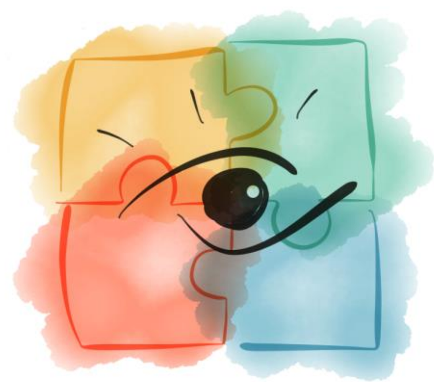
VQA → Visual Dialog



Devi Parikh

Georgia Tech

facebook research


2012 SUMMER
ARTS
FESTIVAL

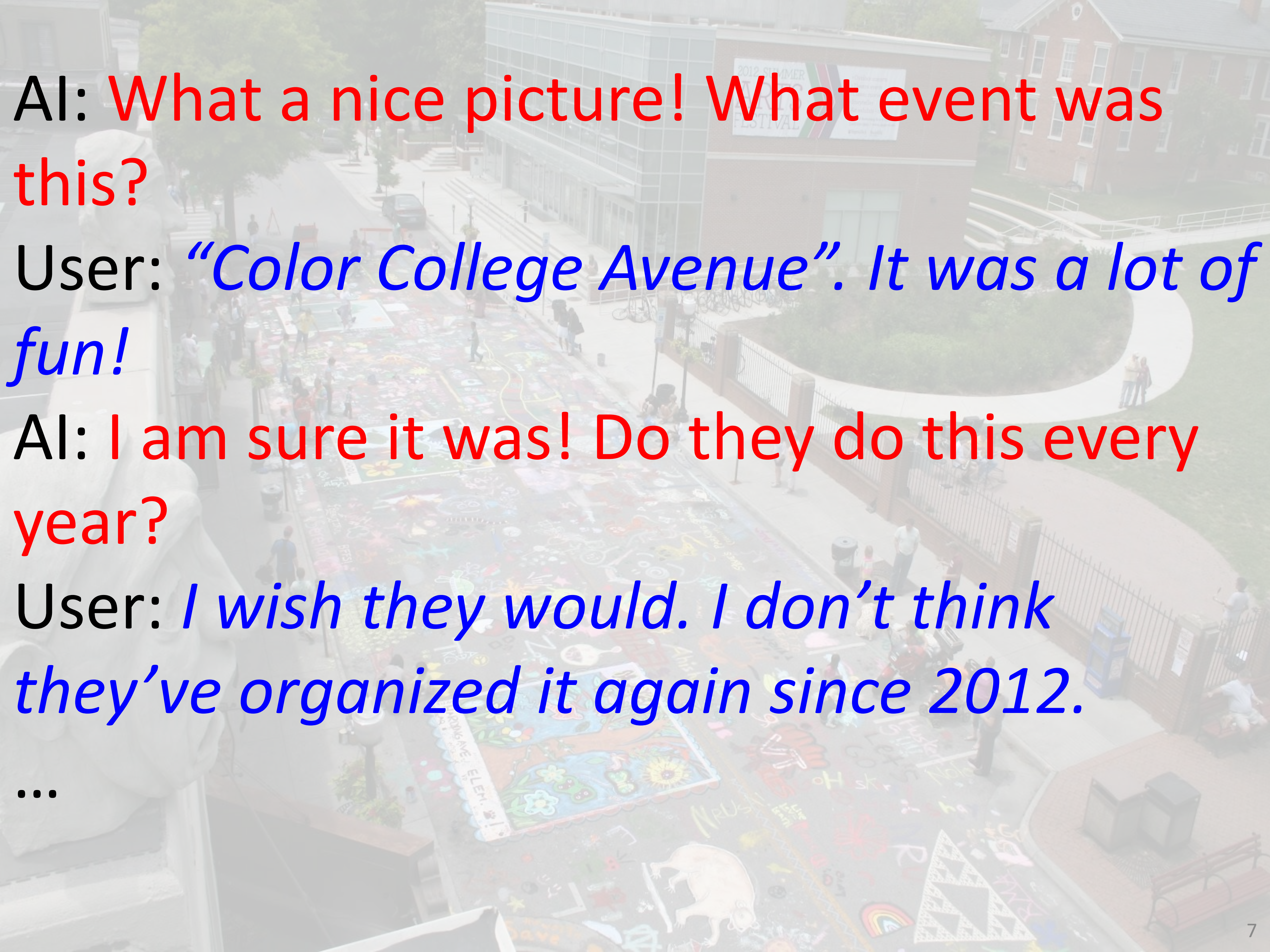
People coloring a street on a college campus

It was a great event! It brought families out, and the whole community together.


2012 SUMMER
ARTS
FESTIVAL

Q. What are they coloring the street with?

A. Chalk



AI: What a nice picture! What event was this?

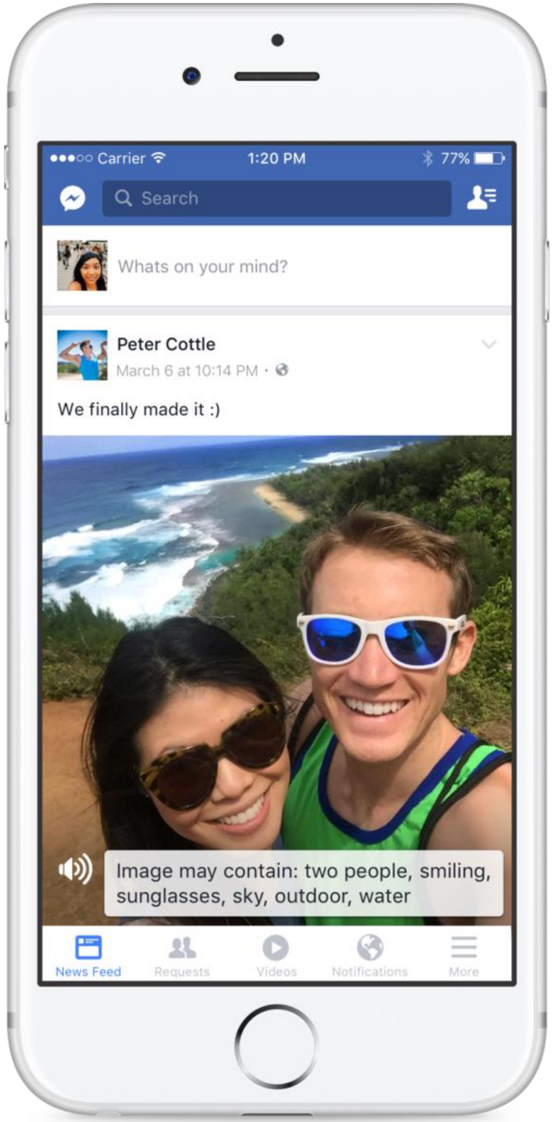
User: *"Color College Avenue". It was a lot of fun!*

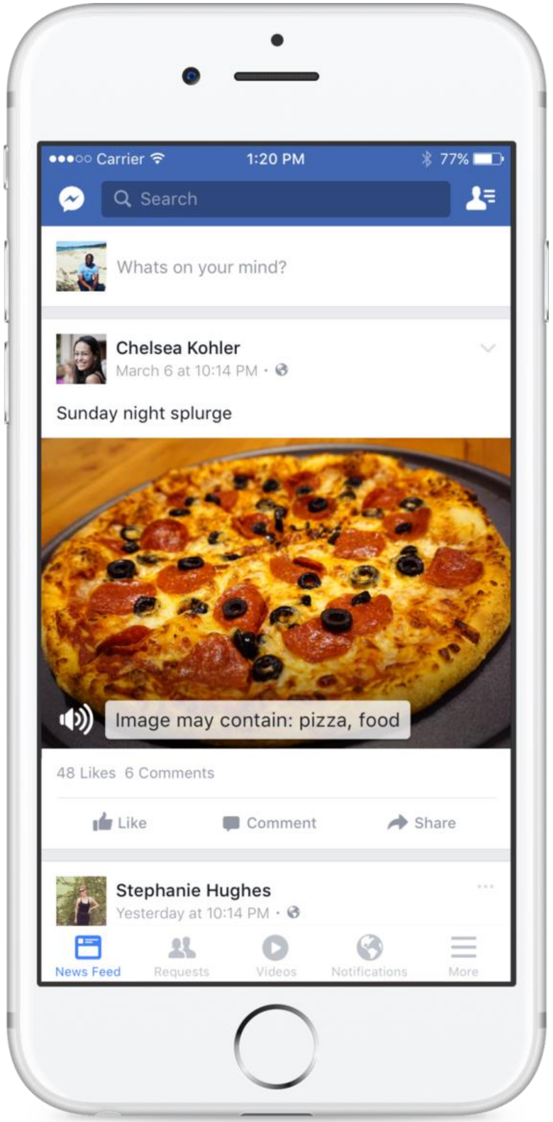
AI: I am sure it was! Do they do this every year?

User: *I wish they would. I don't think they've organized it again since 2012.*

…

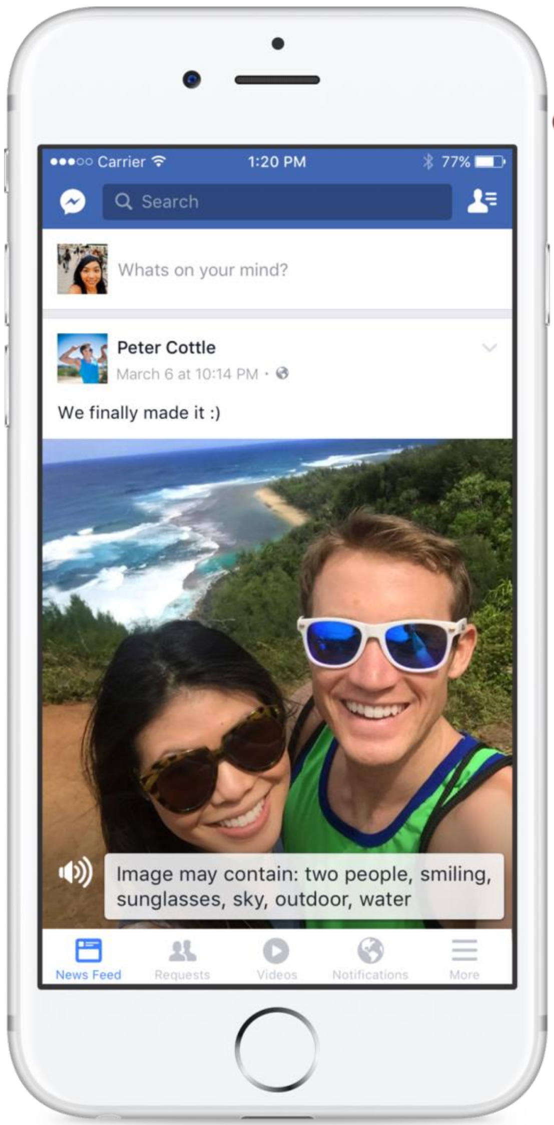
Aid visually-impaired users

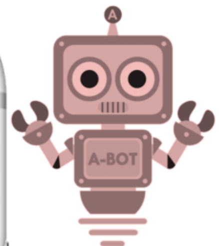




FACEBOOK'S AI CAN CAPTION PHOTOS FOR THE BLIND ON ITS OWN

Aid visually-impaired users





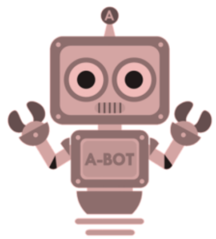
Peter just uploaded a picture from his vacation in Hawaii

Great, is he at the beach?

No, on a mountain

...

Aid ‘situationally-impaired’ analysts



Did anyone enter this room last week?

Yes, 127 instances logged on camera

Were any of them carrying a black bag?

…

Natural language instructions for robots

Is there smoke in any room around you?

Yes, in one room

Go there and look for people

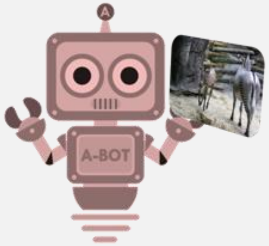
…

Image Credit: Lockheed Martin; DARPA Robotics Challenge

Outline

Visual Question Answering

Visual Dialog



GuessWhat?! Visual object discovery through multi-modal dialogue

Harm de Vries
University of Montreal
mail@harmdevries.com

Florian Strub
Univ. Lille, CNRS, Centrale Lille,
Inria, UMR 9189 CRIStAL
florian.strub@inria.fr

Sarath Chandar
University of Montreal
sarathcse2008@gmail.com

Olivier Pietquin
DeepMind
pietquin@google.com

Hugo Larochelle
Twitter
hlarochelle@twitter.com

Aaron Courville
University of Montreal
aaron.courville@gmail.com

Abstract

We introduce GuessWhat?!, a two-player guessing game as a testbed for research on the interplay of computer vision and dialogue systems. The goal of the game is to locate an unknown object in a rich image scene by asking a sequence of questions. Higher-level image understanding, like spatial reasoning and language grounding, is required to solve the proposed task. Our key contribution is the collection of a large-scale dataset consisting of 150K human-played games with a total of 800K visual question-answer pairs on 66K images. We explain our design decisions in collecting the dataset and introduce the oracle and questioner tasks that are associated with the two players of the game. We prototyped deep learning models to establish initial baselines of the introduced tasks.

Questioner	**Oracle**
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

End-to-end optimization of goal-driven and visually grounded dialogue systems

Florian Strub
florian.strub@inria.fr
Univ. Lille, CNRS, Centrale Lille, Inria,
UMR 9189 - CRIStAL, F-59000 Lille, France

Harm de Vries
mail@harmdevries.com
University of Montreal

Jeremie Mary
jeremie.mary@univ-lille3.fr
Univ. Lille, CNRS, Centrale Lille, Inria,
UMR 9189 - CRIStAL, F-59000 Lille, France

Bilal Piot
piot@google.com
DeepMind

Aaron Courville
aaron.courville@gmail.com
University of Montreal

Olivier Pietquin
pietquin@google.com
DeepMind

Abstract

End-to-end design of dialogue systems has recently become a popular research topic thanks to powerful tools such as encoder-decoder architectures for sequence-to-sequence learning. Yet, most current approaches cast human-machine dialogue management as a supervised learning problem, aiming at predicting the next utterance of a participant given the full history of the dialogue. This vision is too simplistic to render the intrinsic planning problem inherent to dialogue as well as its grounded nature, making the context of a dialogue larger than the sole history. This is why only chit-chat and question answering tasks have been addressed so

Is it a person?	*No*
Is it an item being worn or held?	*Yes*
Is it a snowboard?	*Yes*
Is it the red one?	*No*
Is it the one being held by the person in blue?	*Yes*

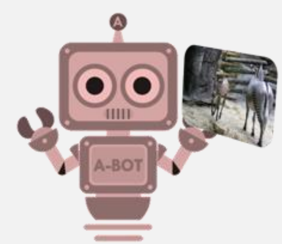
Is it a cow?	*Yes*
Is it the big cow in the middle?	*No*
Is the cow on the left?	*No*
On the right ?	*Yes*
First cow near us?	*Yes*

Figure 1: Two example games of the GuessWhat?! dataset. The correct object is highlighted by a green mask.

Outline

Visual Question Answering

Visual Dialog



Visual Question Answering (VQA)

Visual Question Answering (VQA)

What is the mustache made of?

Visual Question Answering (VQA)

What is the mustache made of?

AI System

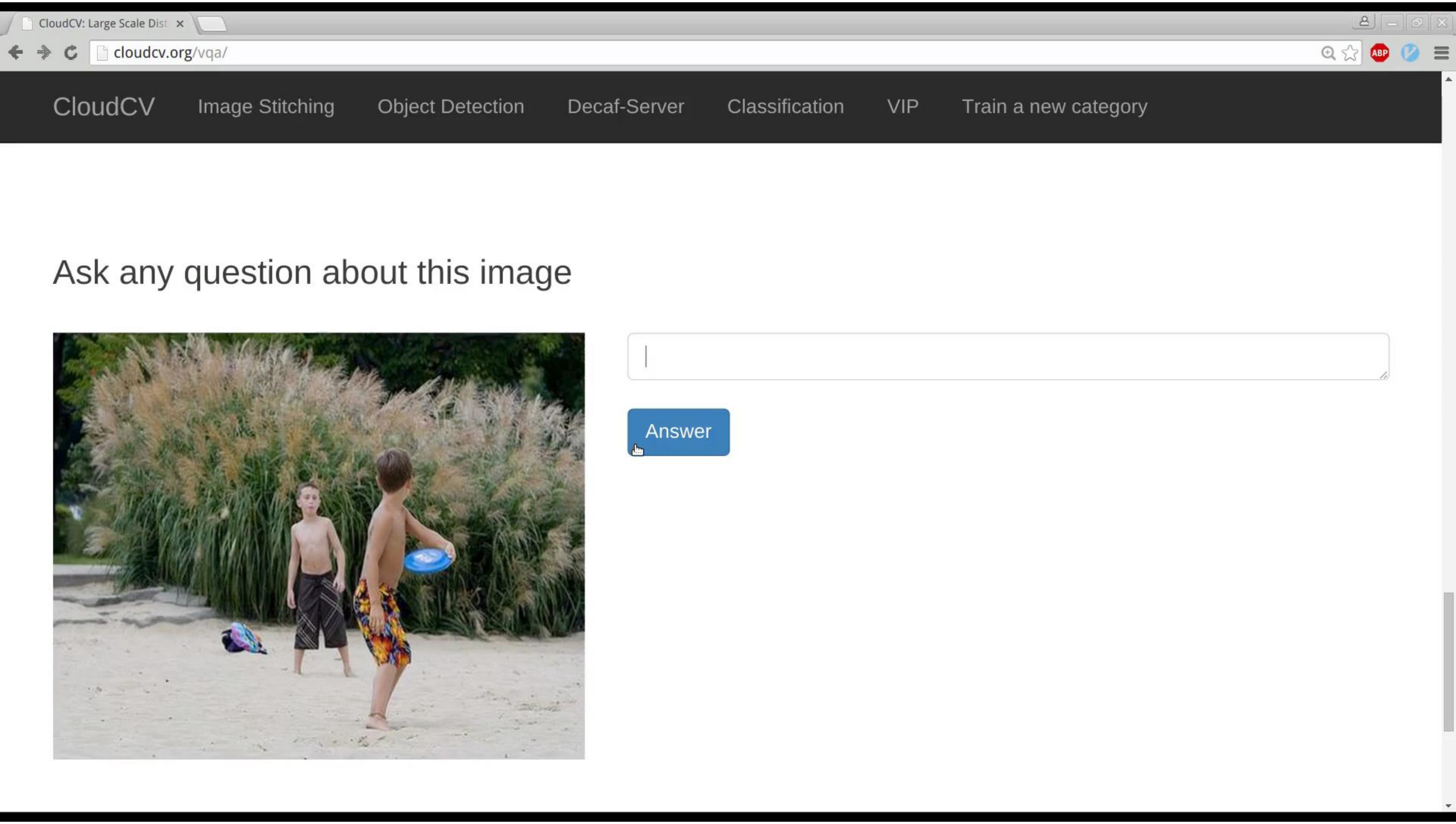
Visual Question Answering (VQA)

What is the mustache made of?

