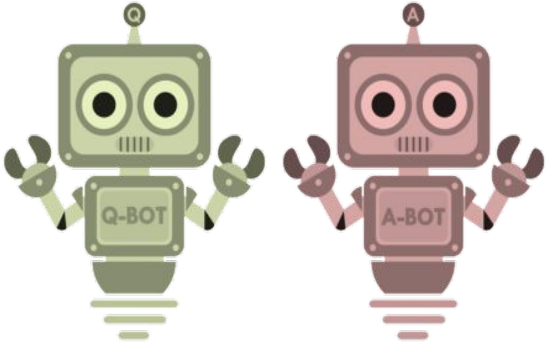
Model cannot steer conversation and doesn't get to see the future consequences of its utterances during training.

Problem-2

Evaluations are infeasible for utterances outside the dataset.

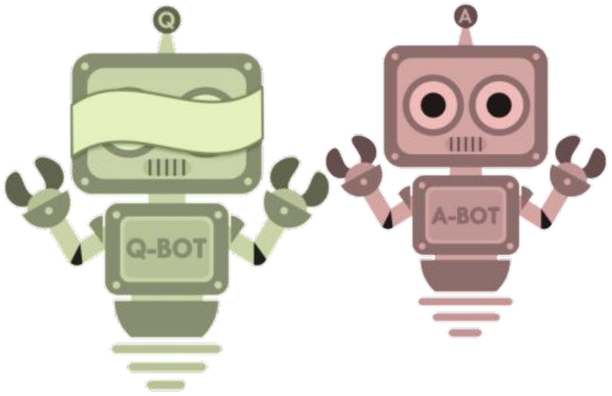
Guess Which

An image guessing game between Q-Bot and A-Bot



Guess Which

An image guessing game between Q-Bot and A-Bot

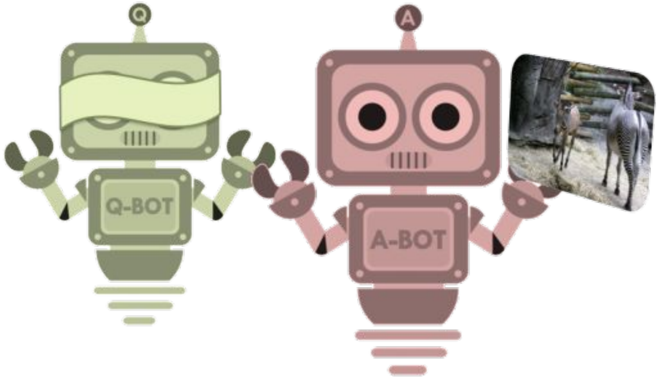


Q-Bot

Questioning Agent
Blind-folded

Guess Which

An image guessing game between Q-Bot and A-Bot



A-Bot

Answering Agent
Access to secret image

Training types

Conducted two types of demonstration with

- Completely ungrounded synthetic world (RL from scratch)
  - Agents communicate via symbols with no pre-specified meanings.

- Large-scale experiment on real images using VisDial dataset
  - Pretrain on dialog data with SL, followed by fine-tuning with RL.

Reinforcement Learning Framework

	Q-Bot	A-Bot
		
Environment	**Other agent + set of all images**	**Other agent + secret image**
Action	**Set of all possible questions**	**Set of valid answers**
State	$s_t=[c, q_1a_1, q_2a_2..q_{t-1}a_{t-1}]$	$s_t=[I, c, q_1a_1, q_2a_2..q_{t-1}a_{t-1}, q_t]$

Reward

Change in distance to the true representation before/after a round of dialog

Training Details

1. Pretrained with supervised learning on Visual Dialog dataset (VisDial)
2. Fine-tuned with REINFORCE

Training Details

1. Pretrained with supervised learning on Visual Dialog dataset (VisDial)
2. Fine-tuned with REINFORCE

Curriculum Learning

Problem: Discrete change in learning landscape

Solution: Gently hand over control to reinforcement learning

Training Details

1. Pretrained with supervised learning on Visual Dialog dataset (VisDial)
2. Fine-tuned with REINFORCE

Curriculum Learning

Problem: Discrete change in learning landscape

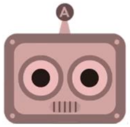
Solution: Gently hand over control to reinforcement learning

Reward Shaping

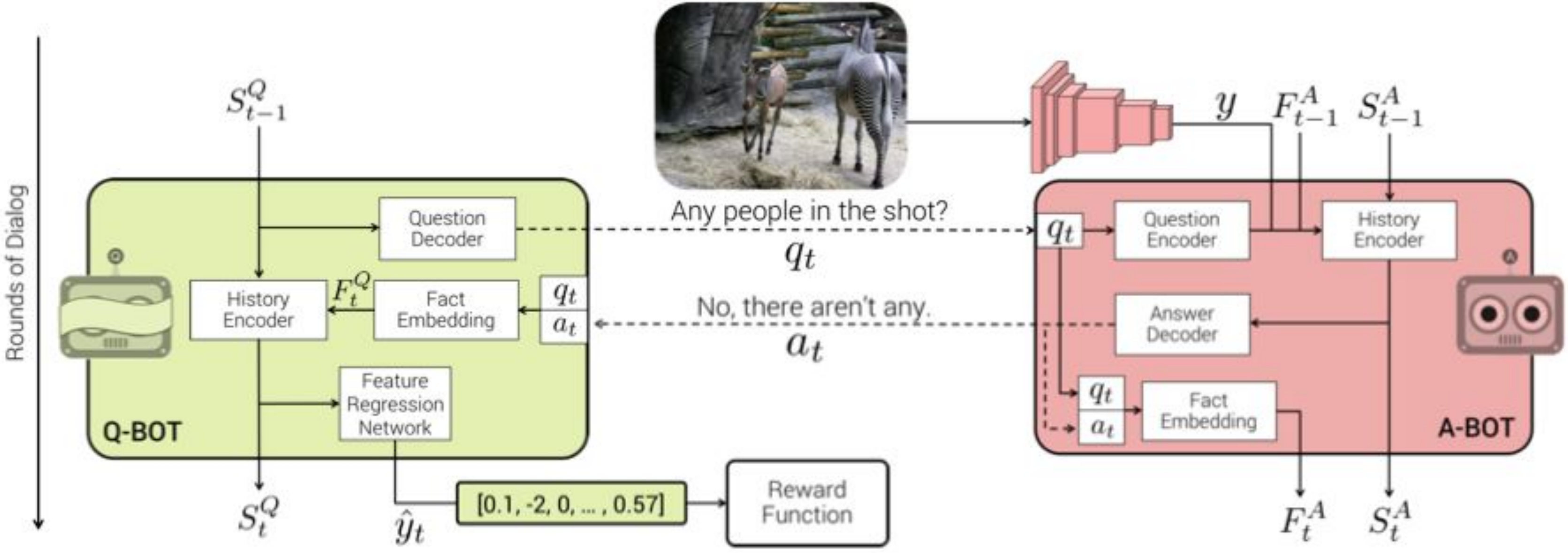
Problem: Delayed reward

Solution: Improvement-based intermediate rewards

Model Internals

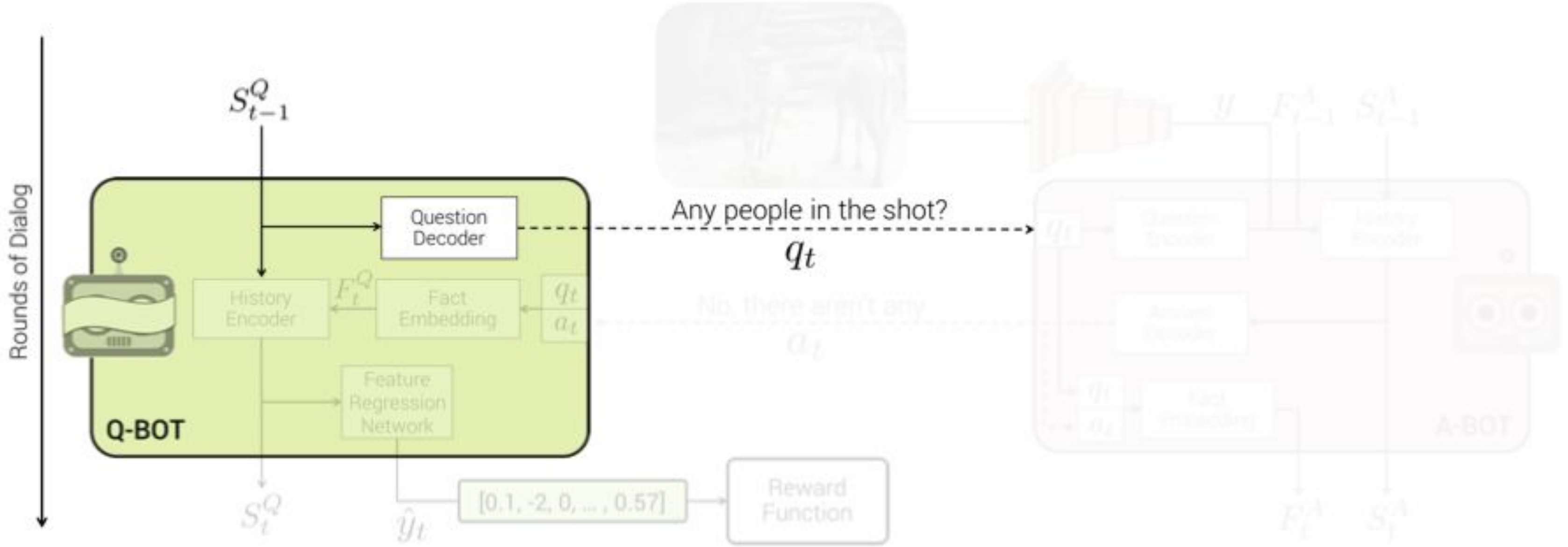


Model Internals



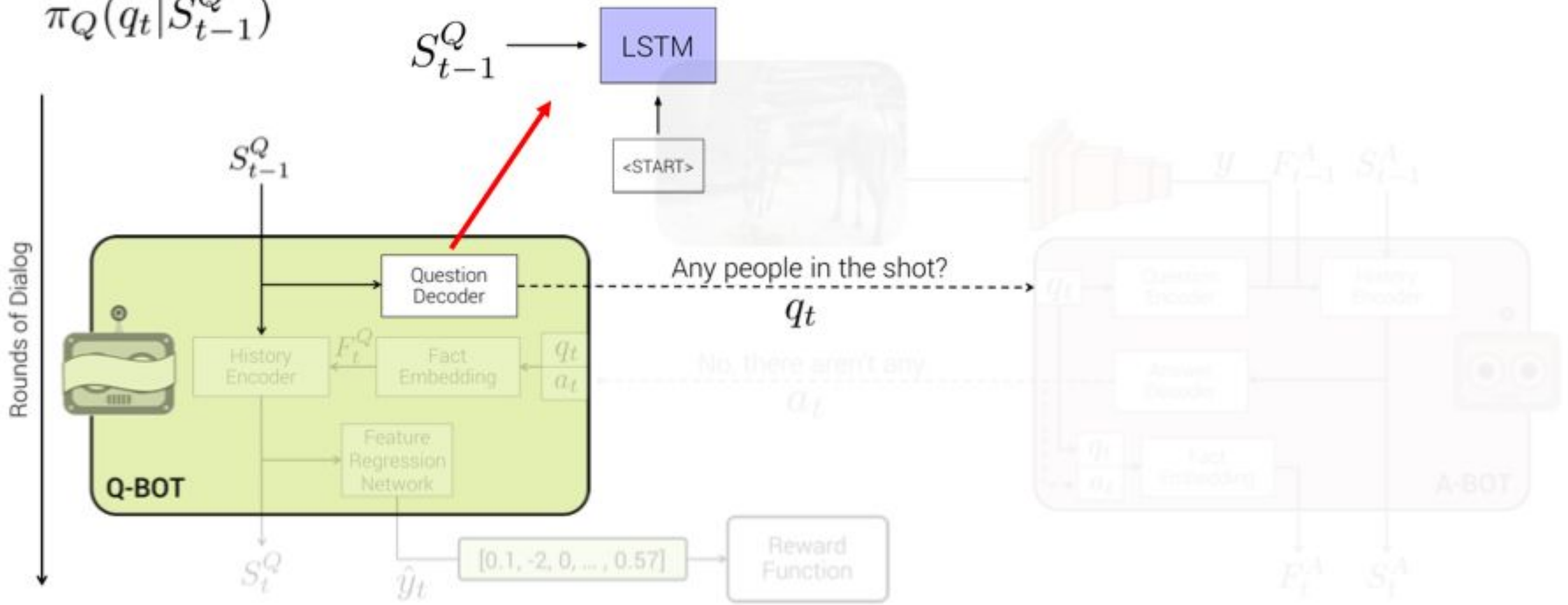
Q-Bot

$$\pi_Q(q_t|S^Q_{t-1})$$



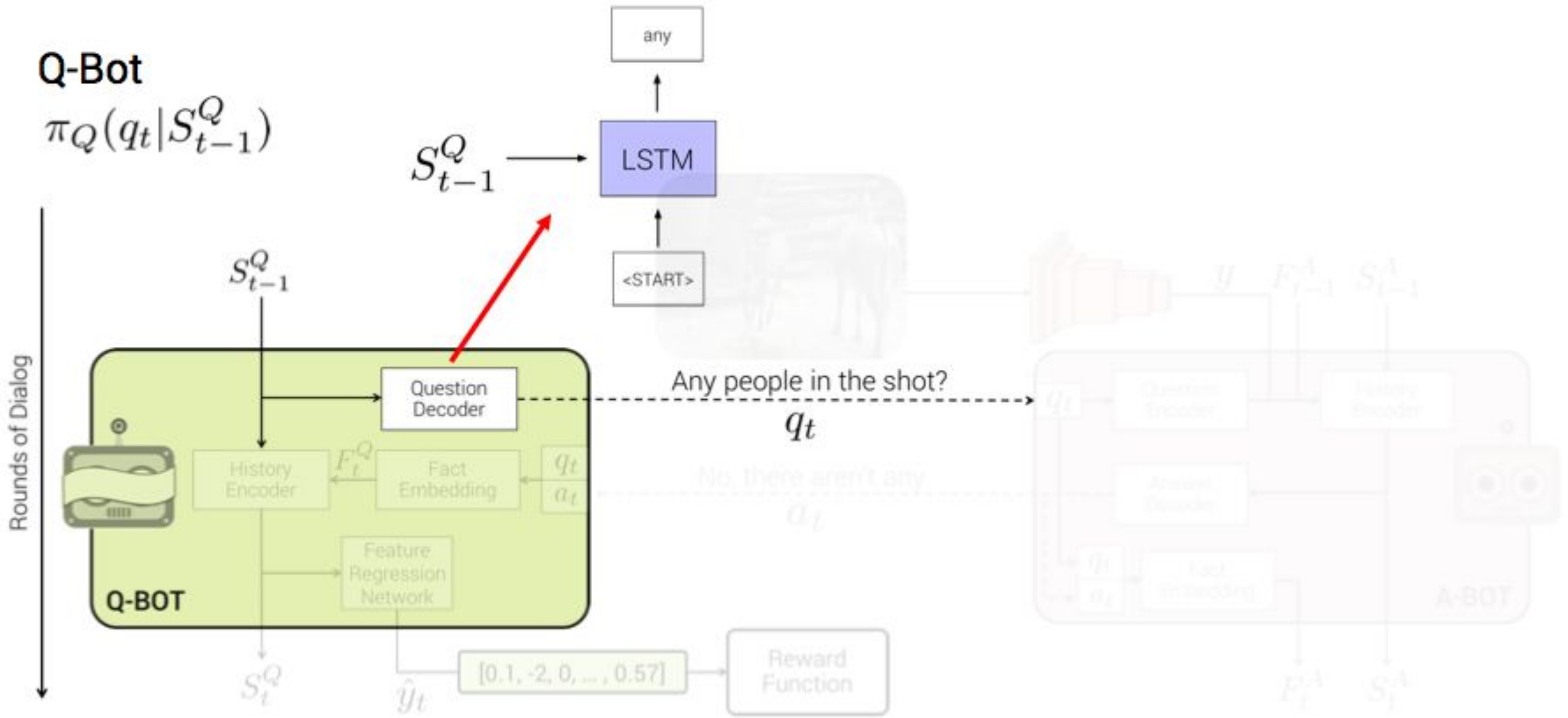
Q-Bot

$$\pi_Q(q_t | S^Q_{t-1})$$



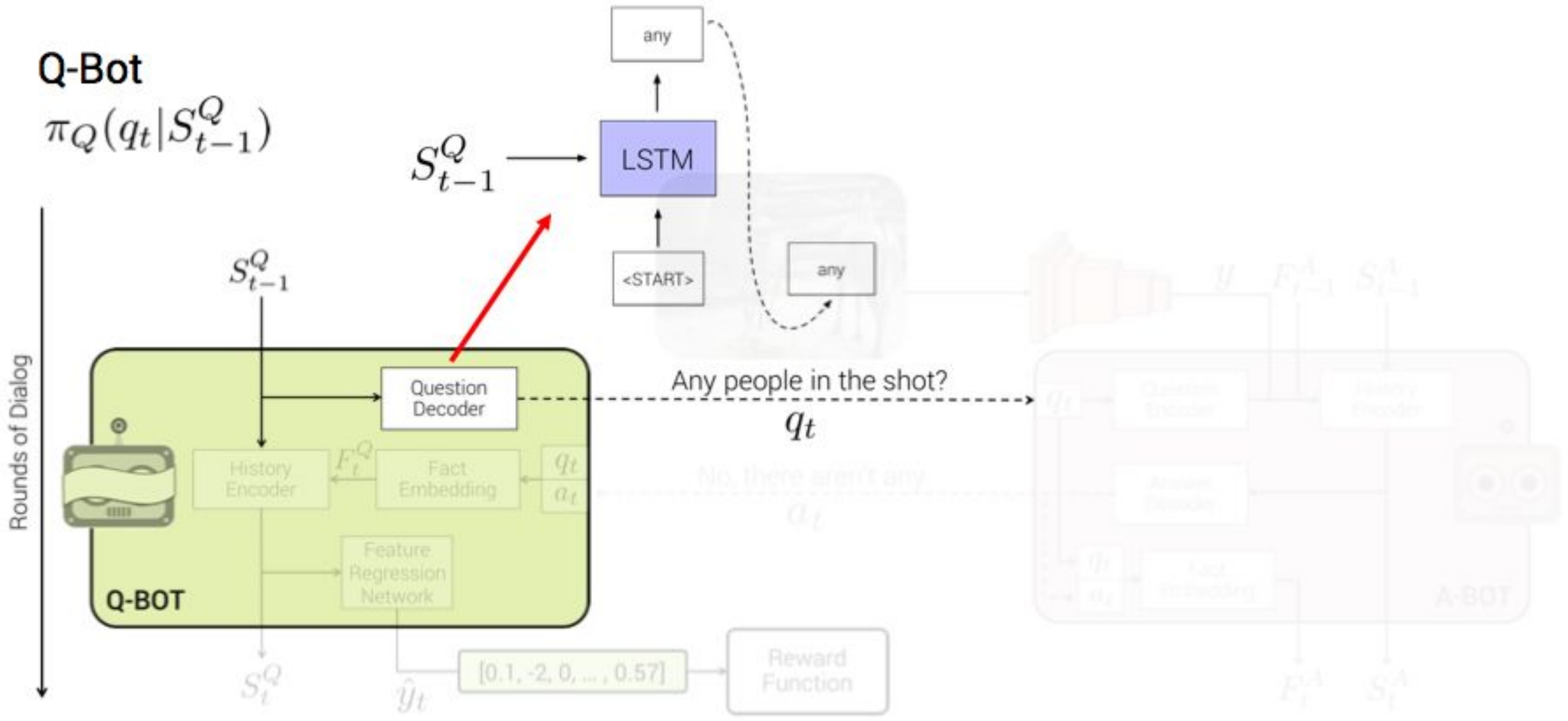
Q-Bot

$$\pi_Q(q_t|S^Q_{t-1})$$



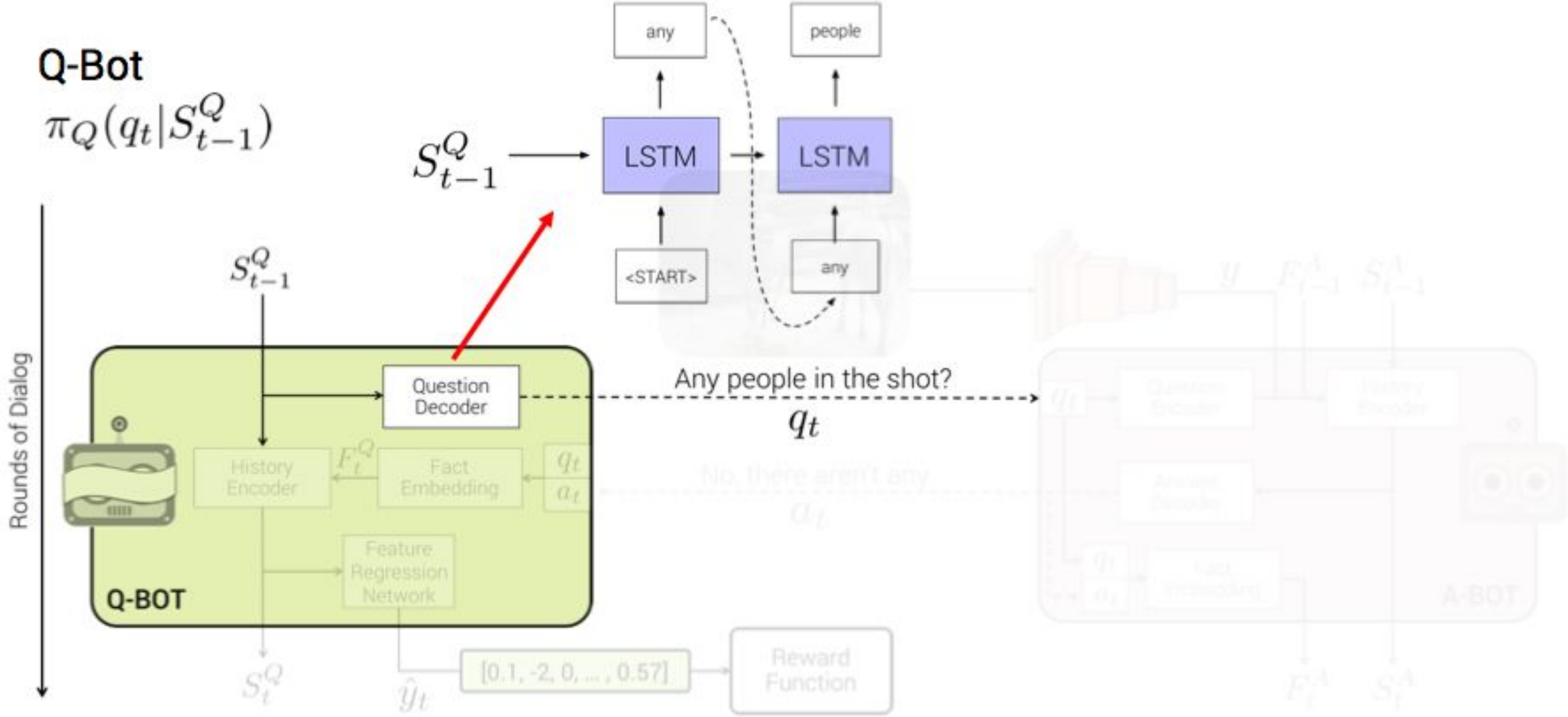