AI System

bananas

Visual Question Answering (VQA)

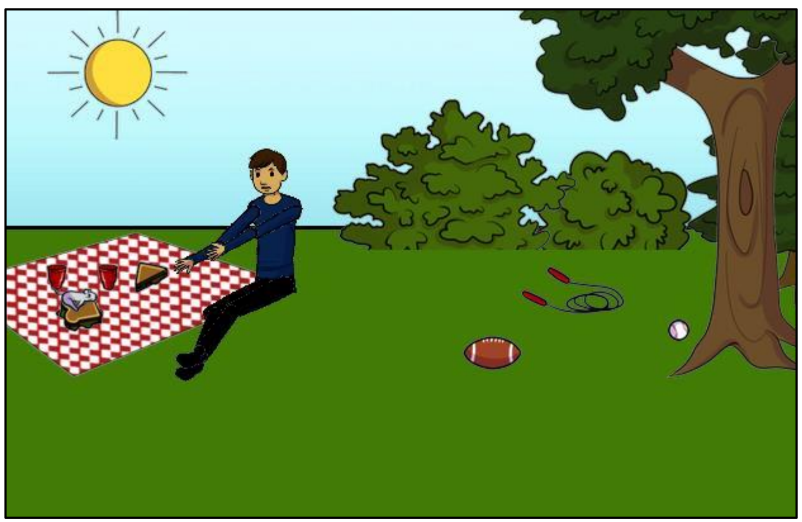


www.visualqa.org

Visual Question Answering (VQA)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

VQA Dataset

VQA Dataset

>0.25 million images

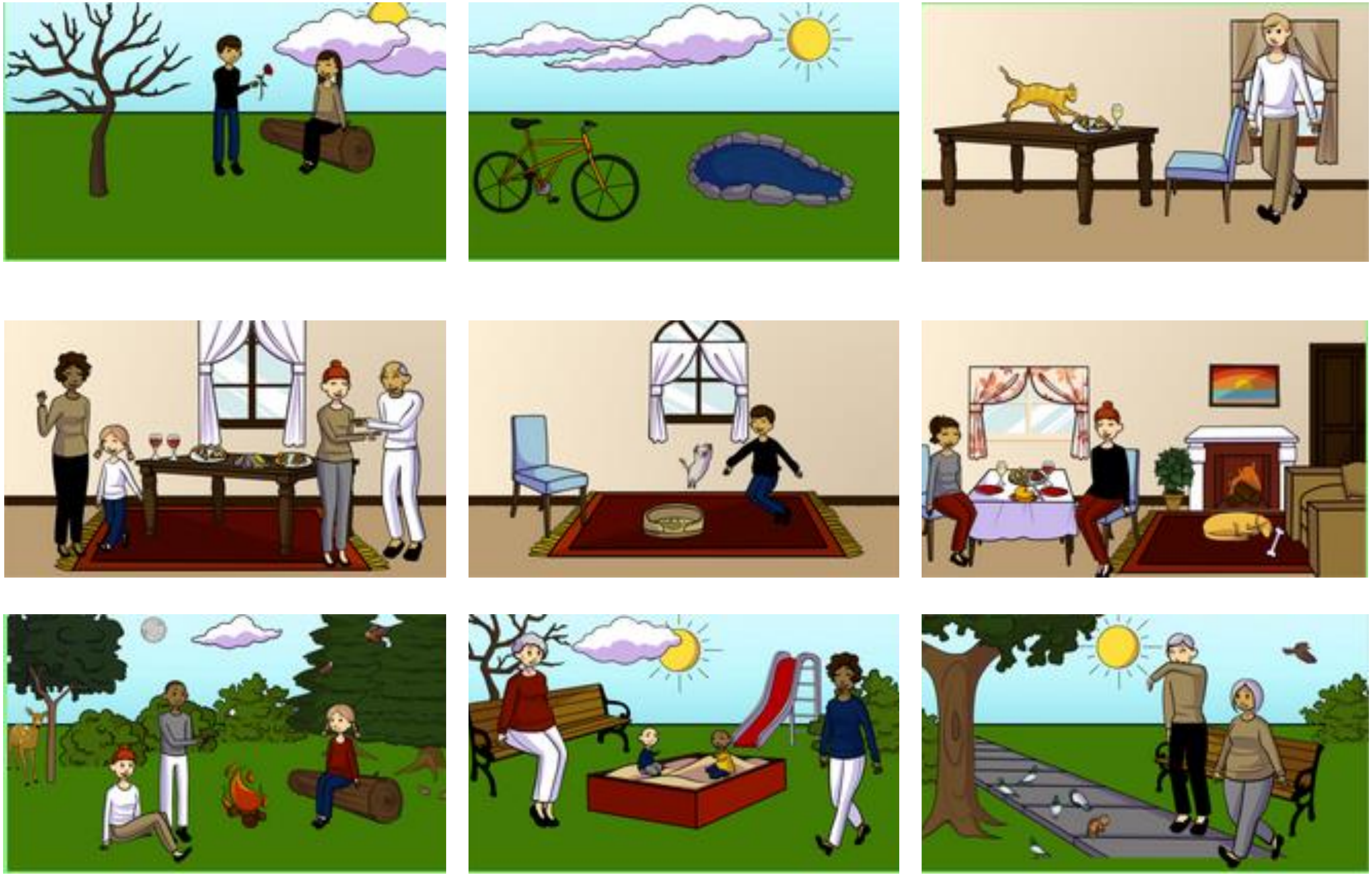








254,721 images (COCO)



50,000 scenes

VQA Dataset

>0.25 million images

>0.76 million questions

Questions

Stump a smart robot! Ask a question about this image that a human can answer, but a smart robot probably can't!

We have built … **can recognize the scene (e.g, kitchen, beach)** … **smart robot!**

Ask a question … **IMPORTANT: T** … **should not be able to answer the question wi** …

Stump a smart robot!
Ask a question that a human can answer,
but a smart robot probably can't!

… **ns below:**

- Do not repeat questions. Do not ask the same questions or the same questions with minor variations over and over again across images. Think of a new question each time specific to each image.
- Each question should be a single question. Do not ask questions that have multiple parts or multiple sub-questions in them.
- Do not ask generic questions that can be asked of many other images. Ask questions specific to each image.

Please ask a question about this image that a human can answer *if* looking at the image (and not otherwise), but would stump this smart robot:

Q1: Write your question here to stump this smart robot.

VQA Dataset

>0.25 million images

>0.76 million questions

~10 million answers

[Antol et al., ICCV 2015]

Papers using VQA

Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick
School of Computer Science, The University of Adelaide
{qi.wu01,p.wang,chunhua.shen,anton.vandenhengel,anthony.dick}@adelaide.edu.au

Simple Baseline for Visual Question Answering

Bolei Zhou[1], Yuandong Tian[2], Sainbayar Sukhbaatar[2], Arthur Szlam[2], and Rob Fergus[2]
[1]Massachusetts Institute of Technology
[2]Facebook AI Research

Compositional Memory for Visual Question Answering

Aiwen Jiang[1,2] Fang Wang[2] Fatih Porikli[2] Yi Li* [2,3]
[1]Jiangxi Normal University [2]NICTA and ANU [3]Toyota Research Institute North America
[1]aiwen.jiang@nicta.com.au [2]{fang.wang, fatih.porikli}@nicta.com.au [3]yi.li@tema.toyota.com

Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering

Huijuan Xu, UMass Lowell, hxu1@cs.uml.edu
Kate Saenko, UMass Lowell, saenko@cs.uml.edu

Deep Compositional Question Answering with Neural Module Networks

Jacob Andreas Marcus Rohrbach Trevor Darrell Dan Klein
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{jda,rohrbach,trevor,klein}@{cs,eecs,eecs,cs}.berkeley.edu

Where To Look: Focus Regions for Visual Question Answering

Kevin J. Shih, Saurabh Singh, and Derek Hoiem
University of Illinois at Urbana-Champaign
{kjshih2, ss1, dhoiem}@illinois.edu

ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering

Kan Chen, University of Southern California, kanchen@usc.edu
Jiang Wang, Baidu Research - IDL, wangjiang03@baidu.com
Liang-Chieh Chen, UCLA, lcchen@cs.ucla.edu
Haoyuan Gao, Baidu Research - IDL, gaohaoyuan@baidu.com
Wei Xu, Baidu Research - IDL, wei.xu@baidu.com
Ram Nevatia, University of Southern California, nevatia@usc.edu

Stacked Attention Networks for Image Question Answering

Zichao Yang[1], Xiaodong He[2], Jianfeng Gao[2], Li Deng[2], Alex Smola[1]
[1]Carnegie Mellon University, [2]Microsoft Research, Redmond, WA 98052, USA
oy@cs.cmu.edu, {xiaohe, jfgao, deng}@microsoft.com, alex@smola.org

VQA Challenge @ CVPR16

Competition

VQA

VQA Real Image Challenge (Open-Ended)

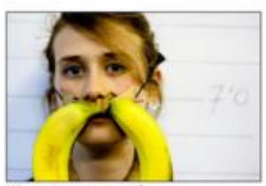
Organized by vqateam - Current server time: March 22, 2016, 5 a.m. UTC

▶ Current	Next
Real Challenge test2015 (oe)	Real test2015 (oe)
Oct. 21, 2015, midnight UTC	Oct. 21, 2015, midnight UTC

Learn the Details | Phases | Participate | Results | Forums

- **Overview**
- Evaluation
- Terms and Conditions

Visual Question Answering (VQA)



What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Recent progress in computer vision and natural language processing has demonstrated that lower-level tasks are much closer to being solved. We believe that the time is ripe to pursue

VQA Challenge @ CVPR16

	By Answer Type			Overall
	Yes/No	Number	Other	
UC Berkeley & Sony[14]	83.24	39.47	58	66.47
Naver Labs[10]	83.31	38.7	54.62	64.79
DLAIT[5]	83.25	40.07	52.09	63.68
snubi-naverlabs[25]	83.16	39.14	51.33	63.18
POSTECH[11]	81.67	38.16	52.79	63.17
san[24]	79.11	36.41	46.42	58.85
UC Berkeley (NMN)[16]	81.16	37.7	44.01	58.66
global_vision[22]	78.24	36.27	46.32	58.43
vqateam-deeperLSTM_NormlizeCNN[27]	80.56	36.53	43.73	58.16
Mujtaba hasan[9]	80.28	36.92	42.24	57.36
RIT[12]	78.82	35.97	42.13	56.61
Bolei[2]	76.76	34.98	42.62	55.89
UPV_UB[18]	78.88	36.33	40.27	55.77
att[21]	78.1	35.3	40.27	55.34
vqateam-lstm_cnn[28]	79.01	35.55	36.8	54.06
UPC[17]	78.05	35.53	36.7	53.62
vqateam-nearest_neighbor[29]	71.73	24.31	22	42.73
vqateam-prior_per_qtype[30]	71.17	35.63	9.32	37.55
vqateam-all_yes[26]	70.53	0.43	1.26	29.72

Winning entry (MCB)

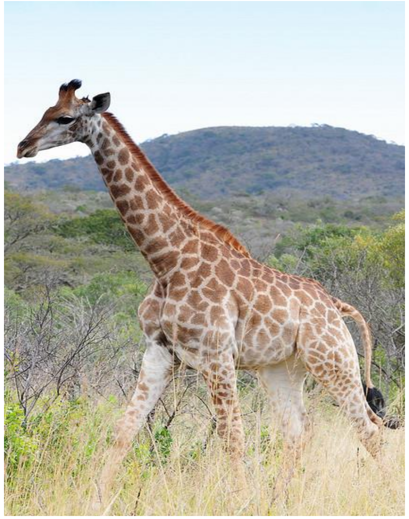
Open-ended: 66%

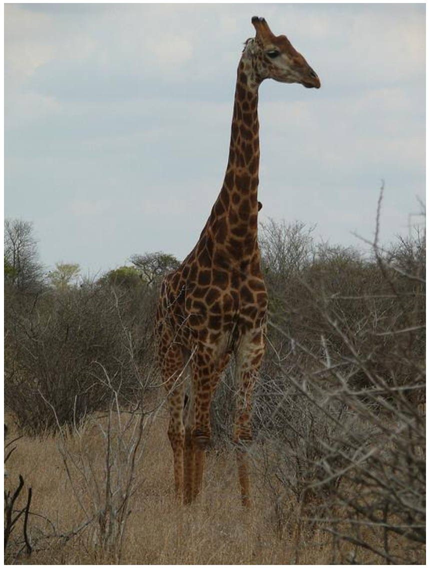
Multiple-choice: 70%

The Power of Language Priors

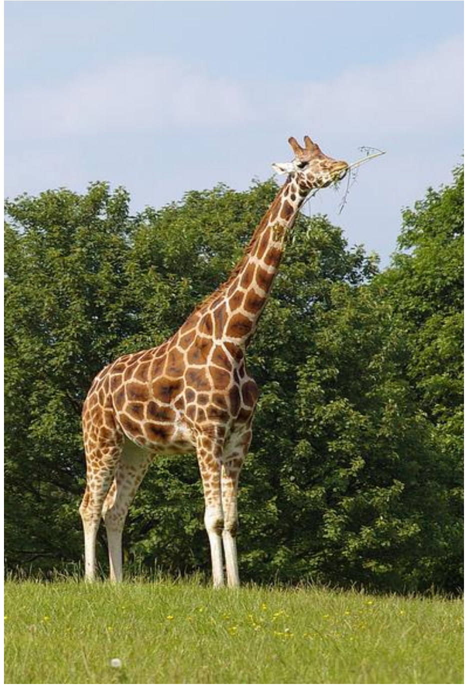
Slide credit: Yash Goyal and Peng Zhang

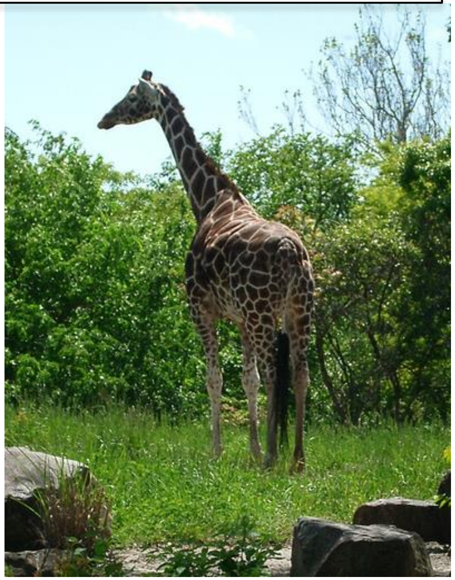
The Power of Language Priors





A giraffe is standing in grass next to a tree





Slide credit: Yash Goyal and Peng Zhang

The Power of Language Priors

Is there a clock … ?

'yes' 98%



……


Is the man wearing glasses … ?

'yes' 94%



……


Are the lights on … ?

'yes' 85%



……


Do you see a … ?

'yes' 87%



……


Slide credit: Yash Goyal and Peng Zhang

The Power of Language Priors

Is the man standing … ?

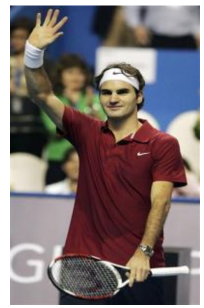
‘no’ 69%



......


What sport is … ?

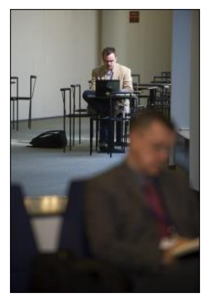
‘tennis’ 41%



......


How many … ?

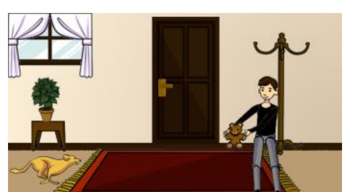
‘2’ 39%



......


What animal is … ?

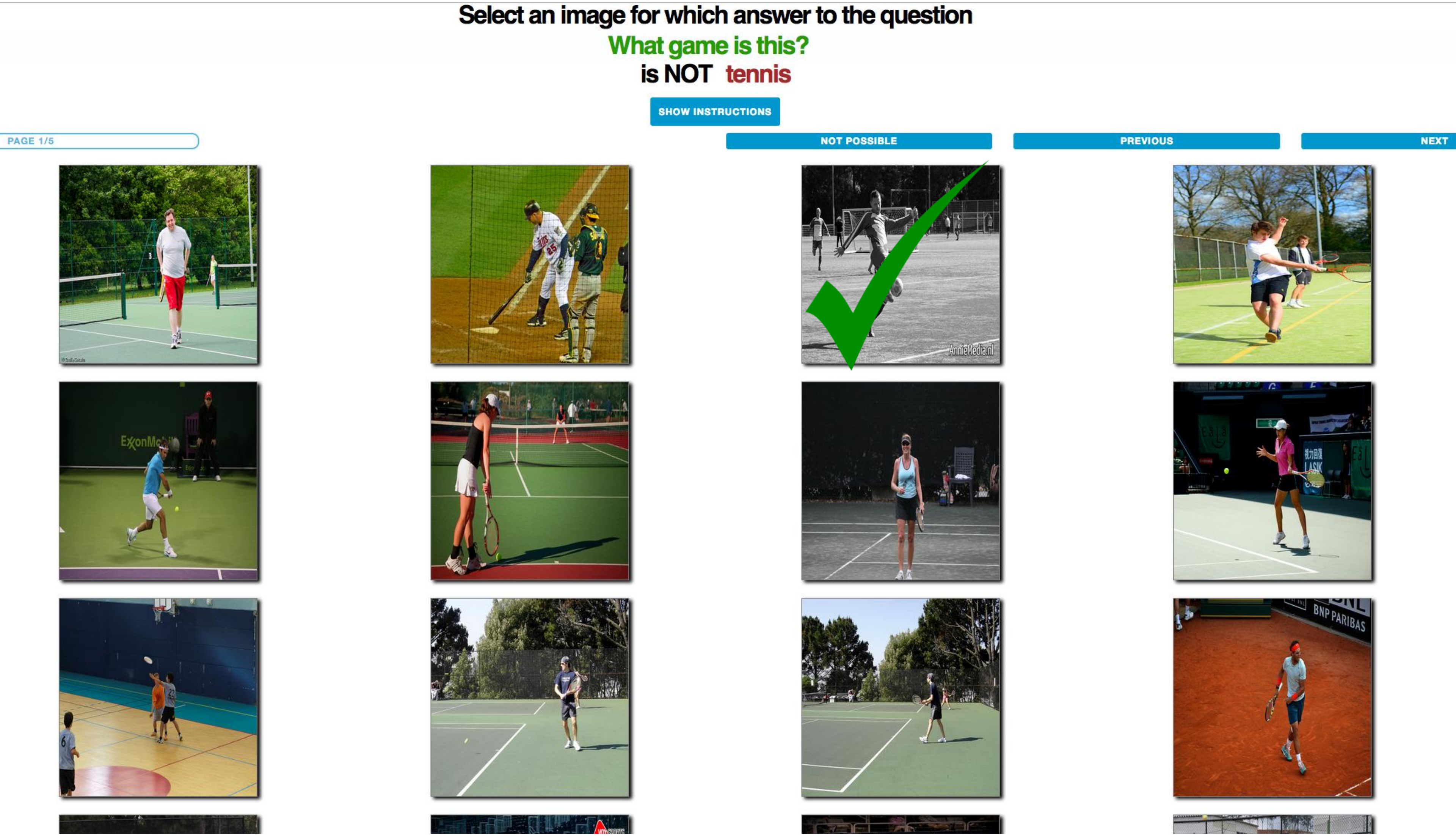
‘dog’ 35%



......


Slide credit: Yash Goyal and Peng Zhang

Balancing the VQA dataset



Balancing the VQA dataset

Is the TV on?

yes

no

Balancing the VQA dataset

How many pets are present?

2

1

Balancing the VQA dataset

What sign is this?

handicap

one way

Balancing the VQA dataset

Where is the child sitting?

fridge

arms

Balancing the VQA dataset

What is the cat doing on the rug?

sleeping

sitting

Balancing the VQA dataset

What color are the pants?

orange

brown

VQA v2.0

- More balanced than VQA v1.0
  - Entropy of answers increases by 56%

- Bigger than VQA v2.0
  - ~1.8 times image-question pairs

Benchmarking SOTA VQA models

- SOTA VQA models
  - Drop in performance by 7-8%
  - Gain 1-2% back when re-trained on balanced dataset
- By answer types
  - Biggest drop in performance in yes/no (10-12%)
  - Biggest improvement gained by re-training in yes/no (3-4%) and number (2-3%)

Trends

	By Answer Type			Overall
	Yes/No	Number	Other	
UC Berkeley & Sony[14]	83.79	38.9	58.64	66.9
Naver Labs[10]	83.78	37.67	54.74	64.89
DLAIT[5]	83.65	39.18	52.62	63.97
snubi-naverlabs[25]	83.64	38.43	51.61	63.4

0.15% 1.51% 7.03% 3.5%

VQA v2.0

2nd VQA Challenge @ CVPR17!

CVPR
July 21-26
HONOLULU
2017

Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (CVPR 2017)

Yash Goyal
(Virginia Tech)

Tejas Khot
(Virginia Tech)

Doug Summers-Stay
(Army Research Lab)

Dhruv Batra
(Georgia Tech / FAIR)

Devi Parikh
(Georgia Tech / FAIR)

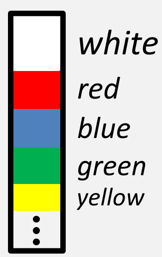
(Another) problem with existing setup

Train

Q: What color is the dog?

A: White

Training Prior



Slide credit: Aishwarya Agrawal

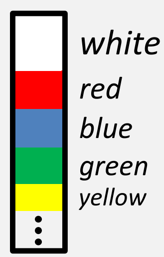
(Another) problem with existing setup

Train

Q: What color is the dog?

A: White

Training Prior



Test

Q: What color is the dog?

A: Black

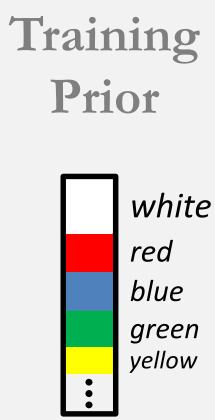
Slide credit: Aishwarya Agrawal

(Another) problem with existing setup

Train

Q: What color is the dog?

A: White



Test

Q: What color is the dog?

A: Black

Prediction:
White

Slide credit: Aishwarya Agrawal

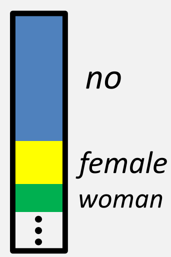
(Another) problem with existing setup

Train

Q: Is the person wearing shorts?

A: No

Training Prior



Test

Q: Is the person wearing shorts?

A: Yes

Prediction:
No

Slide credit: Aishwarya Agrawal

(Another) problem with existing setup

- Similar priors in train and test
- Memorization does not hurt as much
- Problematic for benchmarking progress

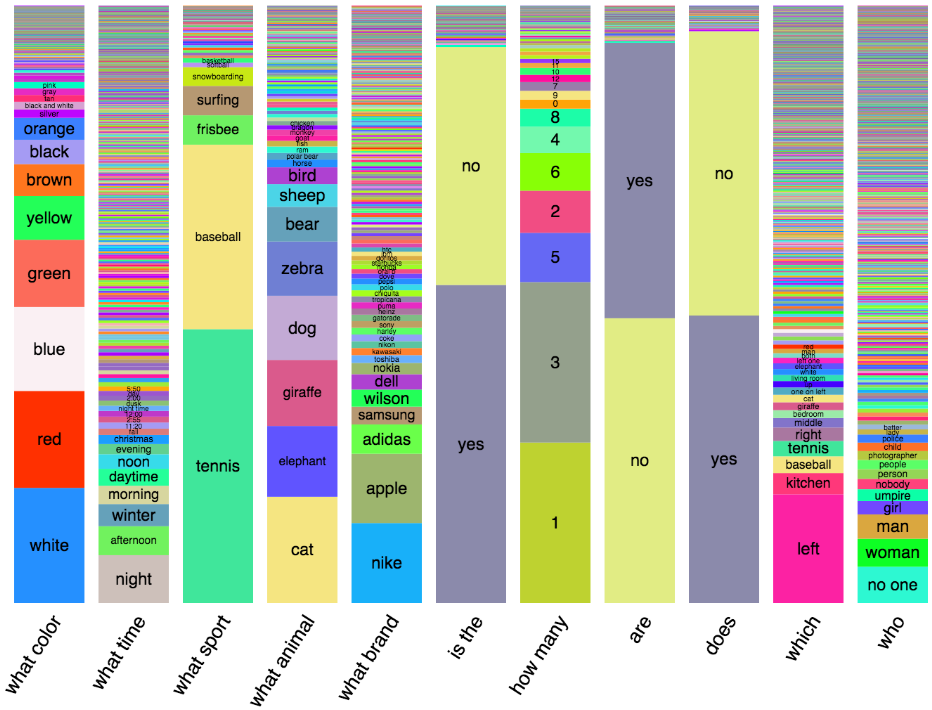
Meet VQA-CP!

- Visual Question Answering under Changing Priors

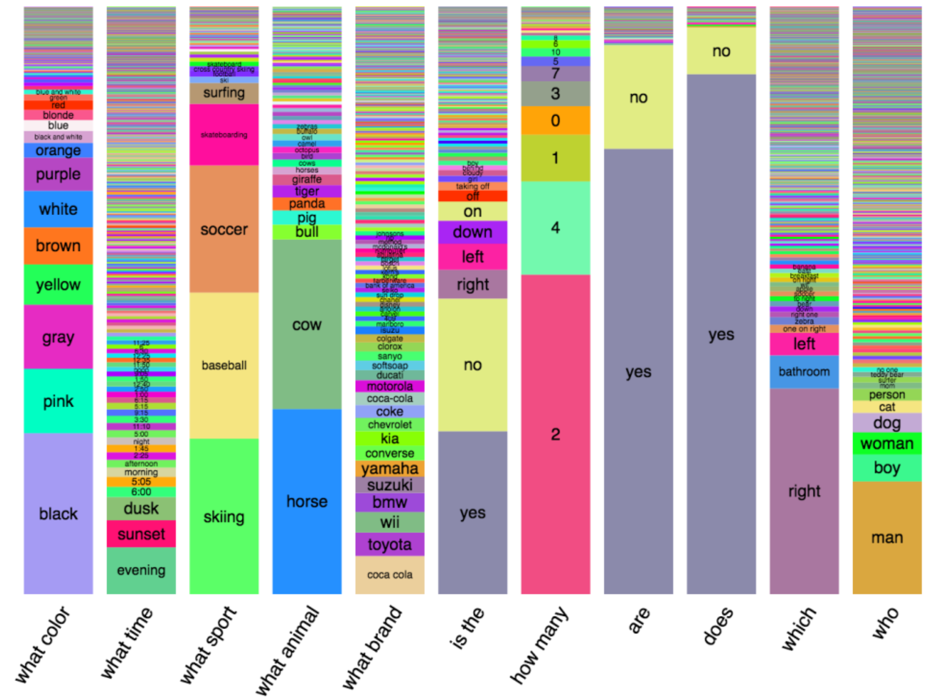
- A new split of the VQA v1.0 dataset (Antol et al., ICCV 2015)

Slide credit: Aishwarya Agrawal

VQA-CP Train Split



VQA-CP Test Split



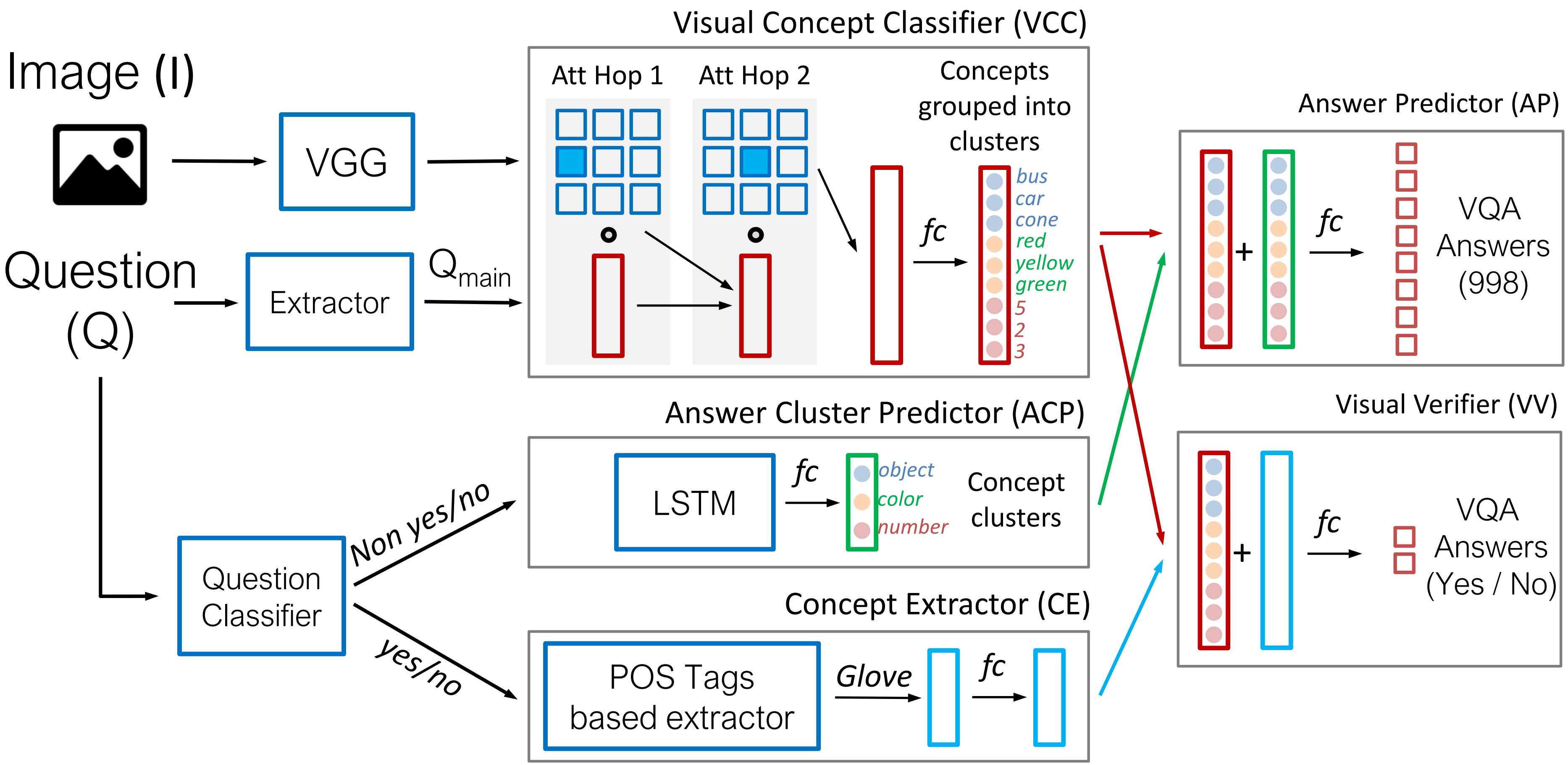
Slide credit: Aishwarya Agrawal

Performance of VQA models on VQA-CP

Model	Dataset	Overall	Yes/No	Number	Other	
d-LSTM Q + norm I	VQA	54.23	79.81	33.26	40.35	31% drop
(Antol et al. ICCV15)	VQA-CP	23.51	34.53	11.40	17.42	
NMN	VQA	54.83	80.39	33.45	41.07	25% drop
(Andreas et al. CVPR16)	VQA-CP	29.64	38.85	11.23	27.88	
SAN	VQA	55.86	78.54	33.46	44.51	29% drop
(Yang et al. CVPR16)	VQA-CP	26.88	35.34	11.34	24.70	
MCB	VQA	60.97	81.62	34.56	52.16	27% drop
(Fukui et al. EMNLP16)	VQA-CP	34.39	37.96	11.80	39.90	

Slide credit: Aishwarya Agrawal

Grounded-VQA (GVQA)



Slide credit: Aishwarya Agrawal

Aishwarya Agrawal
(Virginia Tech)

Dhruv Batra
(Georgia Tech / FAIR)

Devi Parikh
(Georgia Tech / FAIR)

Ani Kembhavi
(AI2)

C-VQA: Compositional VQA

Training

Q: What color is the plate?

A: Green

Q: What color are stop lights?

A: Red

Testing

Q: What color is the stop light?

A: Green

Q: What is the color of the plate?

A: Red

Aishwarya Agrawal
(Virginia Tech)

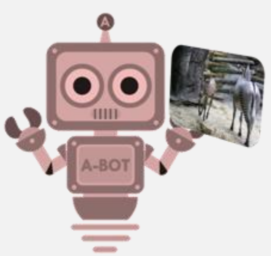
Dhruv Batra
(Georgia Tech / FAIR)

Devi Parikh
(Georgia Tech / FAIR)

Outline

Visual Question Answering

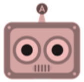
Visual Dialog



Visual Dialog

Visual Dialog

Visual Dialog



A man and a woman are holding umbrellas

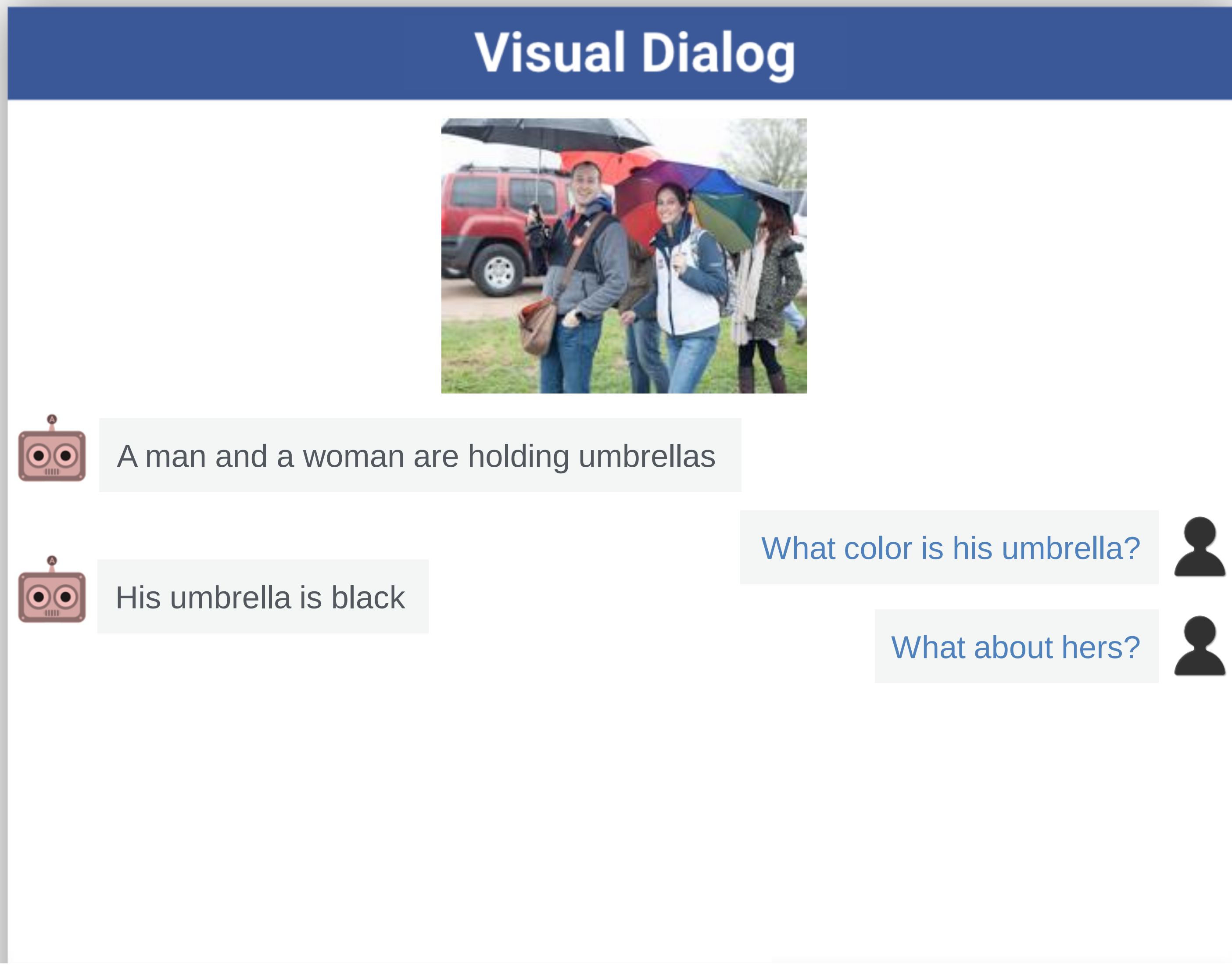

Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?


Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?

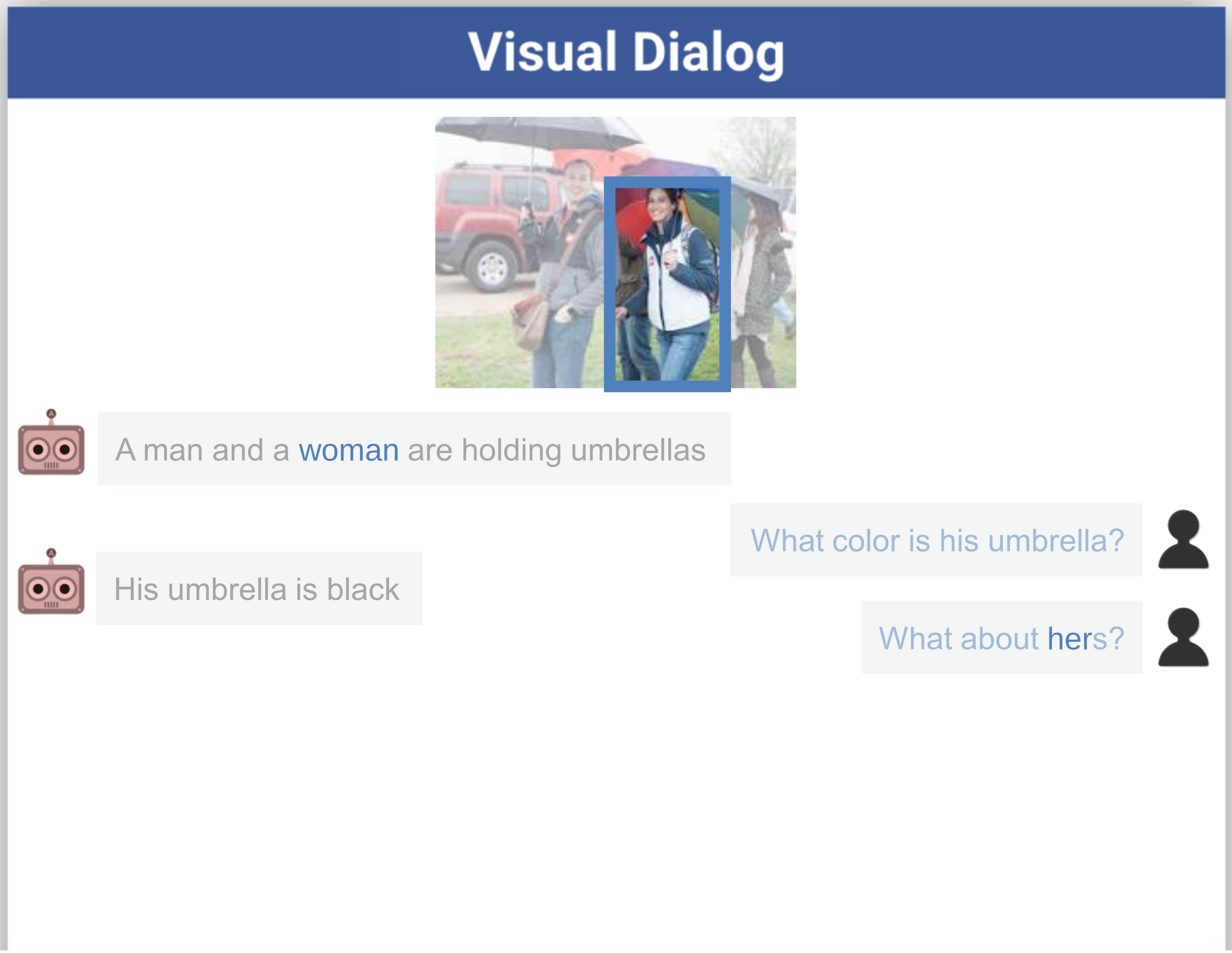

Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?

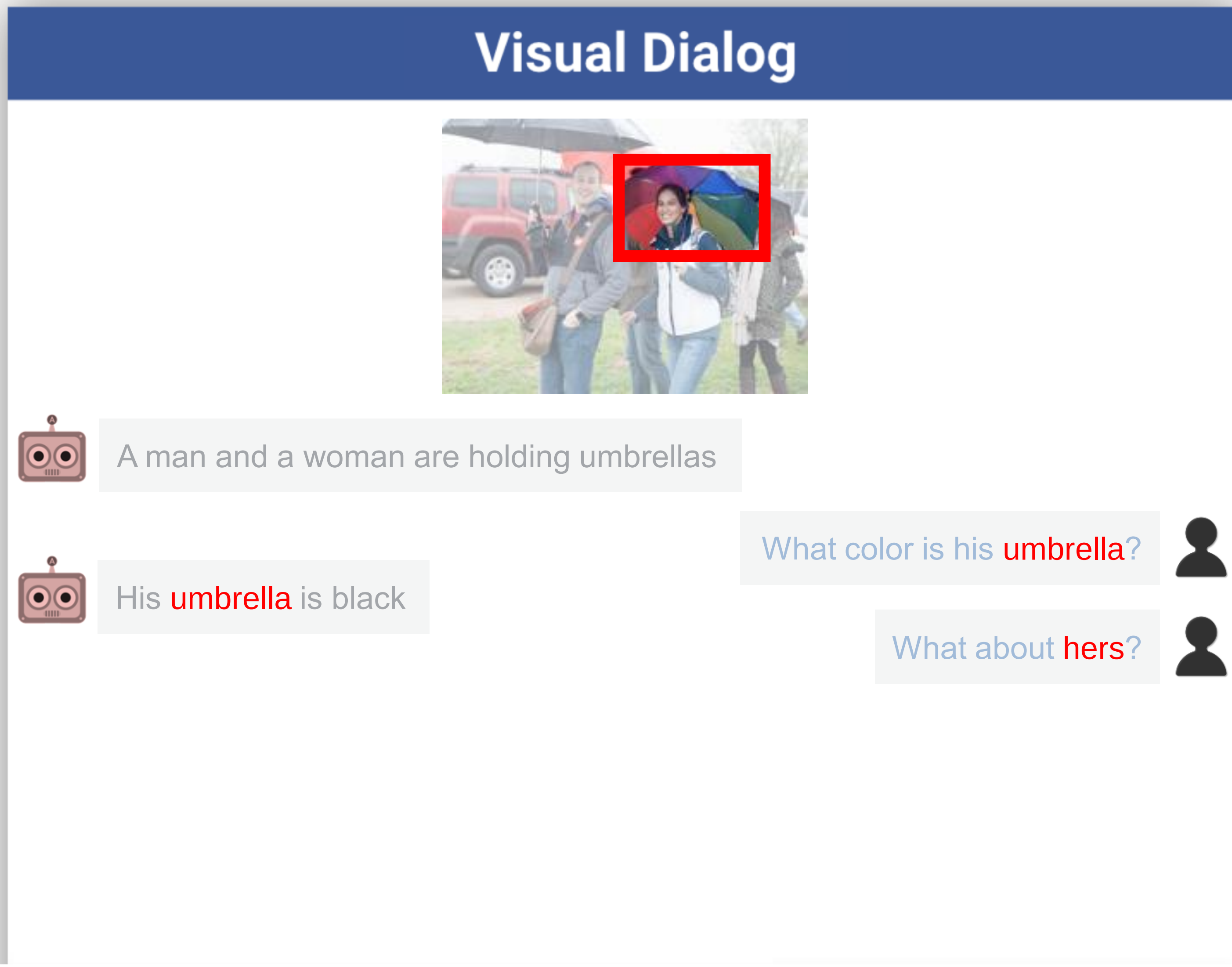
Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?

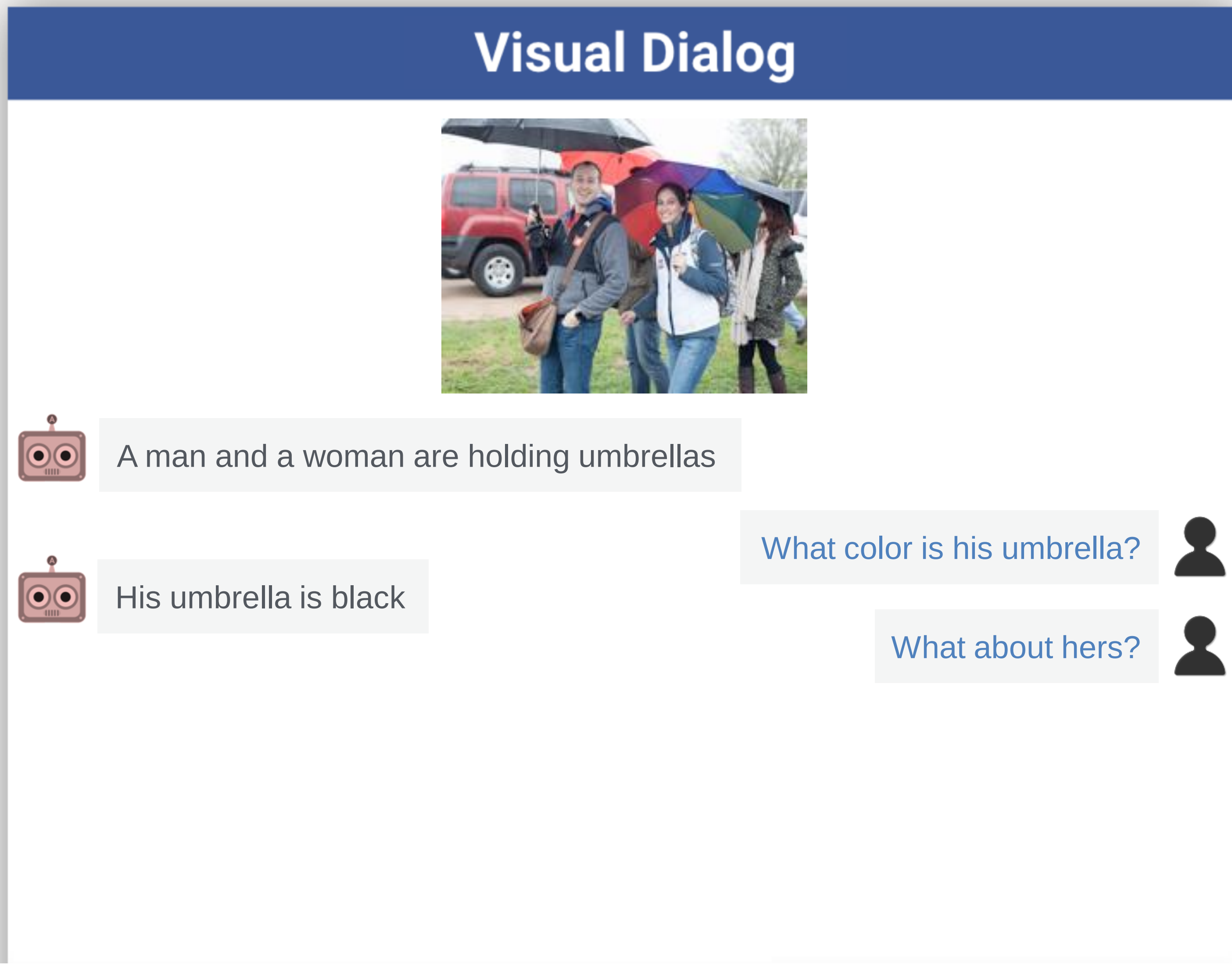

Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black



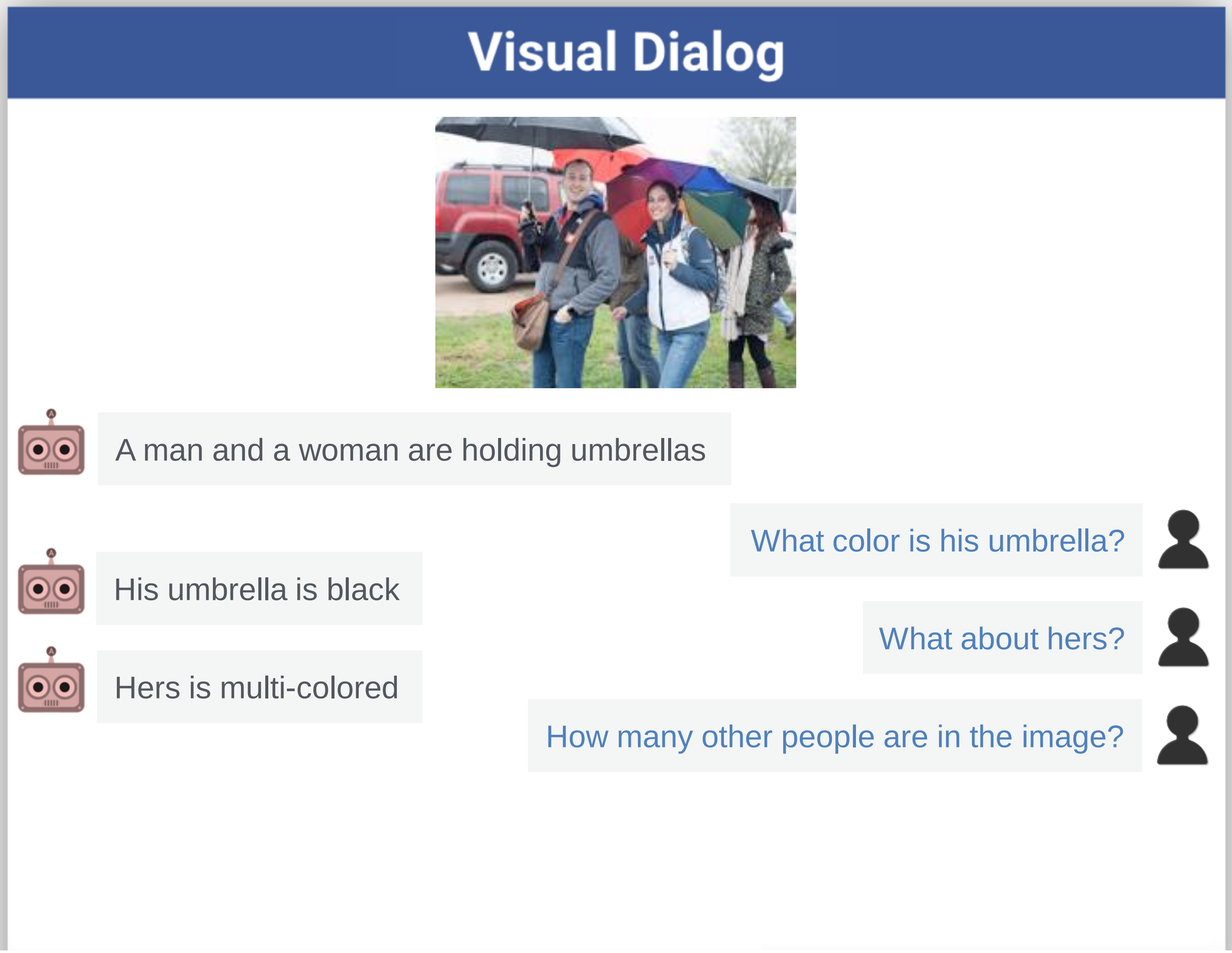
Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?