Q-Bot

$$\pi_Q(q_t|S^Q_{t-1})$$



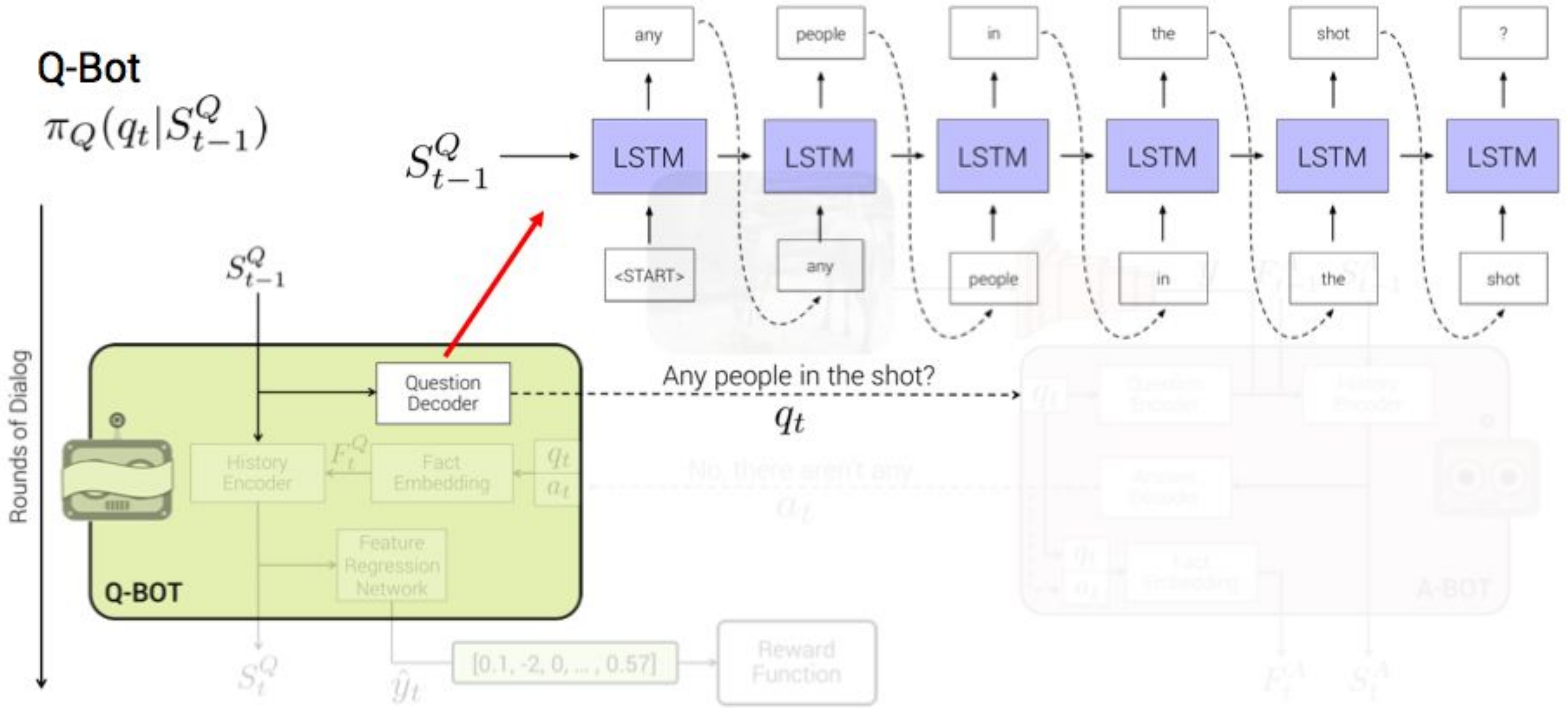
Q-Bot

$\pi_Q(q_t|S^Q_{t-1})$



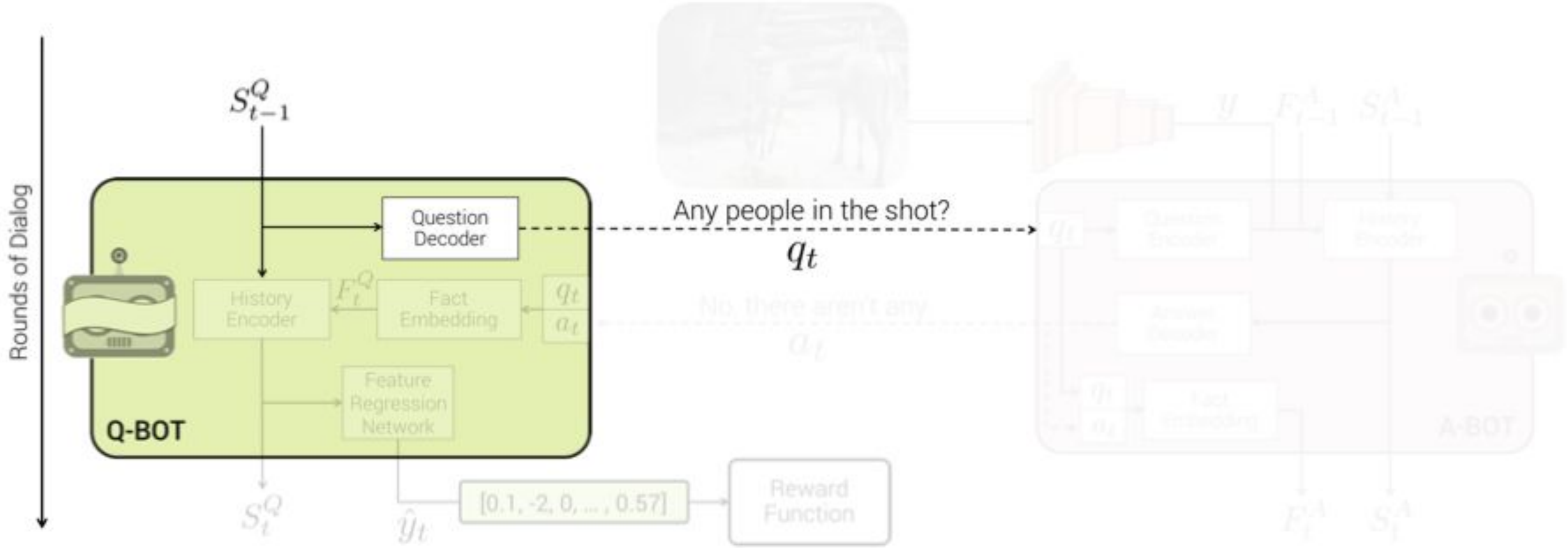
Q-Bot

$\pi_Q(q_t|S^Q_{t-1})$



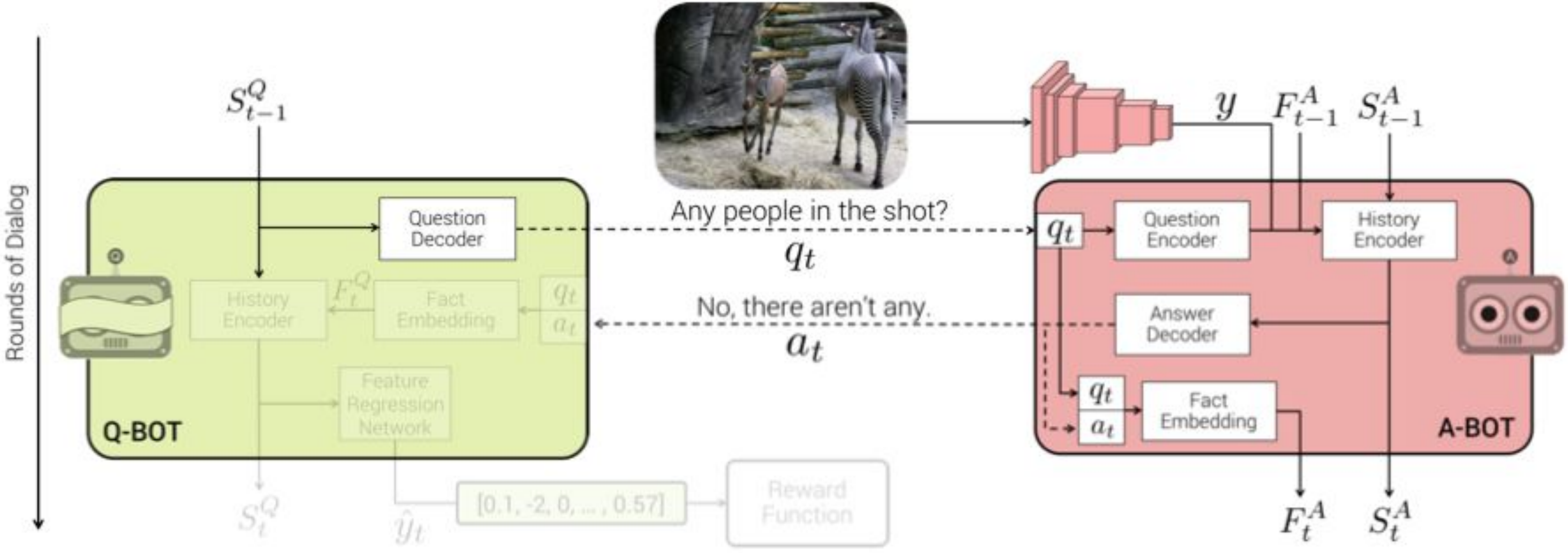
Q-Bot

$$\pi_Q(q_t|S_{t-1}^Q)$$



Q-Bot $\pi_Q(q_t|S^Q_{t-1})$

A-Bot $\pi_A(a_t|S^A_{t-1})$



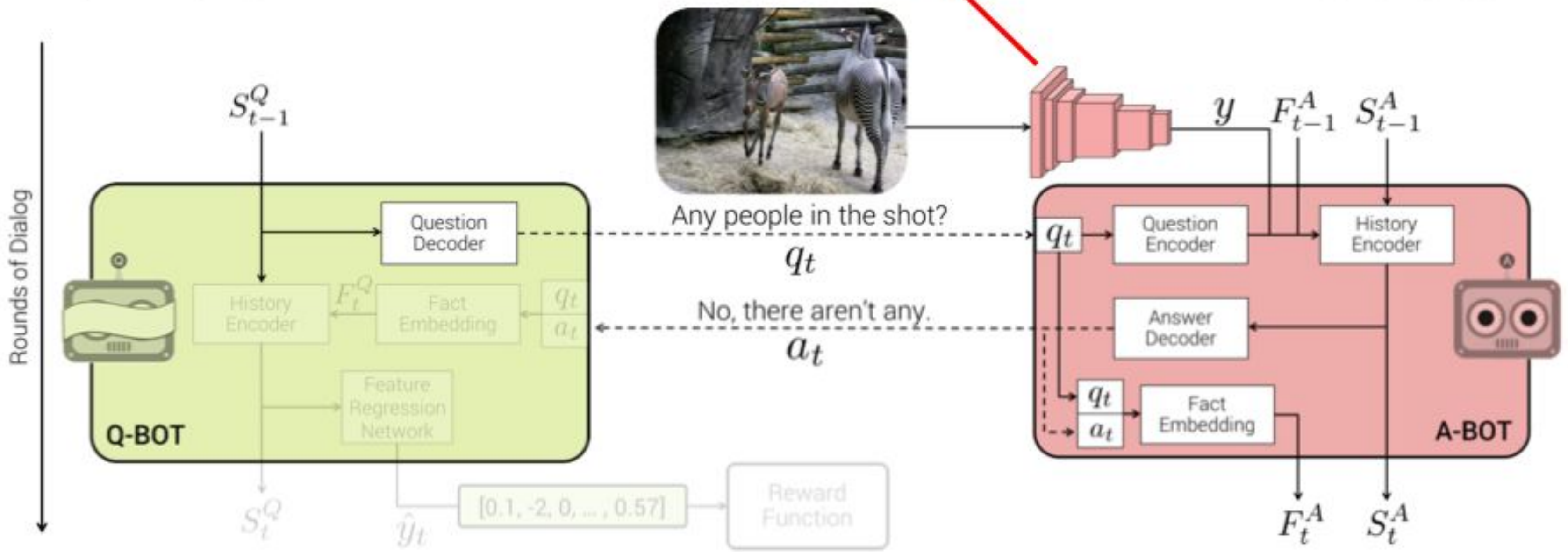
Q-Bot

$\pi_Q(q_t|S_{t-1}^Q)$

A-Bot

$\pi_A(a_t|S_{t-1}^A)$

VGG-16

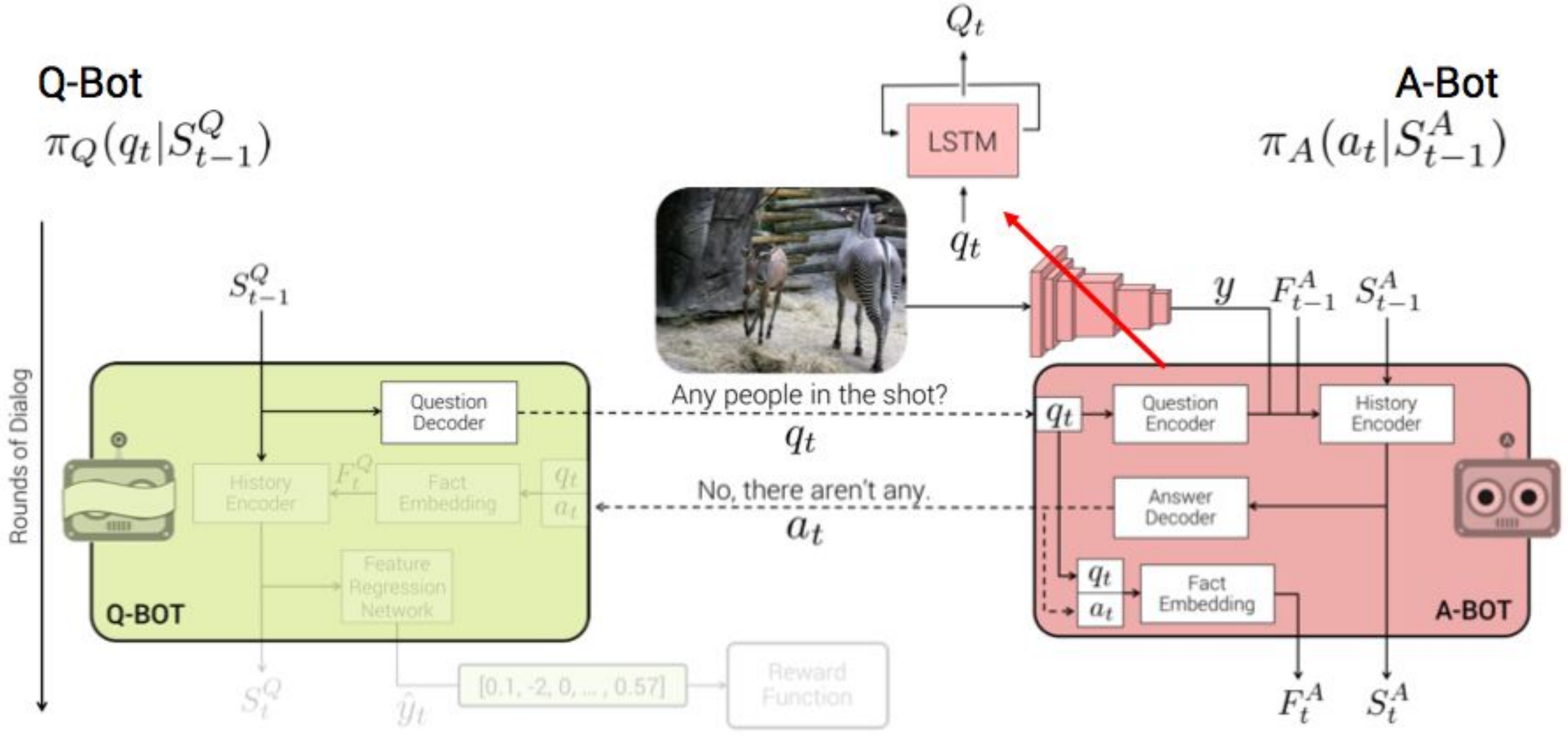