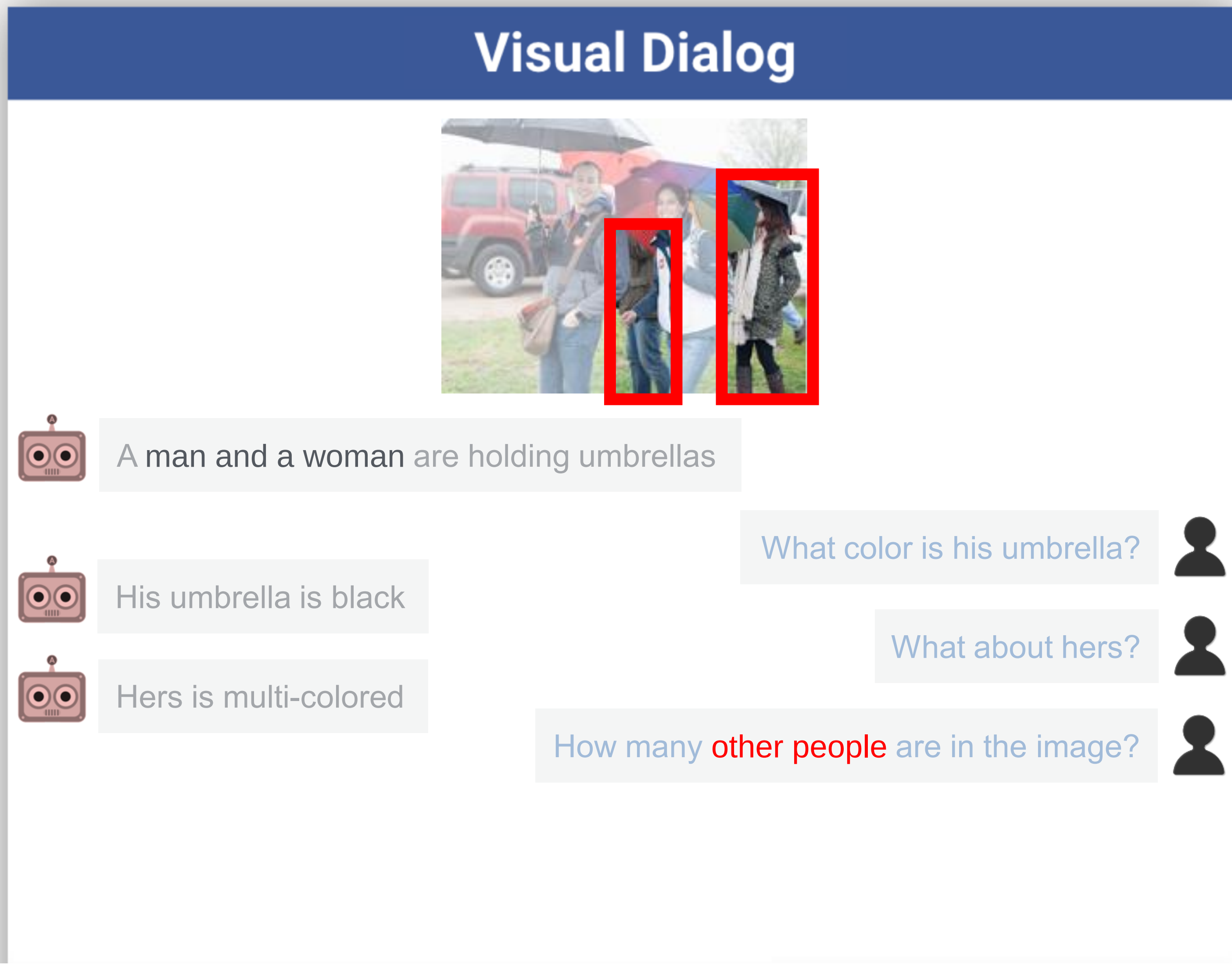

Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?


Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?


Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?


Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?
Hers is multi-colored


Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?
Hers is multi-colored
How many other people are in the image?


Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?
Hers is multi-colored
How many other people are in the image?


Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?
Hers is multi-colored
How many other people are in the image?
I think 3. They are occluded


Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?
Hers is multi-colored
How many other people are in the image?
I think 3. They are occluded
How many are men?


Visual Dialog
A man and a woman are holding umbrellas
What color is his umbrella?
His umbrella is black
What about hers?
Hers is multi-colored
How many other people are in the image?
I think 3. They are occluded
How many are men?

Visual Dialog: Task

- Given
  - Image I
  - History of human dialog $(Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1})$
  - Follow-up Question Q_t
- Task
  - Produce free-form natural language answer A_t

Visual Dialog

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.
Q : Is the other one holding anything?
A : He is not.

Visual Dialog: Evaluation Protocol

- Given
  - Image I
  - History of human dialog
    $(Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1})$
  - Follow-up Question Q_t
  - 100 Answer Options
    - 50 answers from NN questions
    - 30 popular answers
    - 20 random answers

- Evaluation Task
  - Rank the list of 100 options

- Accuracy/Error
  - mean-rank-of-GT, mean-reciprocal-rank

Visual Dialog

Question: Do people look happy ?

GT: Not really

- Yes they do
- I can't tell
- Not facing me
- Yes they look happy
- Yes I can only see 1 of their faces but she looks happy
- Not really but not unhappy either

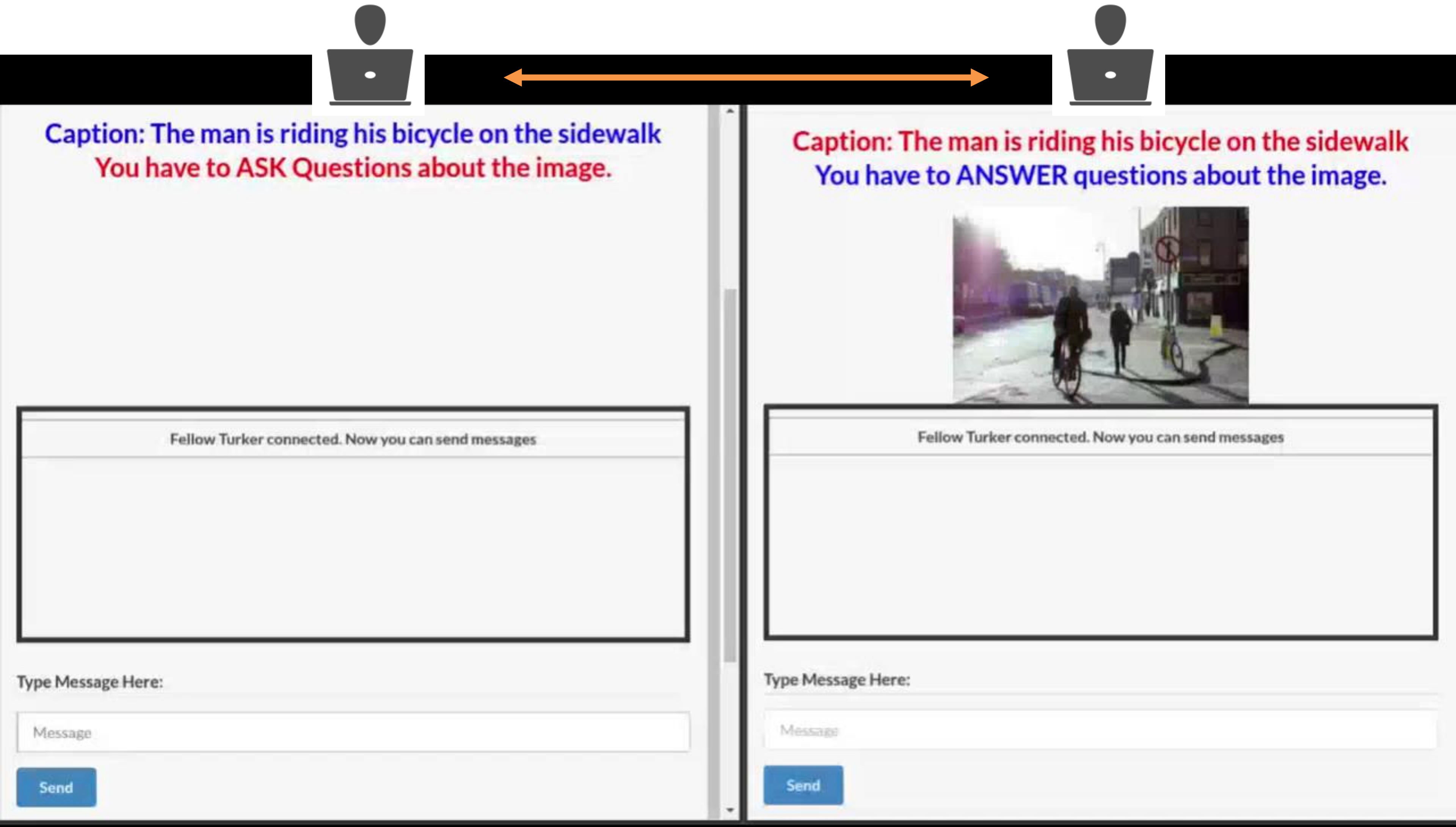
VisDial Dataset

Live Two-Person Chat on Amazon Mechanical Turk

VisDial Dataset

VisDial Dataset

Live Two-Person Chat on Amazon Mechanical Turk



batra-mlp-lab / visdial-amt-chat

Watch 7 | Star 19 | Fork 2

Code | Issues 0 | Pull requests 0 | Projects 0 | Pulse | Graphs

Code for the chat interface used to collect the VisDial dataset on AMT http://visualdialog.org/

1 commit | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Find file | Clone or download

abhshkdz Initial commit — Latest commit 4e7206e 6 days ago

mturk_scripts	Initial commit	6 days ago
nodejs	Initial commit	6 days ago
.gitignore	Initial commit	6 days ago
README.md	Initial commit	6 days ago
schema.sql	Initial commit	6 days ago

README.md

VisDial AMT Chat

Source for the two-person chat interface used to collect the VisDial dataset (arxiv.org/abs/1611.08669) on Amazon Mechanical Turk.

VisDial v0.9 Stats

>120k images (from COCO)

1 dialog/image

10 question-answer rounds/dialog

Total of *>1.2 Million* dialog QA pairs

visualdialog.org

Visual Dialog   Overview   People   Data   Bibtex   Acknowledgements   VirginiaTech   Georgia Tech

VisDial Dataset

Code for the real-time chat interface used to collect the VisDial dataset on Amazon Mechanical Turk

VisDial v0.9

Training set (235M)
82,783 images

Validation set (108M)
40,504 images

Readme

- v0.9 Training is from COCO Training and v0.9 Validation set is from COCO Validation
- Numbers (in papers, etc.) should be reported on v0.9 val

Format

```
[
  {
    'data': {
      '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',
        ...
```

Models for Visual Dialog

Encoder

1. Late Fusion
2. Hierarchical Recurrent Encoder
3. Memory Network

Decoder

1. Generative
   - During training, maximizes LL of human response
   - For evaluation, ranks options by LL scores
2. Discriminative
   - Learn to rank 100 options

Visual Dialog Model #3

Image I

Memory Network Encoder

Slide credit: Abhishek Das

Visual Dialog Model #3

Image I

Do you think the woman is with him?

Question Q_t

Memory Network Encoder

Slide credit: Abhishek Das

Visual Dialog Model #3

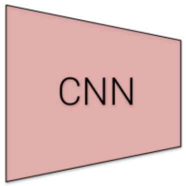


Image I

Do you think the woman is with him?

Question Q_t

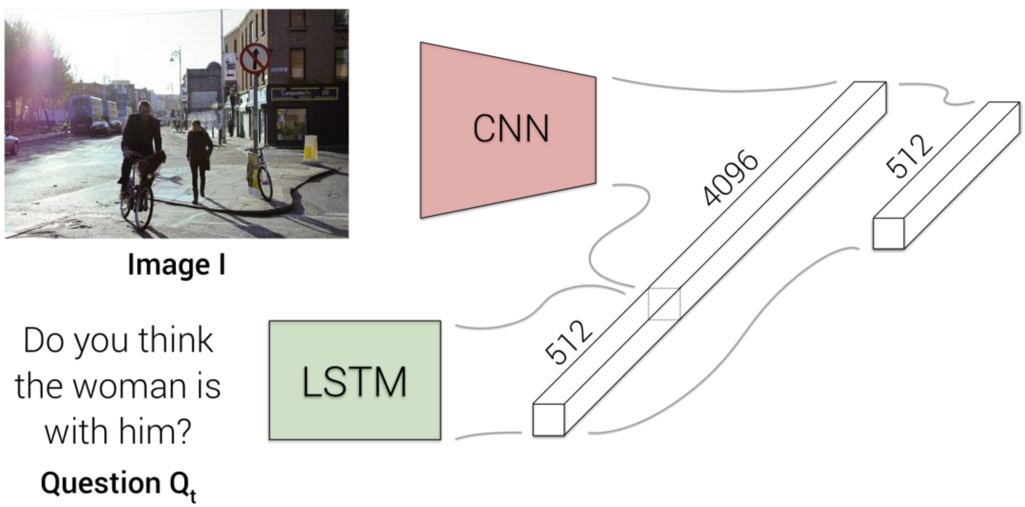
The man is riding his bicycle on the sidewalk.
Is the man wearing a helmet? No he does not have a helmet on.
How old is the man? He looks around 40 years old.
What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
Is anyone else riding a bike? No he's the only one.
Are there any people nearby? Yes there's a woman walking behind him.

t rounds of history
$\{(\text{Caption}), (Q_1,A_1), \ldots, (Q_{t-1}, A_{t-1})\}$

Memory Network Encoder

Slide credit: Abhishek Das

Visual Dialog Model #3



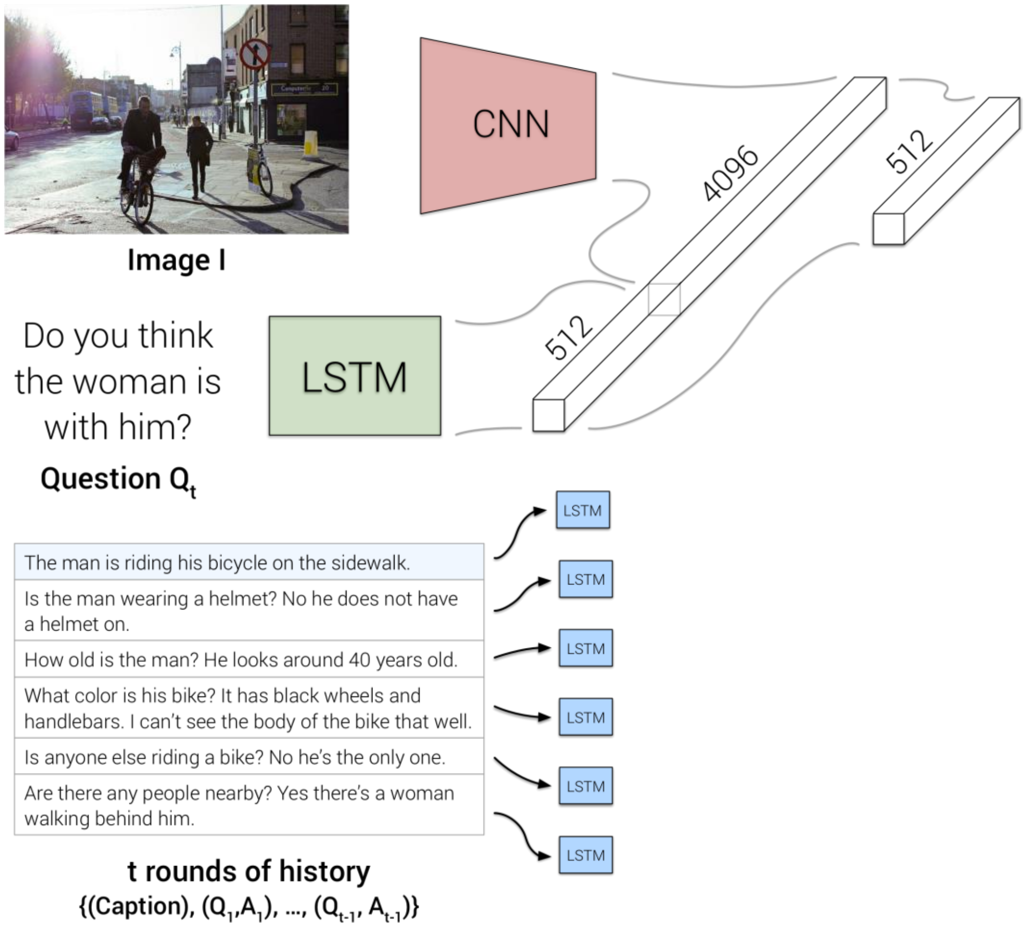
The man is riding his bicycle on the sidewalk.
Is the man wearing a helmet? No he does not have a helmet on.
How old is the man? He looks around 40 years old.
What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
Is anyone else riding a bike? No he's the only one.
Are there any people nearby? Yes there's a woman walking behind him.