Q-Bot

$\pi_Q(q_t|S_{t-1}^Q)$

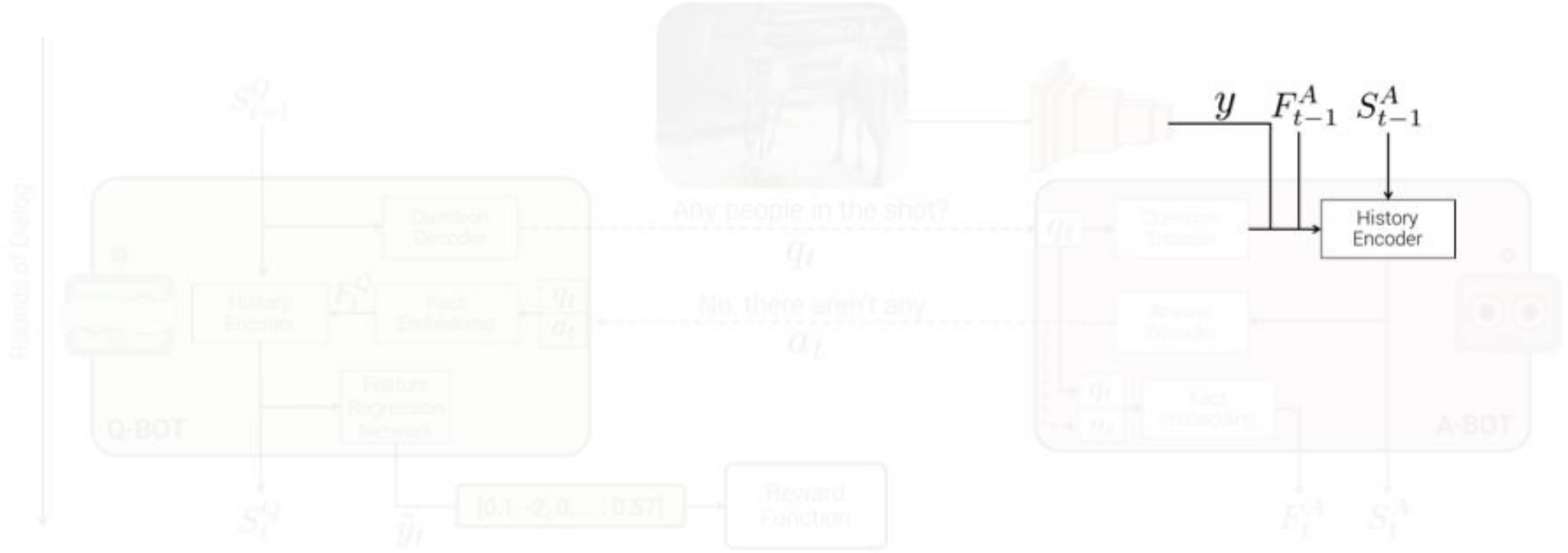
A-Bot

$\pi_A(a_t|S_{t-1}^A)$



A-Bot

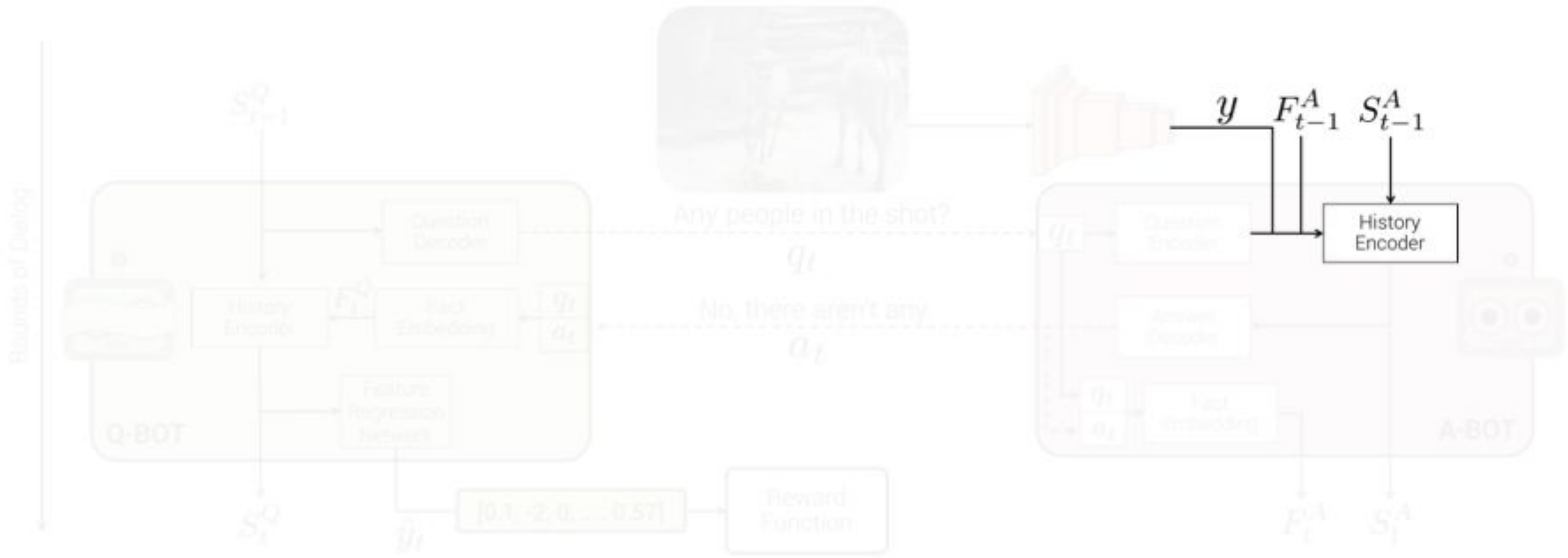
$$\pi_A(a_t|S^A_{t-1})$$



Fact Embedding

A-Bot

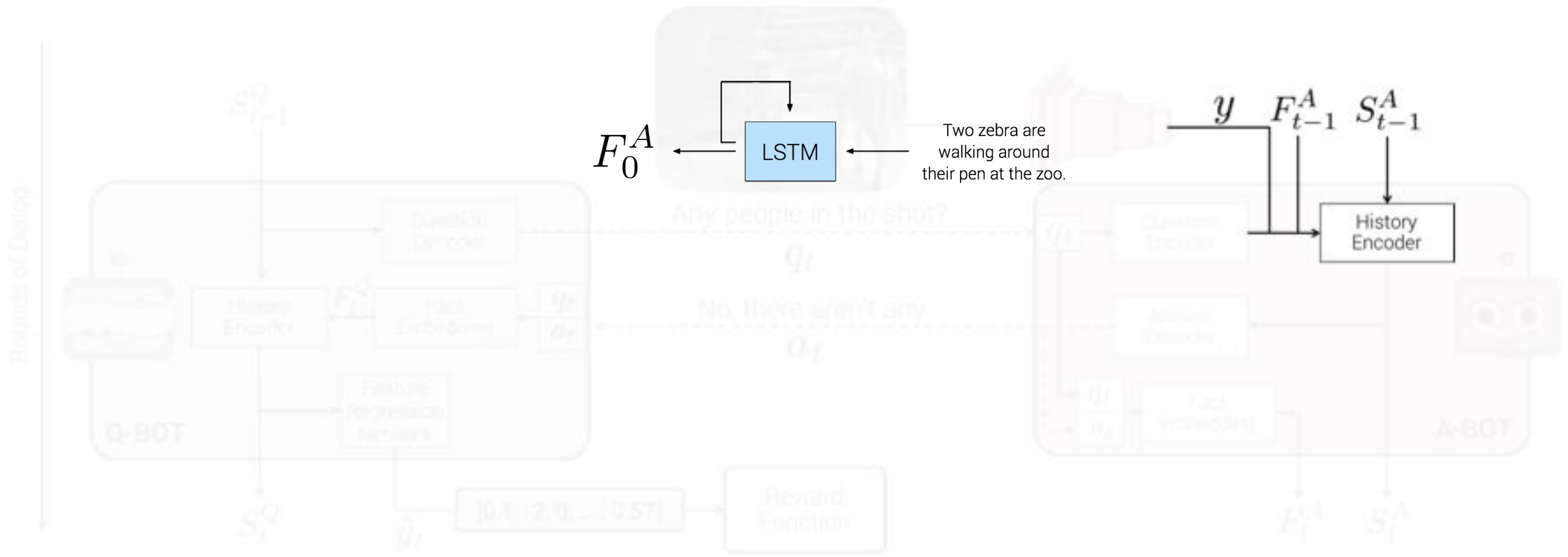
$$\pi_A(a_t|S^A_{t-1})$$



Fact Embedding

A-Bot

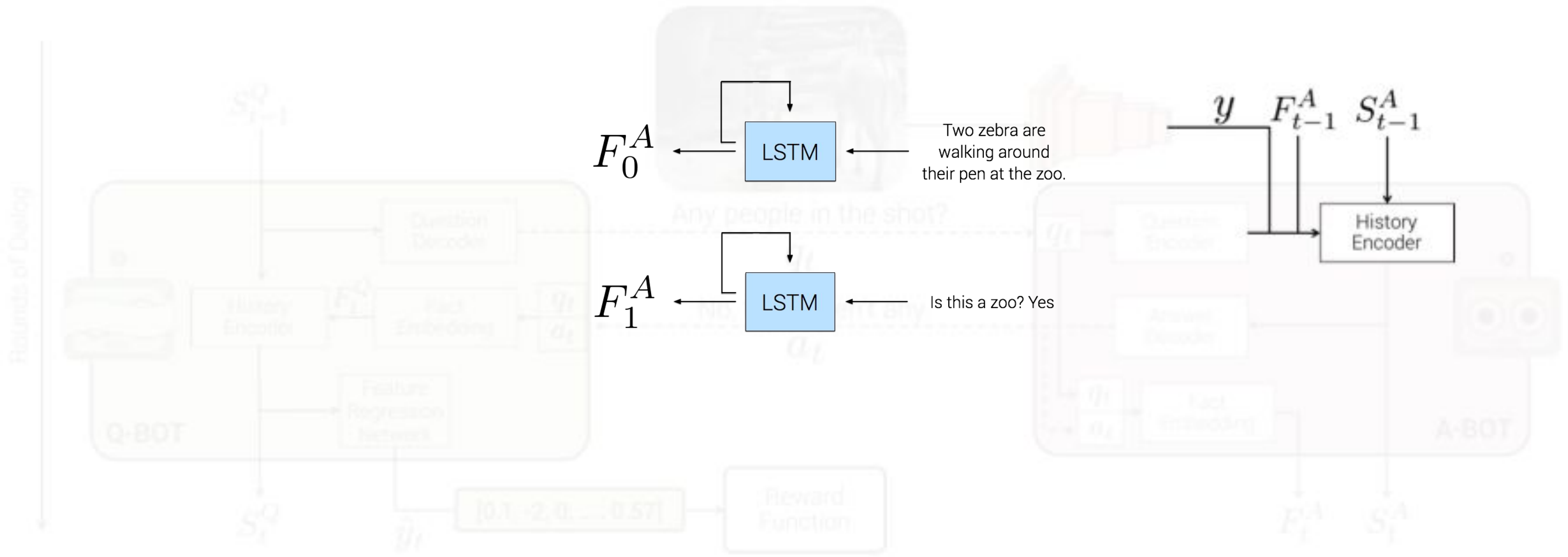
$$\pi_A(a_t|S_{t-1}^A)$$



Fact Embedding

A-Bot

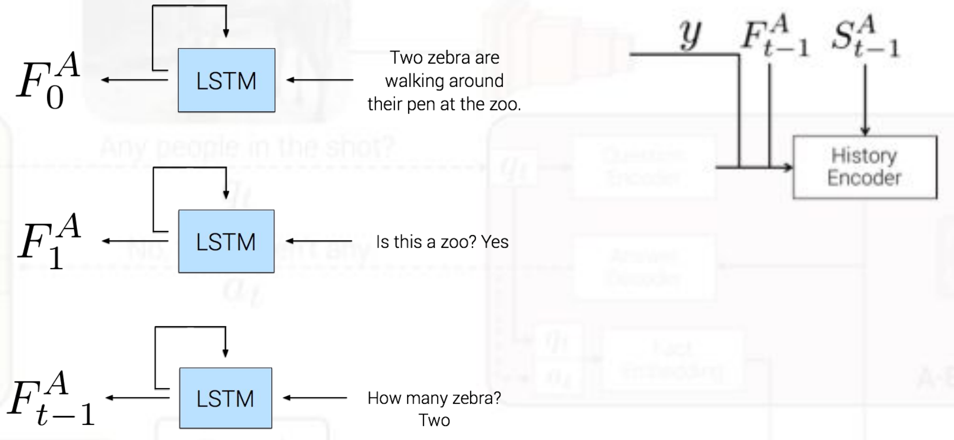
$$\pi_A(a_t|S_{t-1}^A)$$



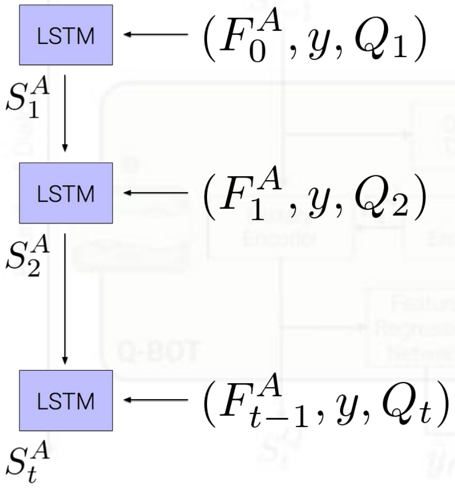
Fact Embedding

A-Bot

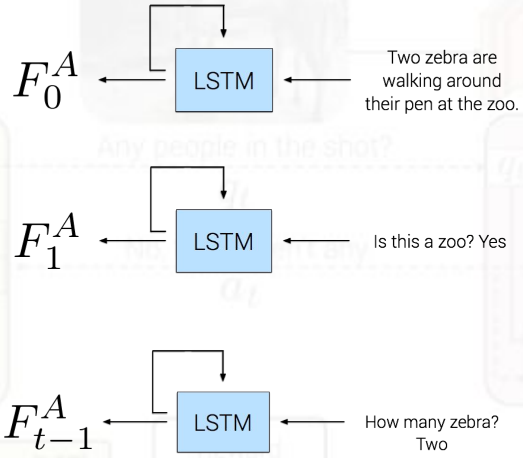
$\pi_A(a_t|S_{t-1}^A)$



History Encoding

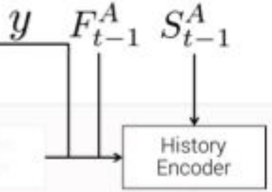


Fact Embedding



A-Bot

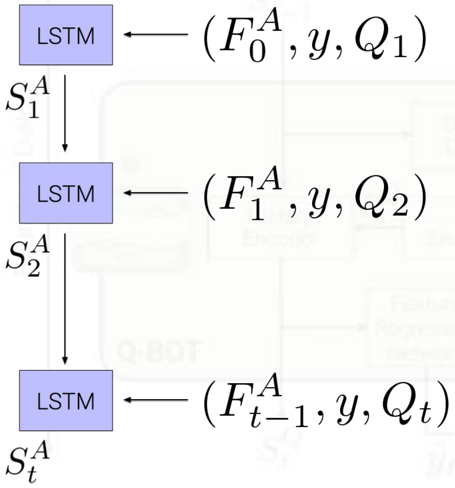
$$\pi_A(a_t|S_{t-1}^A)$$

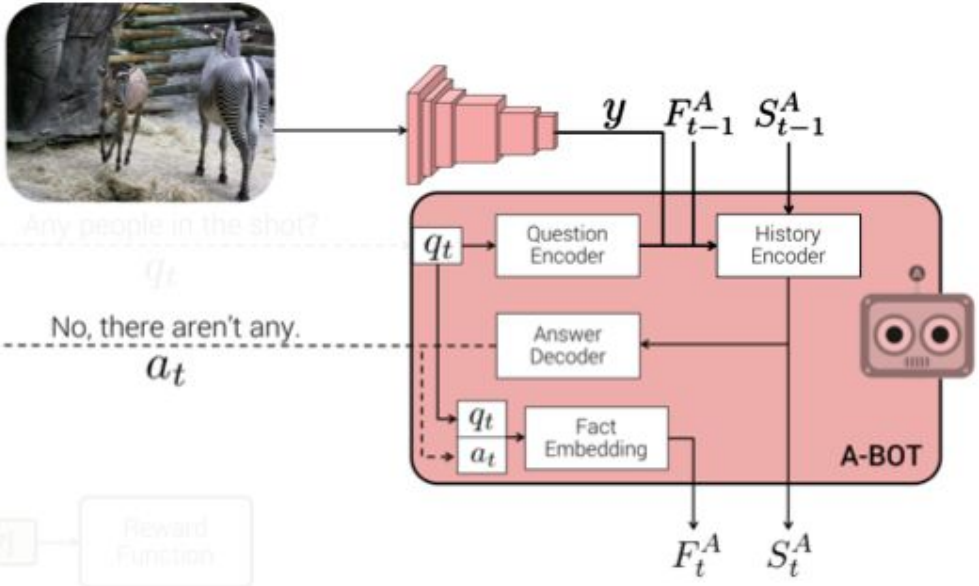


History Encoding

A-Bot

$$\pi_A(a_t | S^A_{t-1})$$



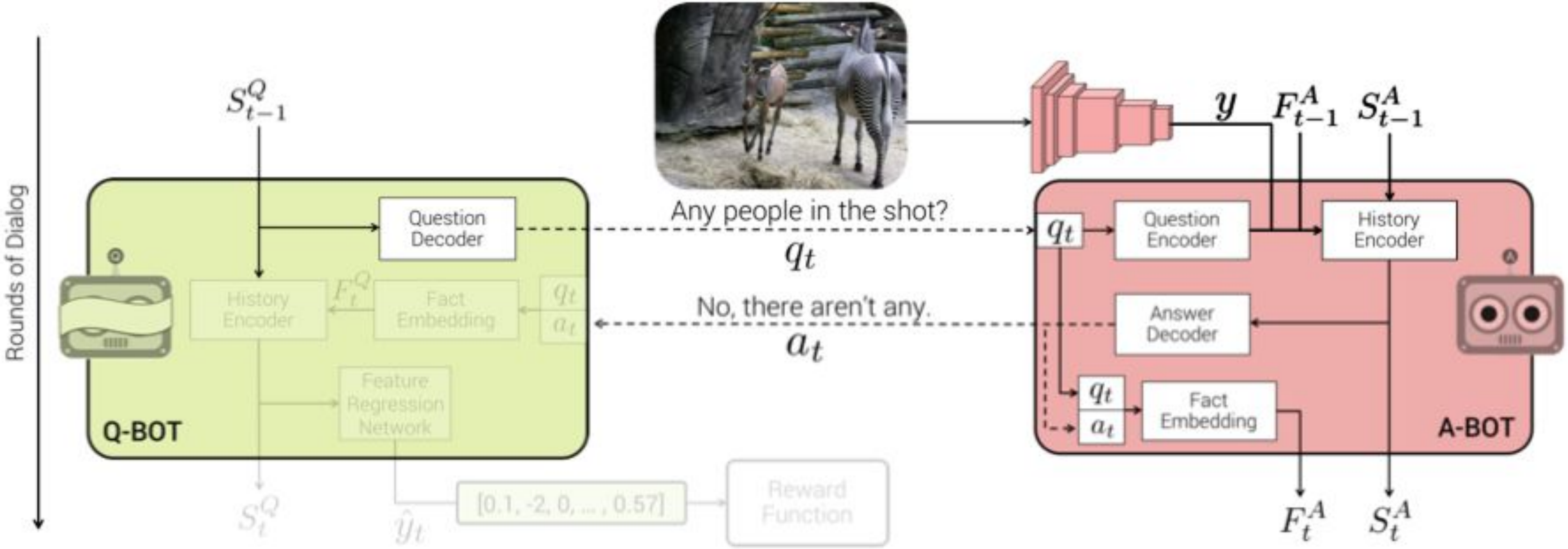


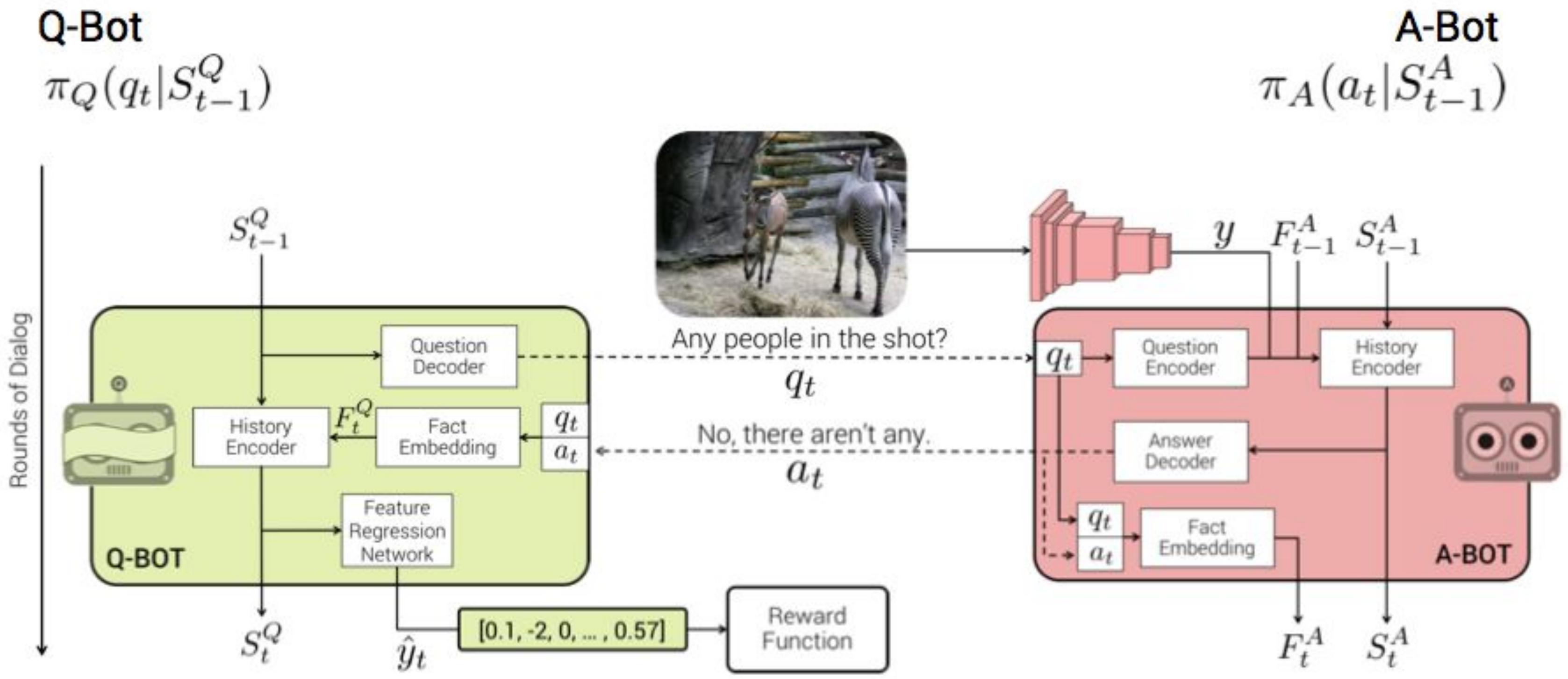
Q-Bot

$\pi_Q(q_t|S^Q_{t-1})$