t rounds of history
{(Caption), (Q_1,A_1), …, (Q_{t-1}, A_{t-1})}

Memory Network Encoder

Slide credit: Abhishek Das

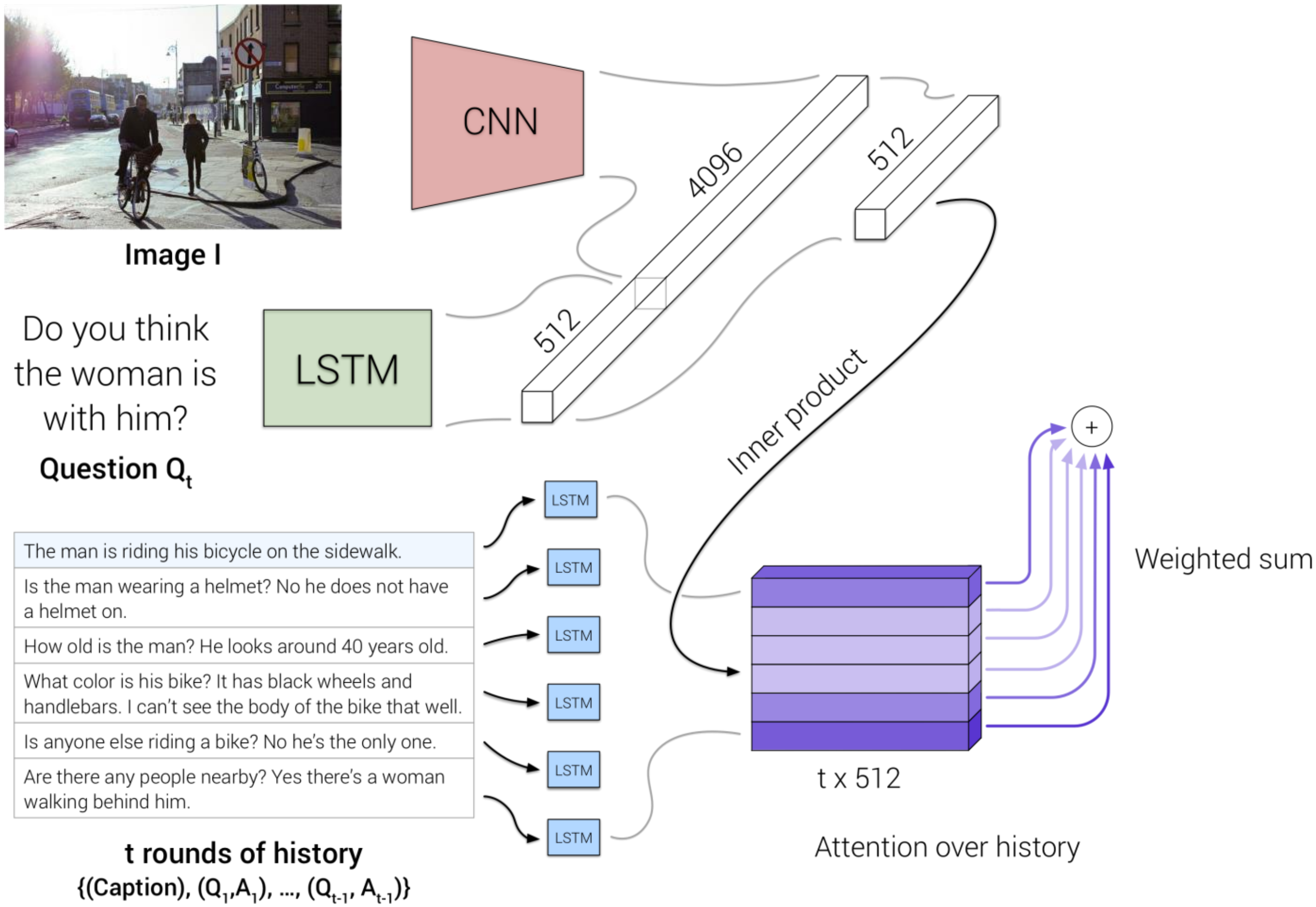
Visual Dialog Model #3



Memory Network Encoder

Slide credit: Abhishek Das

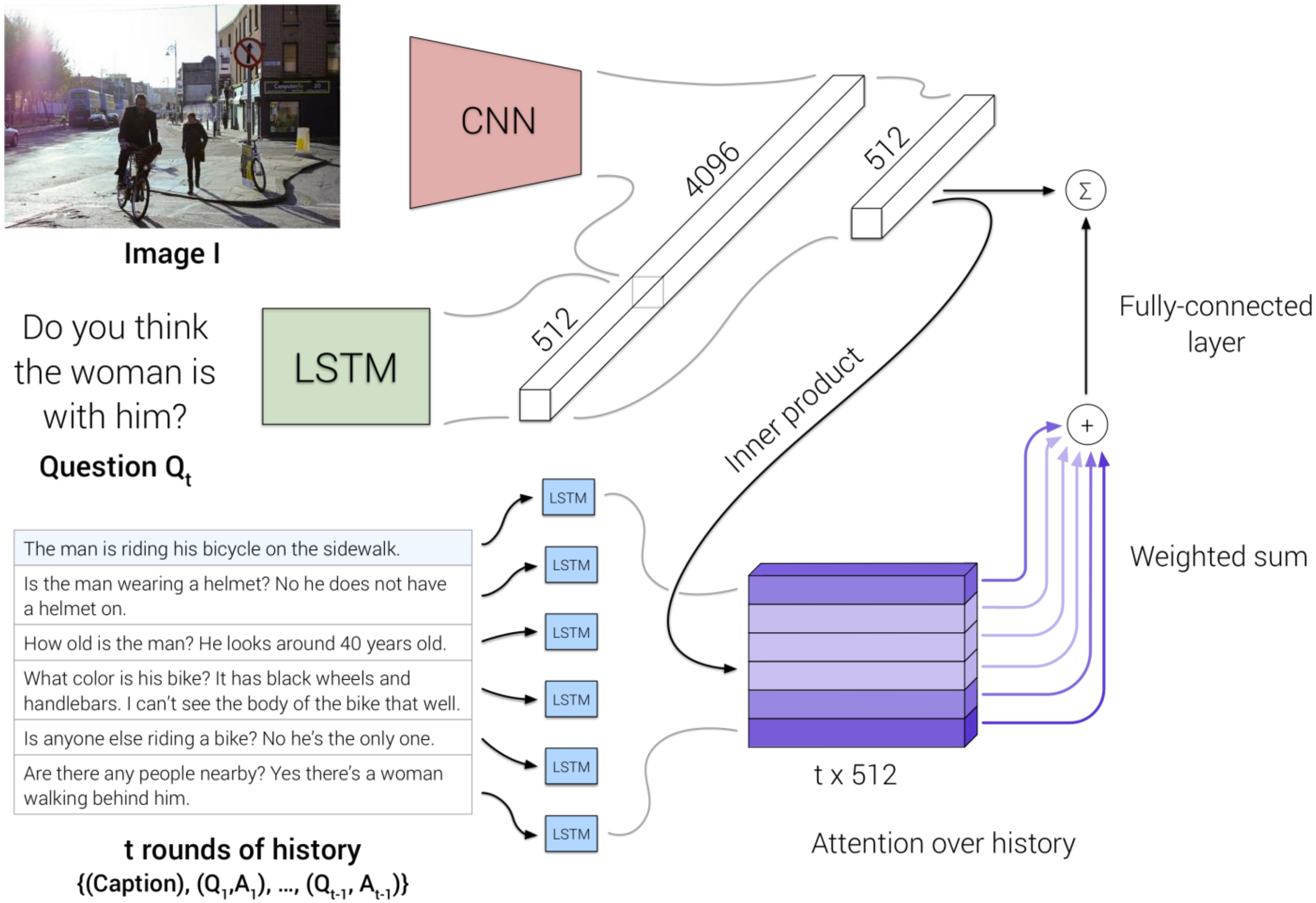
Visual Dialog Model #3



Memory Network Encoder

Slide credit: Abhishek Das

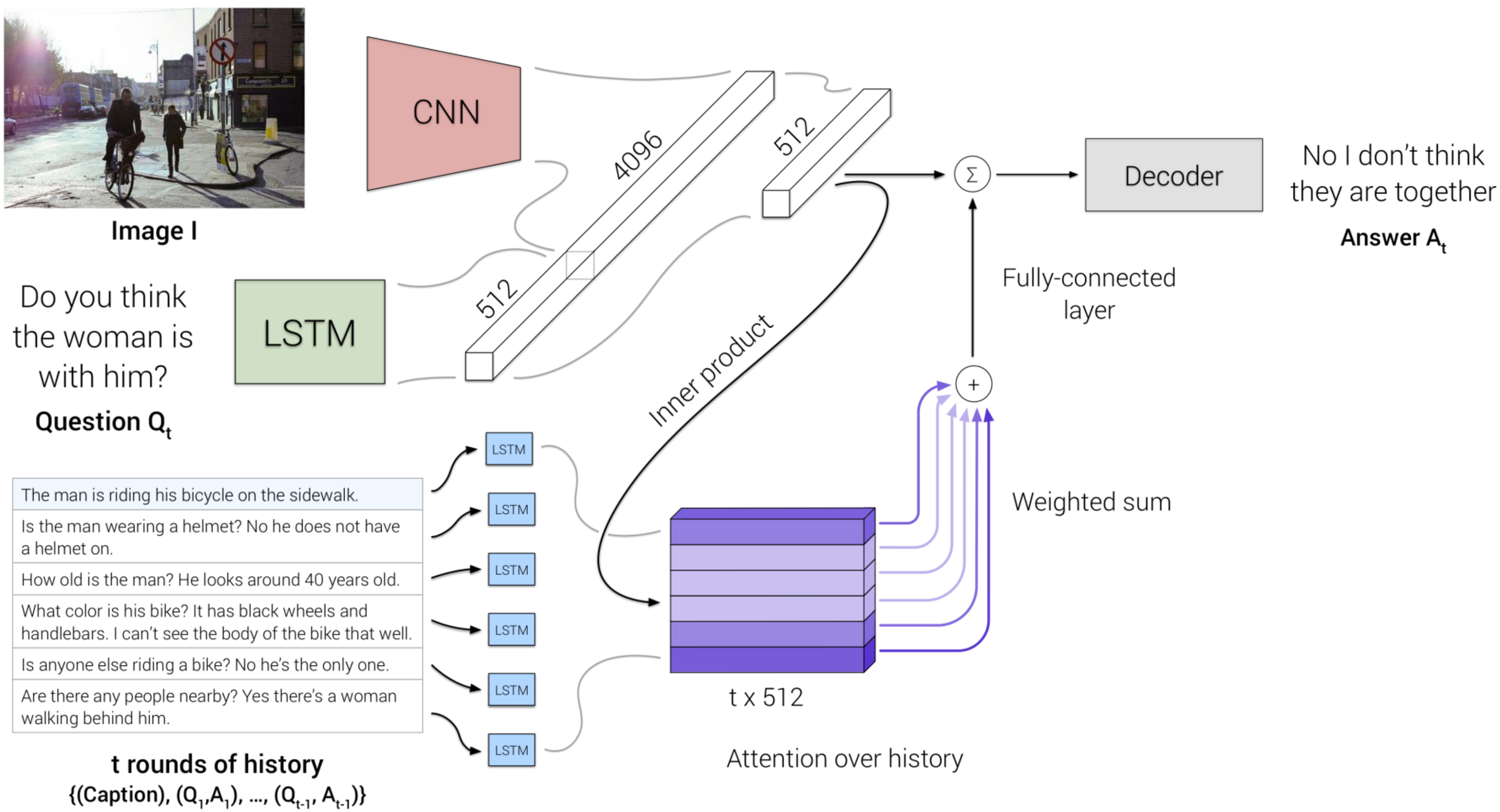
Visual Dialog Model #3



Memory Network Encoder

Slide credit: Abhishek Das

Visual Dialog Model #3



Memory Network Encoder

Slide credit: Abhishek Das

	Model	MRR	R@1	R@5	R@10	Mean
Baseline	Answer prior	0.3735	23.55	48.52	53.23	26.50
	NN-Q	0.4570	35.93	54.07	60.26	18.93
	NN-QI	0.4274	33.13	50.83	58.69	19.62
Generative	LF-Q-G	0.5048	39.78	60.58	66.33	17.89
	LF-QH-G	0.5055	39.73	60.86	66.68	17.78
	LF-QI-G	0.5204	42.04	61.65	67.66	16.84
	LF-QIH-G	0.5199	41.83	61.78	67.59	17.07
	HRE-QH-G	0.5102	40.15	61.59	67.36	17.47
	HRE-QIH-G	0.5237	**42.29**	62.18	67.92	17.07
	HREA-QIH-G	0.5242	42.28	62.33	68.17	**16.79**
	MN-QH-G	0.5115	40.42	61.57	67.44	17.74
	MN-QIH-G	**0.5259**	**42.29**	**62.85**	**68.88**	17.06
Discriminative	LF-Q-D	0.5508	41.24	70.45	79.83	7.08
	LF-QH-D	0.5578	41.75	71.45	80.94	6.74
	LF-QI-D	0.5759	43.33	74.27	83.68	5.87
	LF-QIH-D	0.5807	43.82	74.68	84.07	5.78
	HRE-QH-D	0.5695	42.70	73.25	82.97	6.11
	HRE-QIH-D	0.5846	44.67	74.50	84.22	5.72
	HREA-QIH-D	0.5868	44.82	74.81	84.36	5.66
	MN-QH-D	0.5849	44.03	75.26	84.49	5.68
	MN-QIH-D	**0.5965**	**45.55**	**76.22**	**85.37**	**5.46**
VQA	SAN1-QI-D	0.5764	43.44	74.26	83.72	5.88
	HieCoAtt-QI-D	0.5788	43.51	74.49	83.96	5.84

Results

- Memory Network (generally) performs best
  - 0.53 MRR / ~17 mean rank (Generative)
  - 0.60 MRR / ~5.5 mean rank (Discriminative)

Slide credit: Abhishek Das

batra-mlp-lab / **visdial**

Watch 15 | Star 82 | Fork 16

<> Code | Issues 2 | Pull requests 0 | Projects 0 | Pulse | Graphs

Visual Dialog code in Torch https://arxiv.org/abs/1611.08669

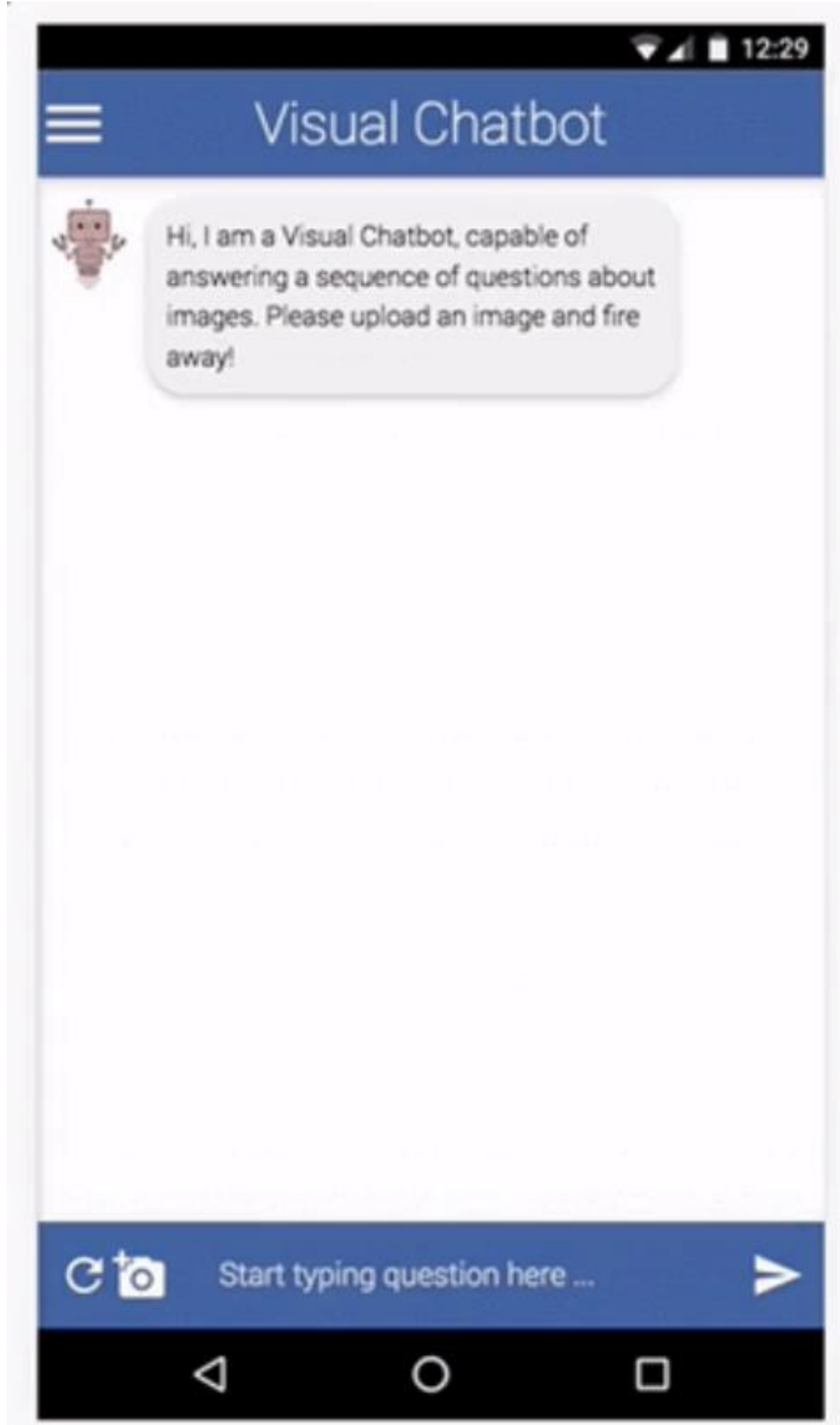
torch | computer-vision | natural-language-processing | deep-learning

3 commits | **1** branch | **0** releases | **1** contributor

Branch: **master** | New pull request | Find file | Clone or download

abhshkdz committed on **GitHub** Updates demo link — Latest commit 5b97b5b on Apr 19

data	Changes type to `int` for lengths; Refs #1	a month ago
decoders	Initial commit	a month ago
encoders	Initial commit	a month ago
model_utils	Initial commit	a month ago
scripts	Initial commit	a month ago
vis	Initial commit	a month ago
.gitignore	Initial commit	a month ago
README.md	Updates demo link	a month ago
dataloader.lua	Initial commit	a month ago
evaluate.lua	Initial commit	a month ago


12:29
Visual Chatbot
Hi, I am a Visual Chatbot, capable of answering a sequence of questions about images. Please upload an image and fire away!
Start typing question here ...

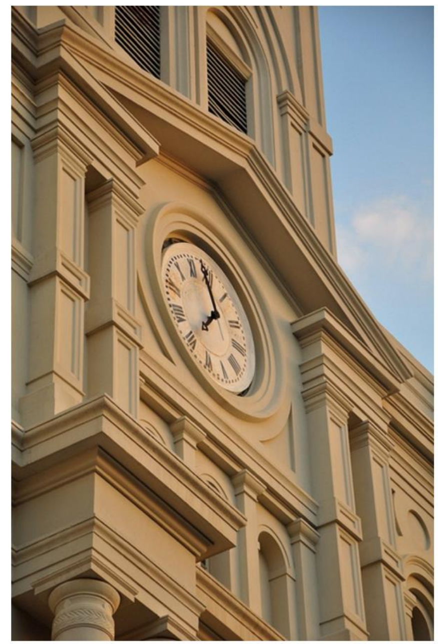
Results

Q: Is the bottle open?
A: Yes
Q: Red or white wine?
A: Red
Q: Is there only 1 glass?
A: Yes
Q: Any food?
A: No
Q: What color is the table?
A: Brown
Q: Is it in a restaurant?
A: No
Q: Is it in the kitchen?
A: Yes

Slide credit: Abhishek Das

Results

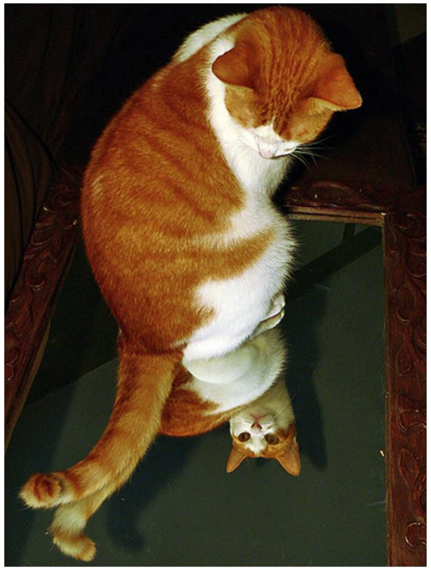
Q: Is the bottle open?
A: Yes
Q: Red or white wine?
A: Red
Q: Is there only 1 glass?
A: Yes
Q: Any food?
A: No
Q: What color is the table?
A: Brown
Q: Is it in a restaurant?
A: No
Q: Is it in the kitchen?
A: Yes



Q: What color is the building?
A: Brown
Q: What color is the clock?
A: Black and white
Q: Is it a digital or analog clock?
A: Analog
Q: What color are the numbers?
A: Black
Q: Are all of the numbers visible?
A: Yes
Q: Is it an outdoor view?
A: Yes
Q: Are there any people?
A: No
Q: Can you see other buildings
A: No

Slide credit: Abhishek Das

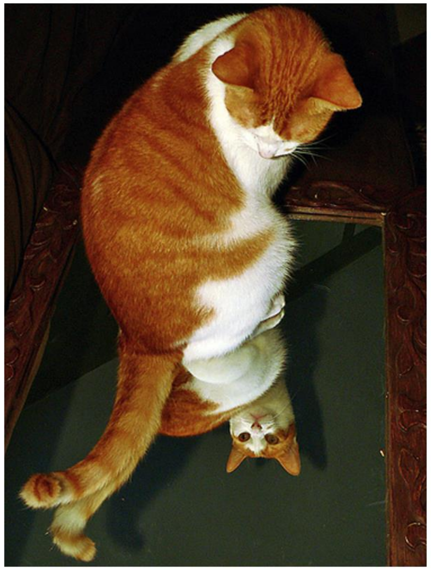
Results



Q: What color is the cat?
A: Orange and white
Q: Is the cat on the floor?
A: Yes
Q: What is the cat on?
A: I can't tell
Q: Can you see anything in the mirror?
A: No, just the cat and the cat
Q: Can you see a ceiling light?
A: No
Q: What color are the cat's eyes?
A: I can't see the eyes
Q: What color is the cat's nose?
A: Orange

Slide credit: Abhishek Das

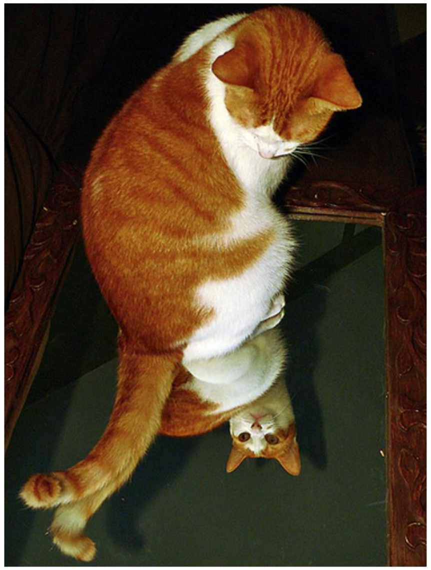
Results



Q: What color is the cat?
A: Orange and white
Q: Is the cat on the floor?
A: Yes
Q: What is the cat on?
A: I can't tell
Q: Can you see anything in the mirror?
A: No, just the cat and the cat
Q: Can you see a ceiling light?
A: No
Q: What color are the cat's eyes?
A: I can't see the eyes
Q: What color is the cat's nose?
A: Orange

Slide credit: Abhishek Das

Results



Q: What color is the cat?
A: Orange and white
Q: Is the cat on the floor?
A: Yes
Q: What is the cat on?
A: I can’t tell
Q: Can you see anything in the mirror?
A: No, just the cat and the cat
Q: Can you see a ceiling light?
A: No
Q: What color are the cat’s eyes?
A: I can’t see the eyes
Q: What color is the cat’s nose?
A: Orange

Slide credit: Abhishek Das

Results

Q: What color is the vase?
A: White
Q: Can you tell what kind of flowers?
A: I can’t tell
Q: Is there any water in the vase?
A: Yes
Q: How much water is there?
A: A small amount
Q: What is the vase on?
A: Can’t tell
Q: Does this look like someone’s home?
A: Can’t tell

Slide credit: Abhishek Das

Results

Q: What color is the vase?
A: White
Q: Can you tell what kind of flowers?
A: I can't tell
Q: Is there any water in the vase?
A: Yes
Q: How much water is there?
A: A small amount
Q: What is the vase on?
A: Can't tell
Q: Does this look like someone's home?
A: Can't tell

Slide credit: Abhishek Das

Visual Dialog (CVPR 2017)

Abhishek Das
(Georgia Tech)

Satwik Kottur
(CMU)

Khushi Gupta
(CMU)

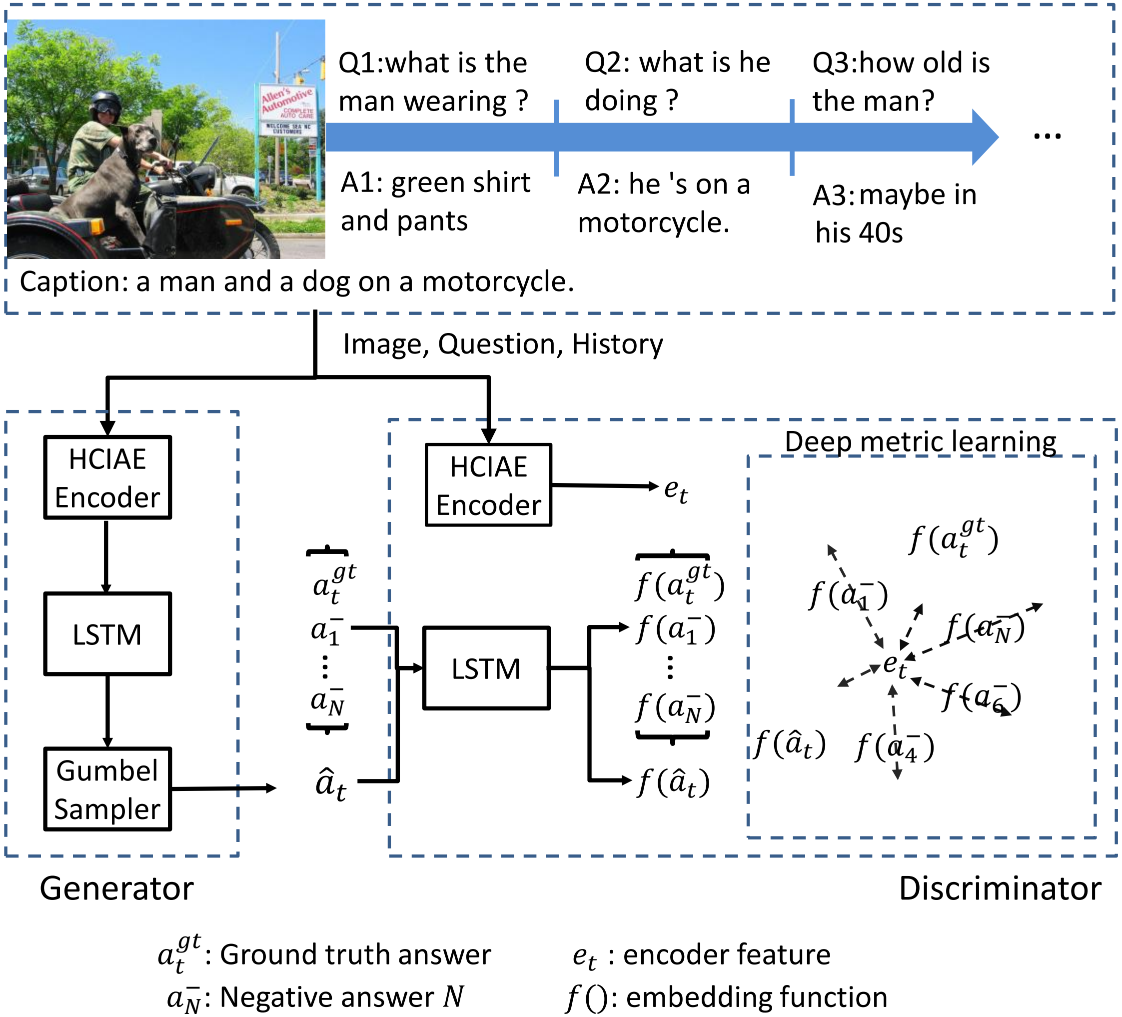
Avi Singh
(UC Berkeley)

Deshraj Yadav
(Virginia Tech)

José Moura
(CMU)

Devi Parikh
(Georgia Tech / FAIR)

Dhruv Batra
(Georgia Tech / FAIR)

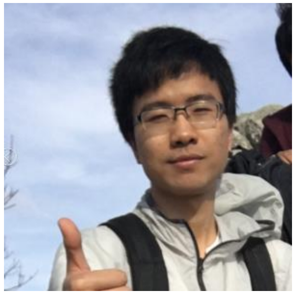


- Quantitative:
  - Ground truth response scores higher more often
- Qualitative:
  - Responses are more informative
  - Responses are longer
  - Responses are more diverse

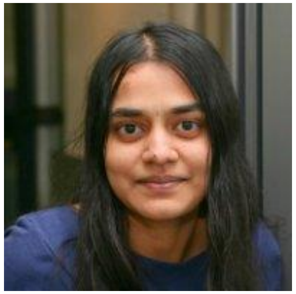
Slide credit: Jiasen Lu

Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model (arXiv)

Jiasen Lu
(Virginia Tech)



Jianwei Yang
(Georgia Tech)



Anitha Kannan
(Facebook AI Research)

Dhruv Batra
(Georgia Tech / FAIR)

Devi Parikh
(Georgia Tech / FAIR)

Open directions

- Improve dialog agents via self-talk
  - No additional human intervention
  - Are these agents better at human-bot interaction?

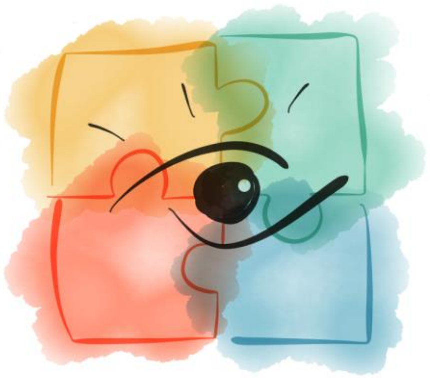
- Domain adaptation via self-talk
  - No need to collect a new dataset for each domain

- Dialog rollouts, future prediction, theory of mind, …

Conclusion

- Natural progression in Vision+Language
  - Captioning → VQA → Visual Dialog
- VQA: Elevating the role of image understanding
  - Balancing
  - Changing priors
  - Compositional
- Visual Dialog
  - New AI task
  - Challenges: Memory, history, reasoning over time
  - VisDial dataset
    - Live 2-person Chat on AMT
    - 120k COCO images, 1 dialog/image, ~*1.2 Million dialog QA pairs*
  - Visual Dialog Models (Neural Encoder-Decoders)
    - Late Fusion, Hierarchical Recurrent Encoder, Memory Network

Thank you.

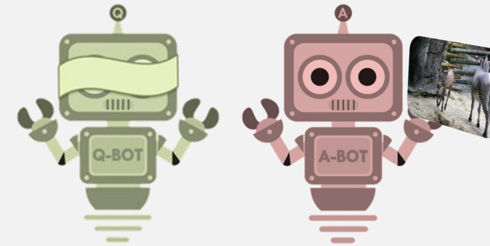


Visual Dialog:
Towards AI agents that can see, talk, and act

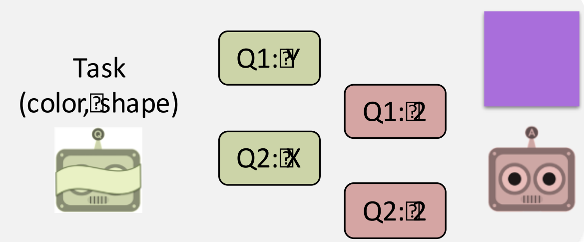
Dhruv Batra

Outline

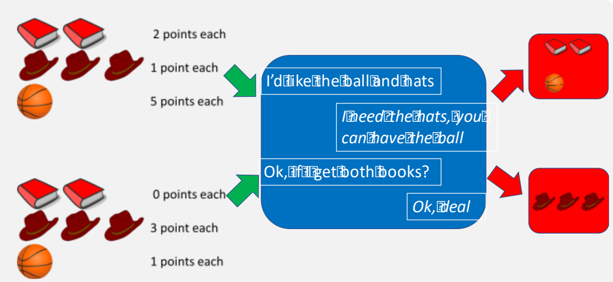
Cooperative Visual Dialog Agents



Emergence of Grounded Dialog



Negotiation Dialog Agents



Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

[ICCV ‘17]

Abhishek Das*
(Georgia Tech)

Satwik Kottur*
(CMU)

José Moura
(CMU)

Stefan Lee
(Virginia Tech)

Dhruv Batra
(Georgia Tech)

Visual Dialog: Task

- Given
  - Image I
  - History of human dialog $(Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1})$
  - Follow-up Question Q_t
- Task
  - Produce free-form natural language answer A_t

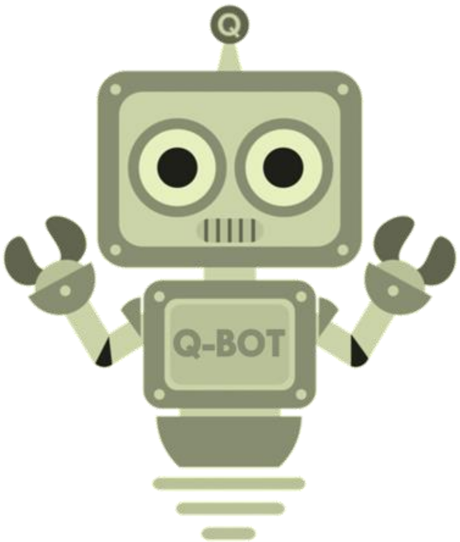
Visual Dialog

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.
Q : Is the other one holding anything?
A : He is not.

Problems

- No goal
  - Why are we talking?
- Agent not in control
  - Artificially injected at every round into a human conversation
  - Can’t steer conversation
  - Doesn’t get to see its errors during training
- Learning equivalent utterances
  - Many ways of answering the same question that should be treated equally, but aren’t
  - Is log-likelihood of human response really a good metric?

Image Guessing Game



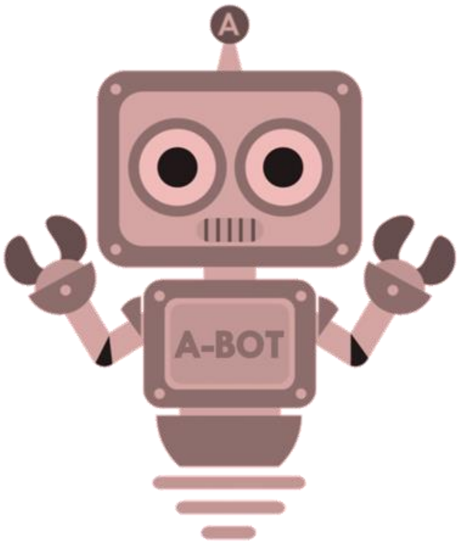
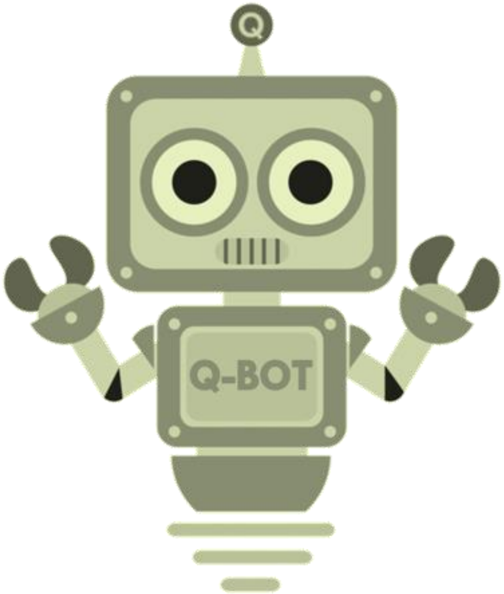
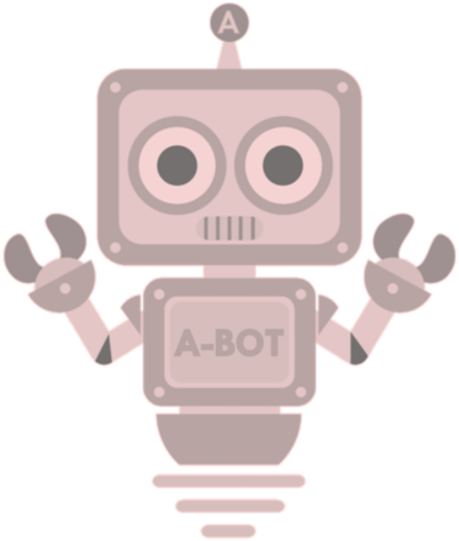