A-Bot

$\pi_A(a_t|S^A_{t-1})$





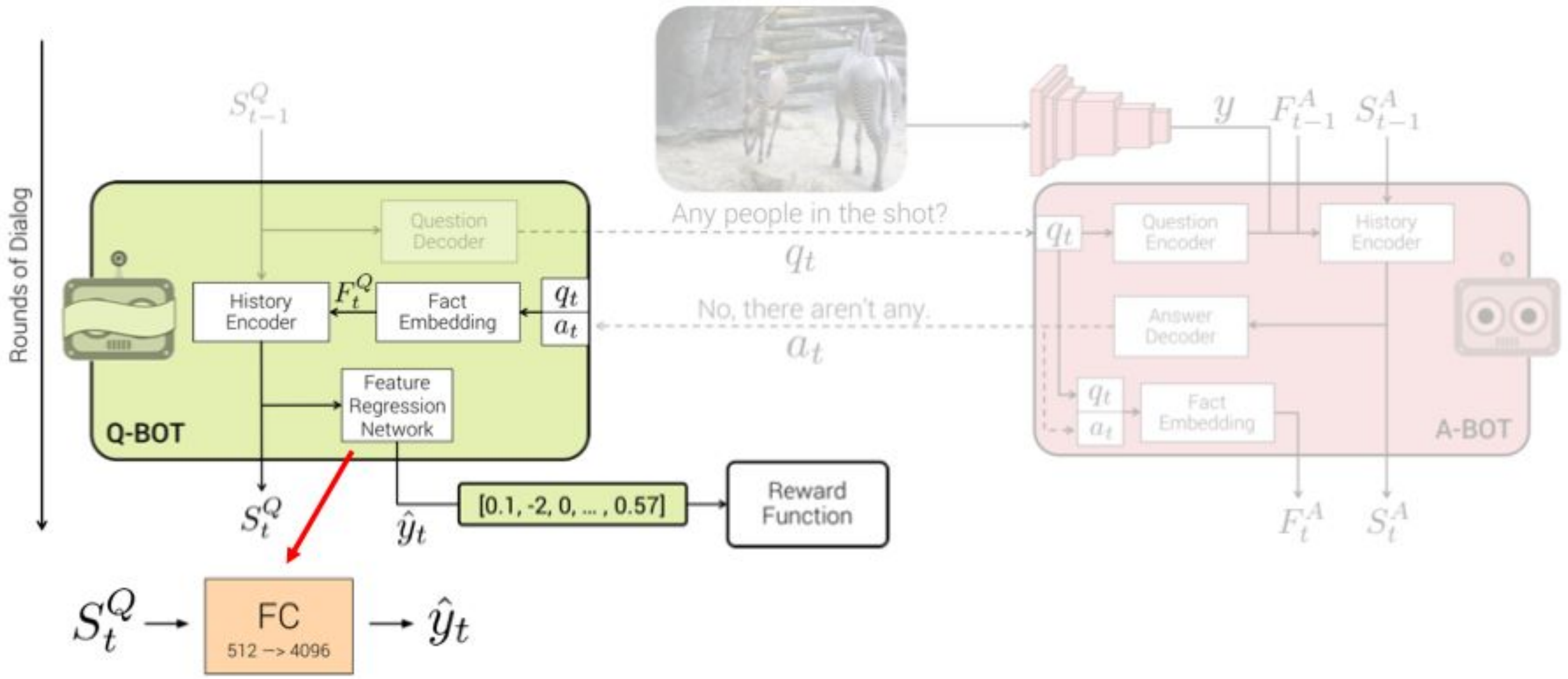
Q-Bot
$\pi_Q(q_t|S^Q_{t-1})$
A-Bot
$\pi_A(a_t|S^A_{t-1})$
Rounds of Dialog
S^Q_{t-1}
Question Decoder
History Encoder
F^Q_t
Fact Embedding
q_t
a_t
Feature Regression Network
Q-BOT
S^Q_t
$\hat{y}_t$
[0.1, -2, 0, ... , 0.57]
Reward Function
Any people in the shot?
q_t
No, there aren't any.
a_t
y
F^A_{t-1}
S^A_{t-1}
Question Encoder
History Encoder
Answer Decoder
Fact Embedding
A-BOT
F^A_t
S^A_t

Q-Bot

$\pi_Q(q_t|S^Q_{t-1})$

A-Bot

$\pi_A(a_t|S^A_{t-1})$

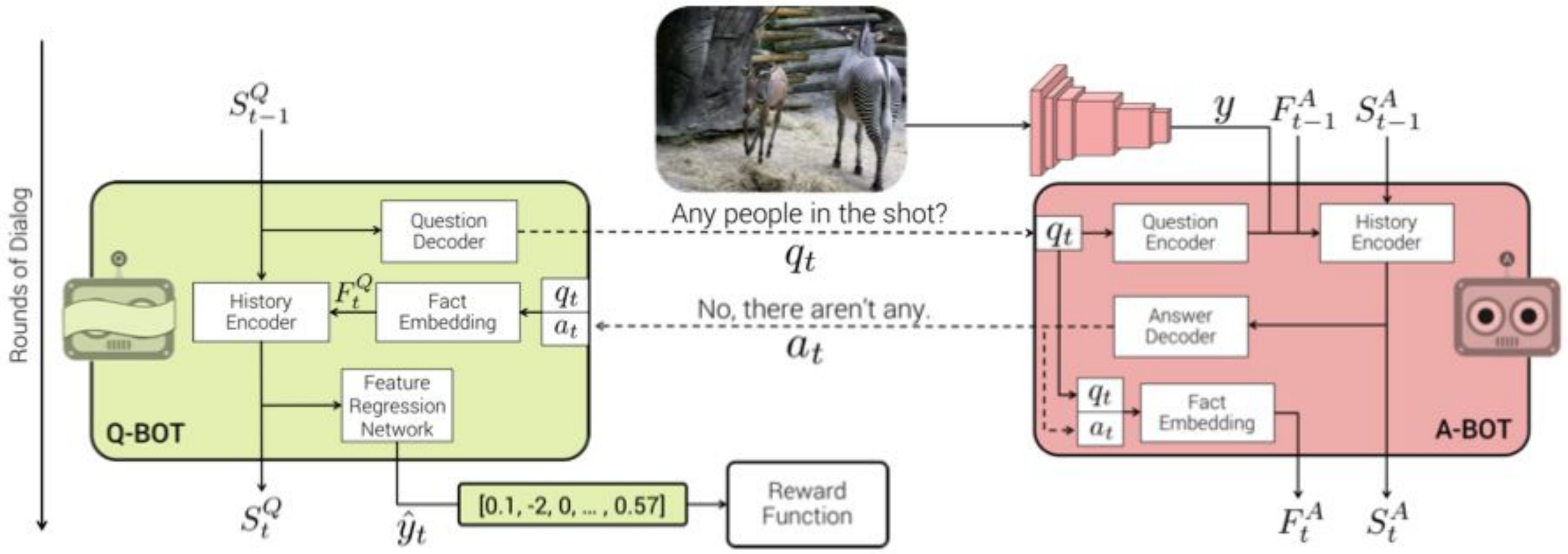


Q-Bot

$\pi_Q(q_t | S^Q_{t-1})$

A-Bot

$\pi_A(a_t | S^A_{t-1})$



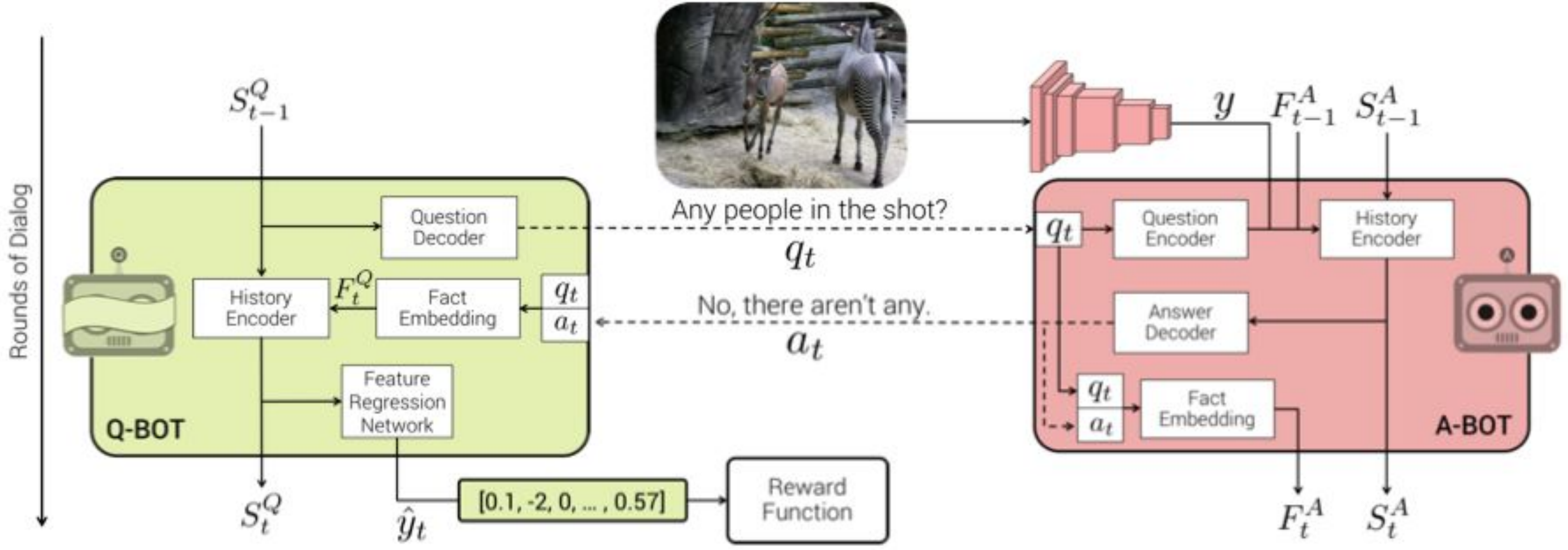
Model Evaluation

Q-Bot

$\pi_Q(q_t|S^Q_{t-1})$

A-Bot

$\pi_A(a_t|S^A_{t-1})$

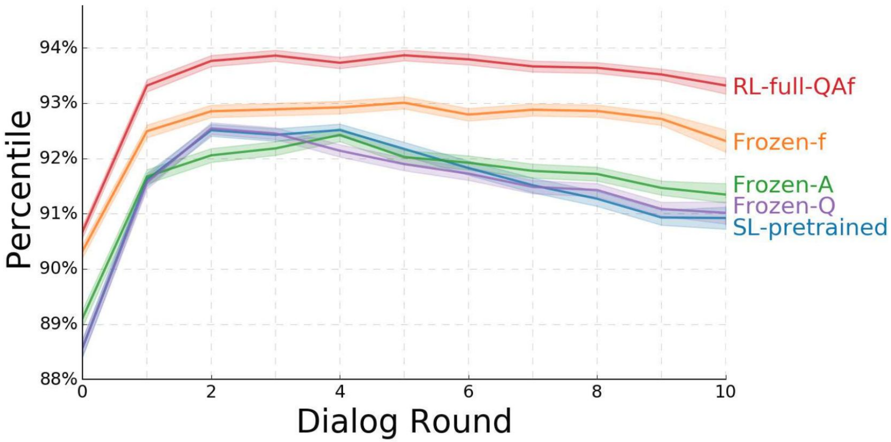


Model Evaluation

1. **Comparison with few natural ablations of the full model (RL-full-QAf)**
    - SL-pretrained
    - Frozen-A
    - Frozen-Q
    - Frozen-F (regression network)
2. **How well the agents perform at guessing game**
3. **How closely they emulate human dialogs**

Evaluation 1

1. **Comparison with few natural ablations of the full model (RL-full-QAf)**
2. How well the agents perform at guessing game
3. How closely they emulate human dialogs



UNIVERSITY OF WATERLOO

Evaluation 2

1. Comparison with few natural ablations of the full model (RL-full-QAf)
2. **How well the agents perform at guessing game**
3. How closely they emulate human dialogs

Evaluation 2

1. Comparison with few natural ablations of the full model (RL-full-QAf)
2. **How well the agents perform at guessing game**
3. How closely they emulate human dialogs

Evaluation 2

1. Comparison with few natural ablations of the full model (RL-full-QAf)
2. **How well the agents perform at guessing game**
3. How closely they emulate human dialogs

Evaluation 2

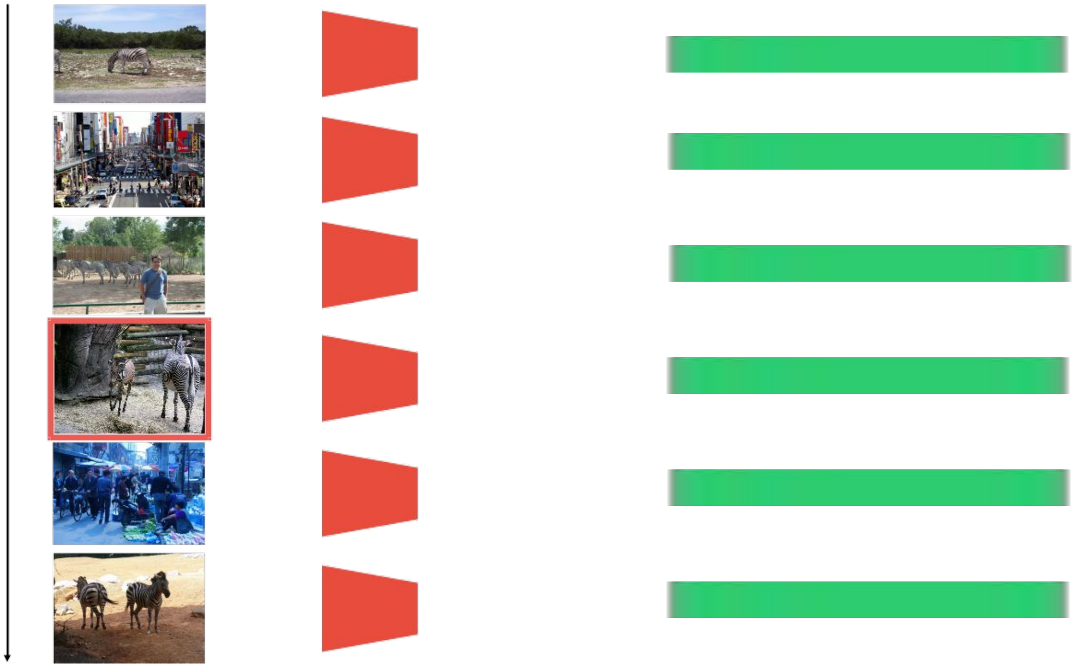
1. Comparison with few natural ablations of the full model (RL-full-QAf)
2. **How well the agents perform at guessing game**
3. How closely they emulate human dialogs

Evaluation 2

Test set
(~10k images)

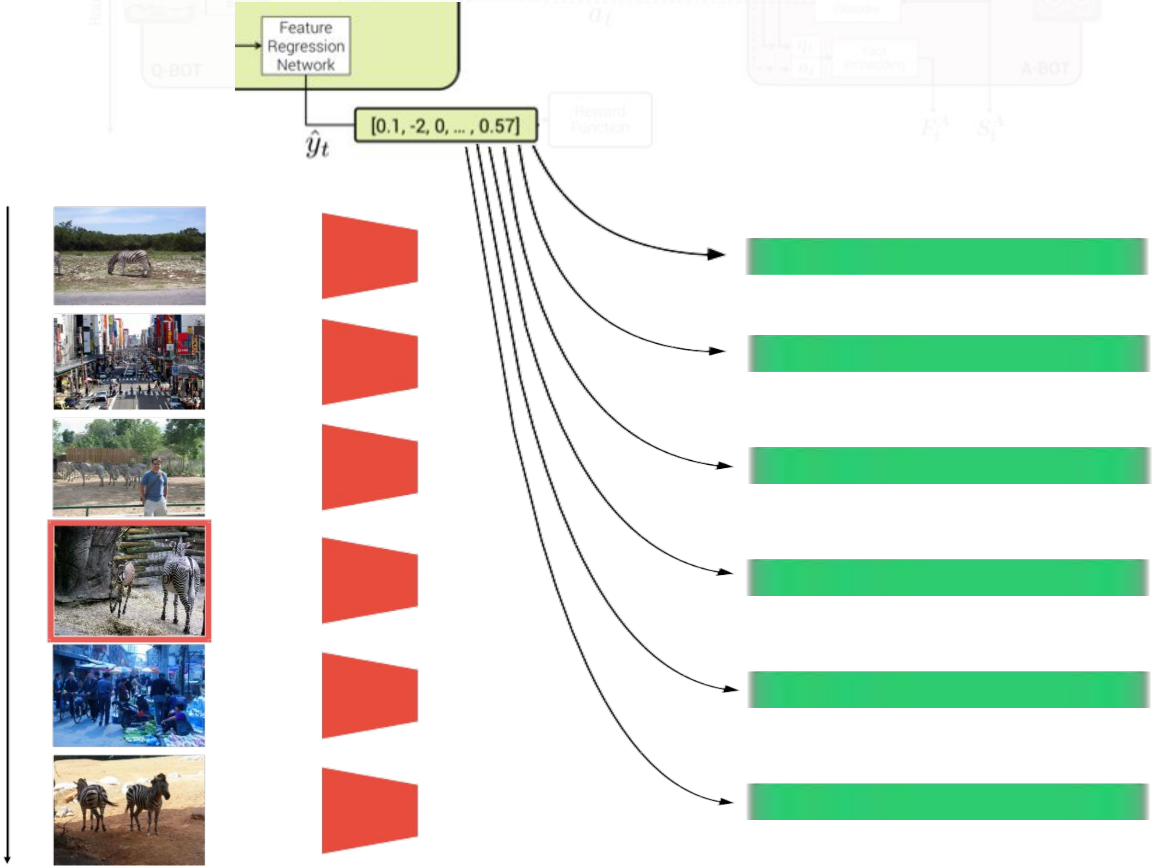
Evaluation 2

Test set
(~10k images)

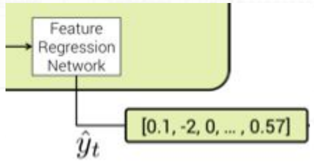


Evaluation 2

Test set
(~10k images)



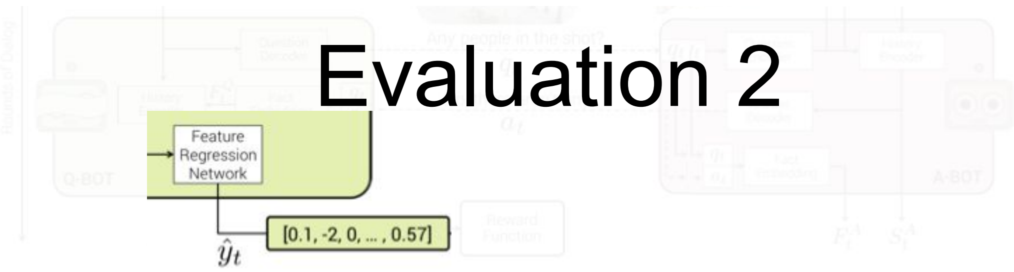
Evaluation 2



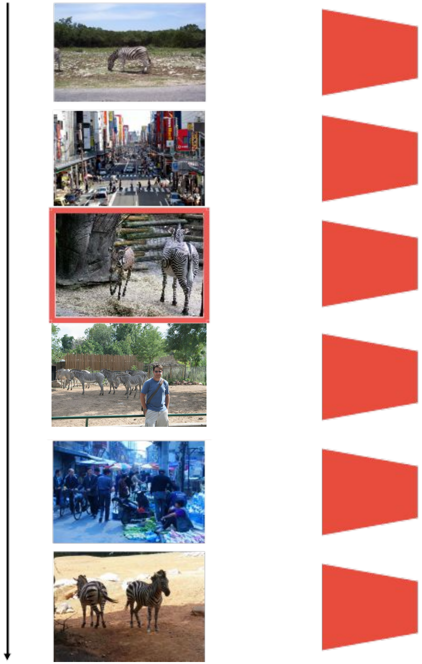
Test set
(~10k images)

Sorting based on distance to fc7 vectors

Evaluation 2

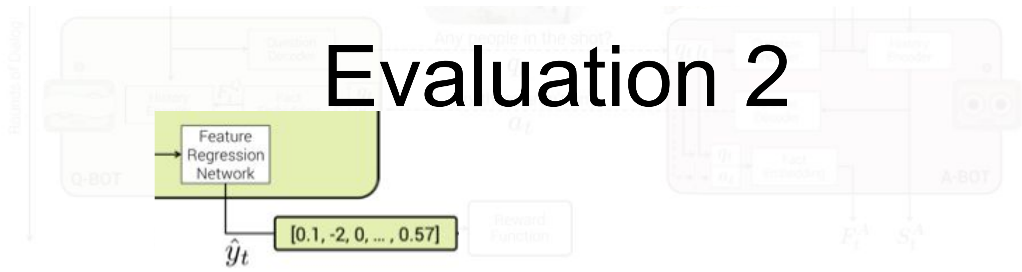


Test set
(~10k images)



Sorting based on distance to fc7 vectors

Evaluation 2



Test set
(~10k images)

Rank of ground truth image = 2

Evaluation 3

1. Comparison with few natural ablations of the full model (RL-full-QAf)
2. How well the agents perform at guessing game
3. **How closely they emulate human dialogs**

Human interpretability study to measure:

- whether humans can easily understand the Q-BOT-A-BOT dialog.
- how image-discriminative the interactions are.

Mean rank for ground-truth image
(lower is better)

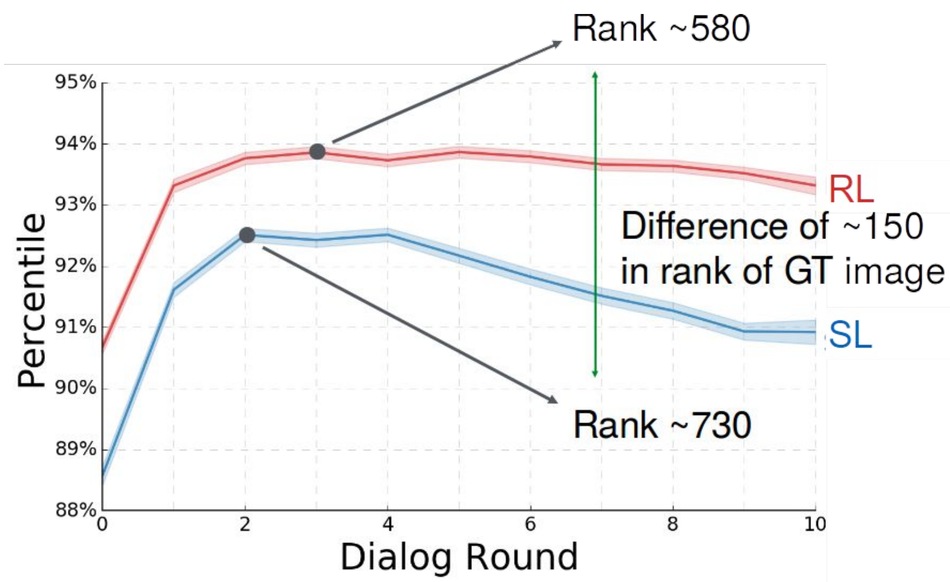
3.70 (SL) vs **2.73** (RL)

Mean Reciprocal Rank
(higher is better)

0.518 (SL) vs **0.622** (RL)

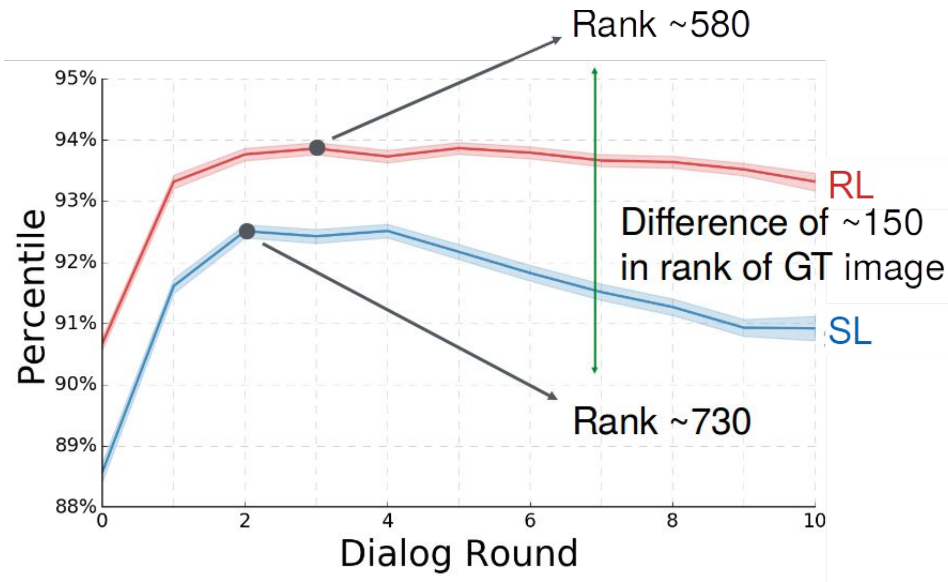
UNIVERSITY OF WATERLOO

Results



Rank ~580
95%
94%
93%
92%
91%
90%
89%
88%
Percentile
RL
Difference of ~150
in rank of GT image
SL
Rank ~730
0
2
4
6
8
10
Dialog Round

Results

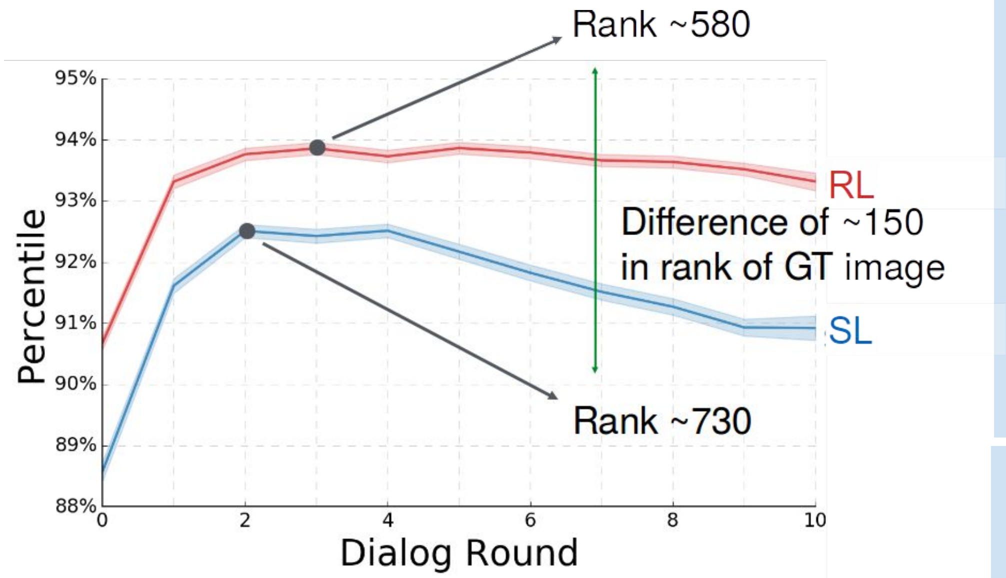


SL vs SL+RL

Supervised Q-BOT seemed to mimic how humans ask questions.

RL trained Q-BOT seemed to shifts strategies and asks questions that the A-BOT was better at answering.

Results



SL vs SL+RL

Supervised Q-BOT seemed to mimic how humans ask questions.

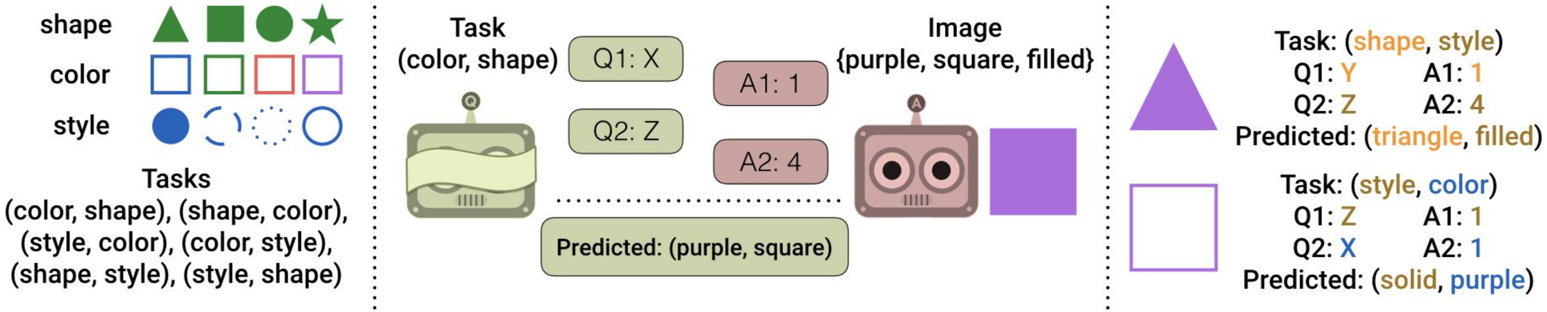
RL trained Q-BOT seemed to shifts strategies and asks questions that the A-BOT was better at answering.

Dialog between the agents were NOT 'hand engineered' to be image discriminative.
It **emerged as a strategy to succeed** at the image-guessing game.

UNIVERSITY OF WATERLOO

Results

- **Emergence of Grounding (RL from scratch)**



The two bots invented their own communication protocol without any human supervision

More details in the follow-up paper:
Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog
Kottur et al., EMNLP 2017

UNIVERSITY OF WATERLOO

Contributions

- Goal-driven training of visual question answering and dialog agents.
  - Self-talk = infinite data
  - Goal-based = evaluation on downstream task
  - Agent-driven = agents learn to deal with consequences of their actions.

- End-to-end learning from pixels to multi-agent multi-round dialog to game reward.
  - Move from SL **on static datasets** to RL on **actual environment.**

UNIVERSITY OF WATERLOO

Class Discussions

- Do you think this approach is limited to goal-driven tasks in dialog systems?
    - If not, how can this be extended to open-ended conversations?
- What other reward models can be used to make SL-RL dialog systems more successful?