Image Guessing Game

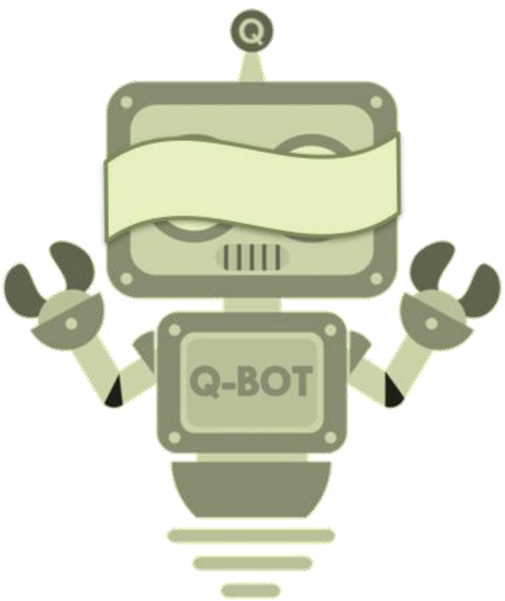


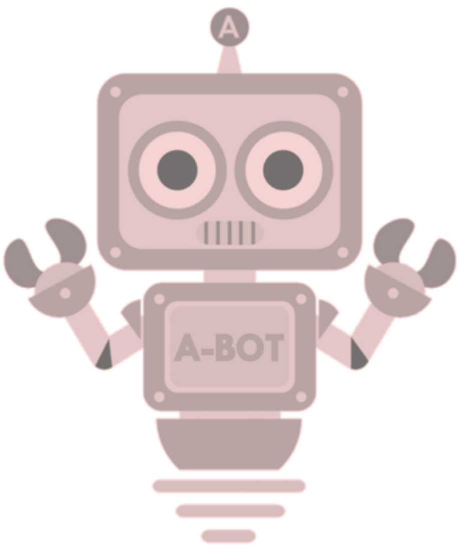


Q-Bot asks questions

Slide Credit: Abhishek Das

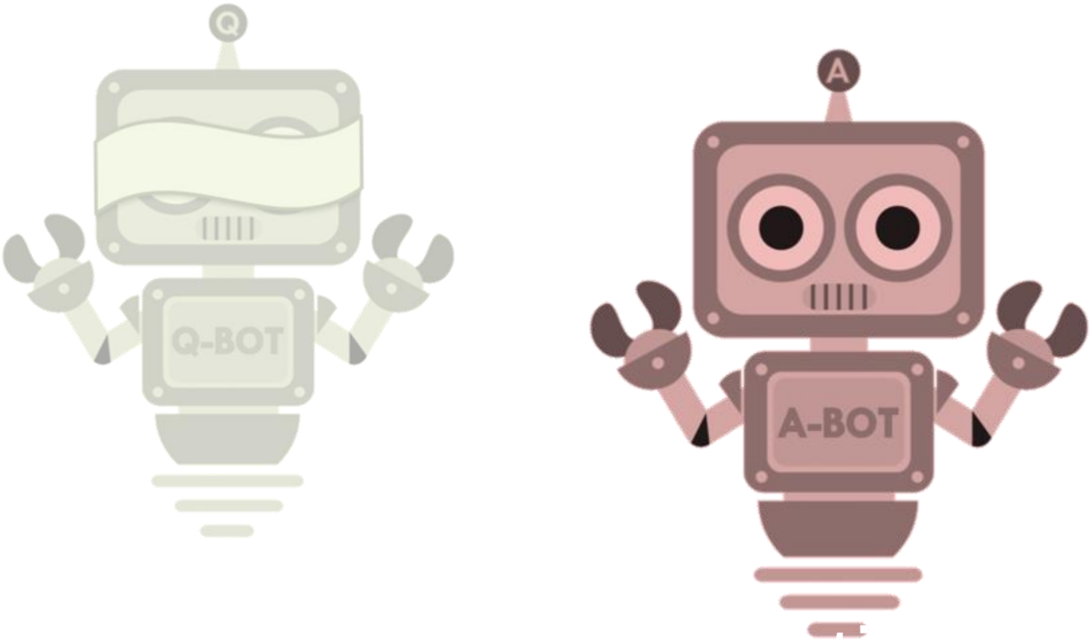
Image Guessing Game





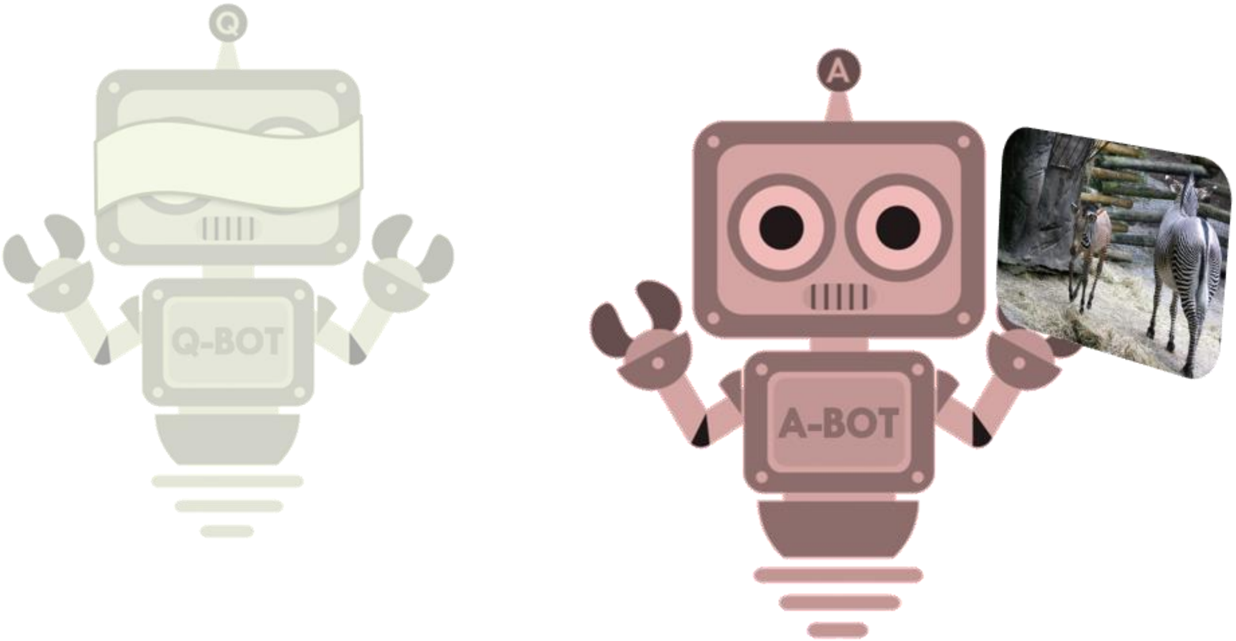
Q-Bot is blindfolded

Image Guessing Game



A-Bot answers questions

Image Guessing Game



A-Bot sees an image

Image Guessing Game

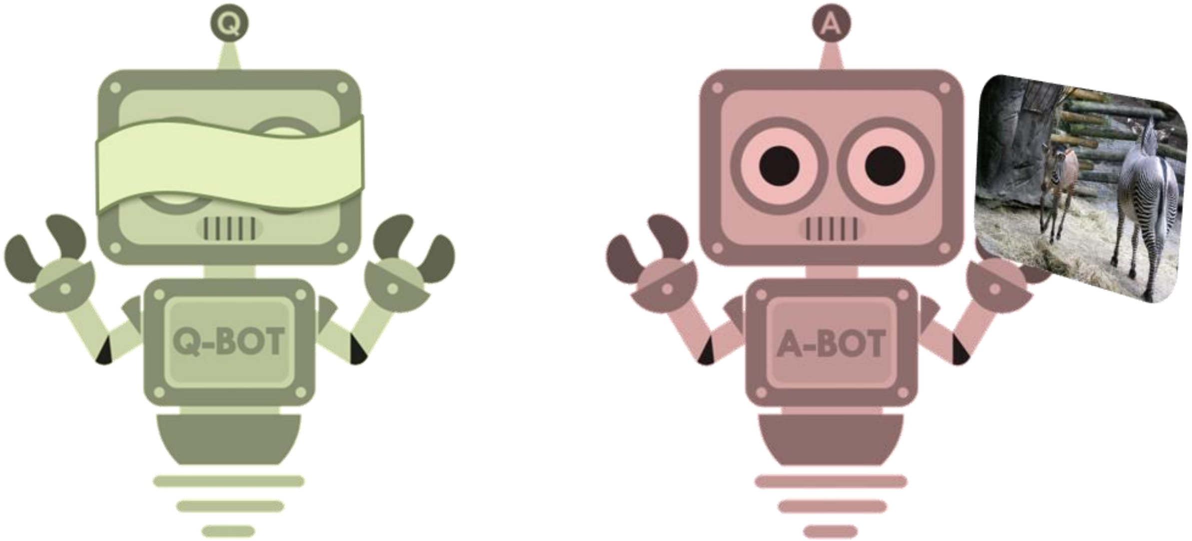


Image Guessing Game

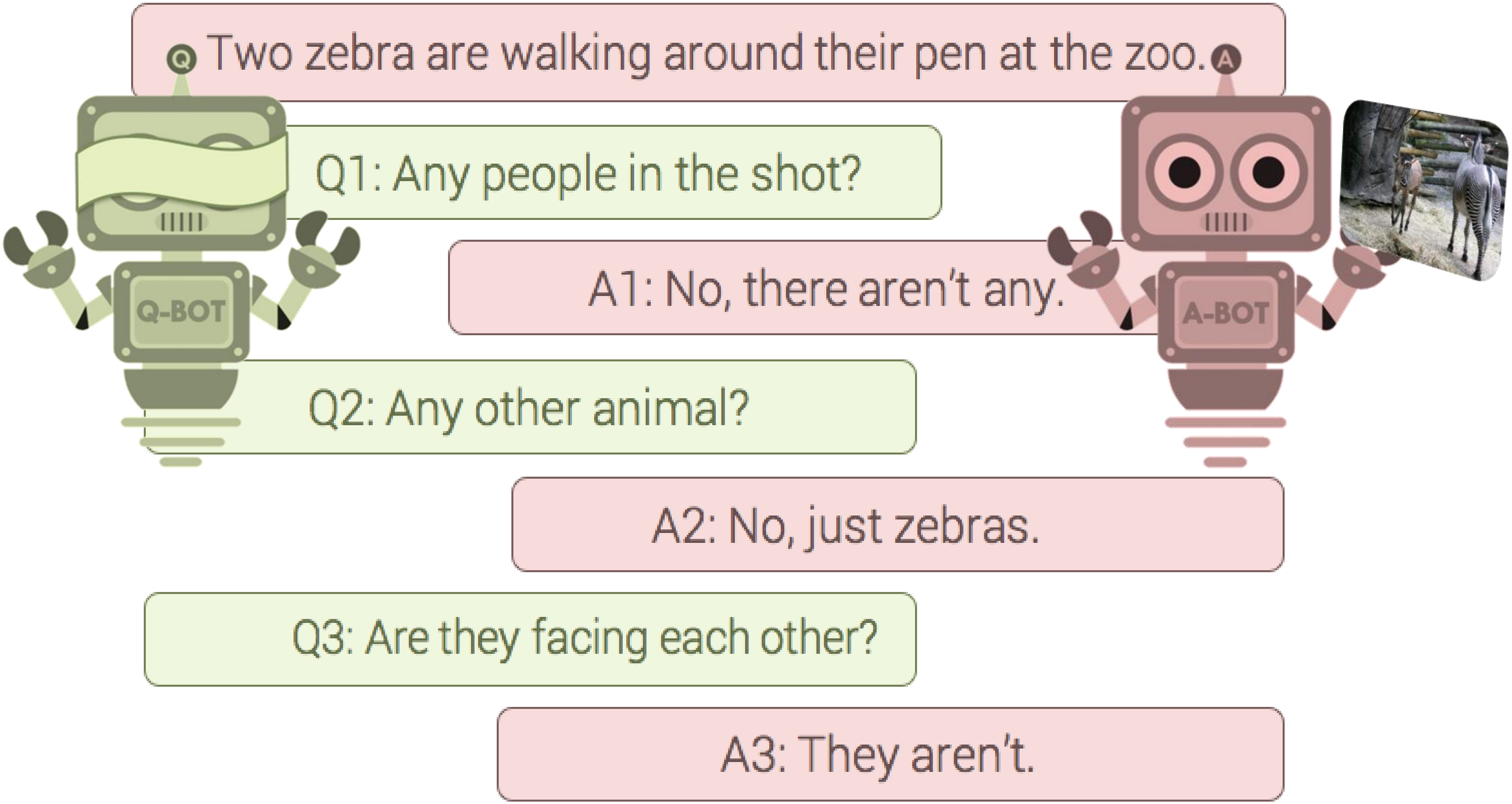
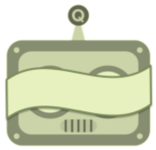
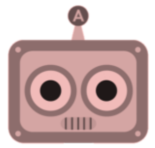


Image Guessing Game

RL for Cooperative Dialog Agents

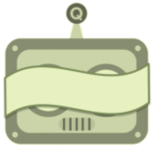


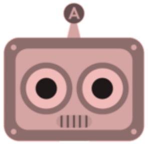


- Agents: (Q-bot, A-bot)

- Environment: Image
- Action:
  - Q-bot: question (symbol sequence) — q_t *Any people in the shot?*
  - A-bot: answer (symbol sequence) — a_t *No, there aren't any.*
  - Q-bot: image regression — $\hat{y}_t \in \mathbb{R}^{4096}$
- State
  - Q-bot: $s_t^Q = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}]$
  - A-bot: $s_t^A = [I, c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t]$

RL for Cooperative Dialog Agents

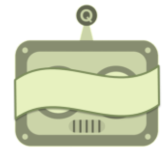


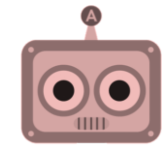


- Action:
  - Q-bot: question (symbol sequence) $\mathbf{q}_t$ *Any people in the shot?*
  - A-bot: answer (symbol sequence) $\mathbf{a}_t$ *No, there aren't any.*
  - Q-bot: image regression $\hat{y}_t \in \mathbb{R}^{4096}$

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

RL for Cooperative Dialog Agents





- Action:
  - Q-bot: question (symbol sequence)
  - A-bot: answer (symbol sequence)
  - Q-bot: image regression

q_t *Any people in the shot?*

a_t *No, there aren't any.*

$\hat{y}_t \in \mathbb{R}^{4096}$

- 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	A-bot
$\pi_Q(q_t \vert S_{t-1}^Q)$	$\pi_A(a_t \vert S_{t-1}^A)$

- Reward

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

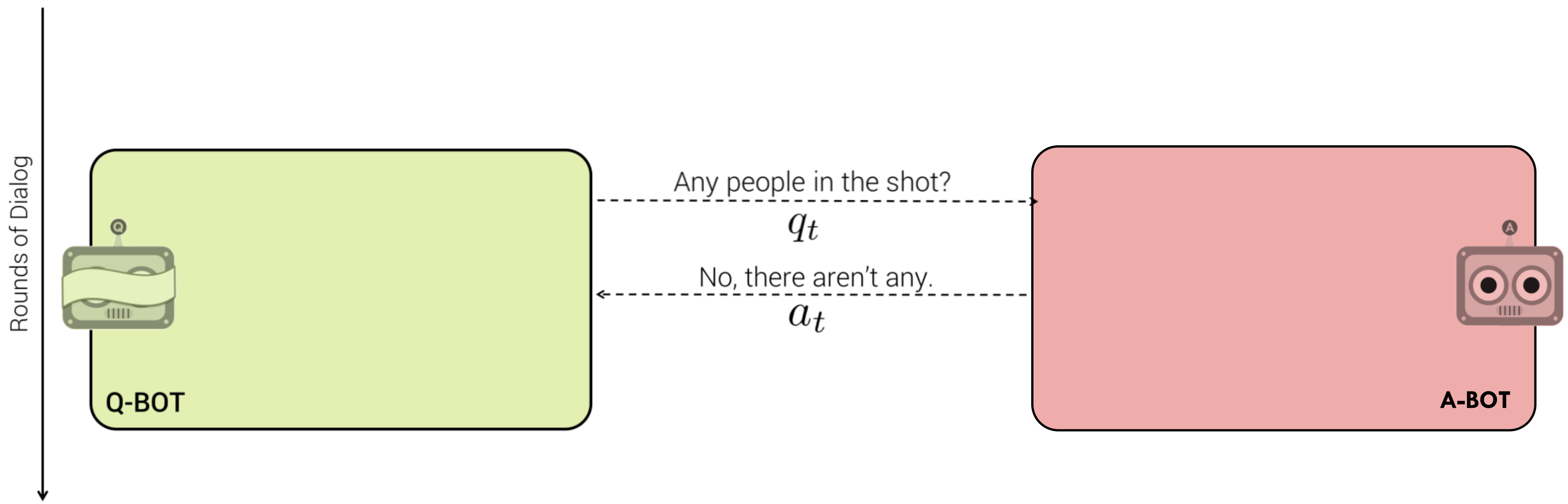
Policy Networks

Q-Bot

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

A-Bot

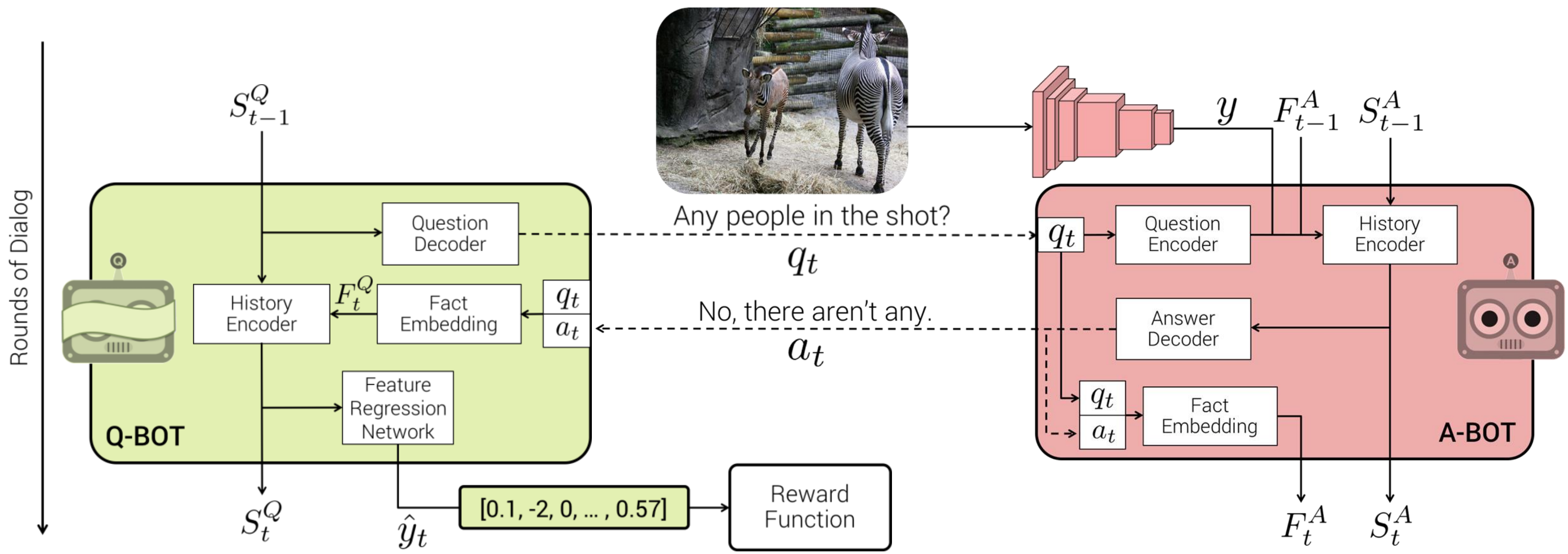
$\pi_A(a_t|S^A_{t-1})$



Policy Networks

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

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

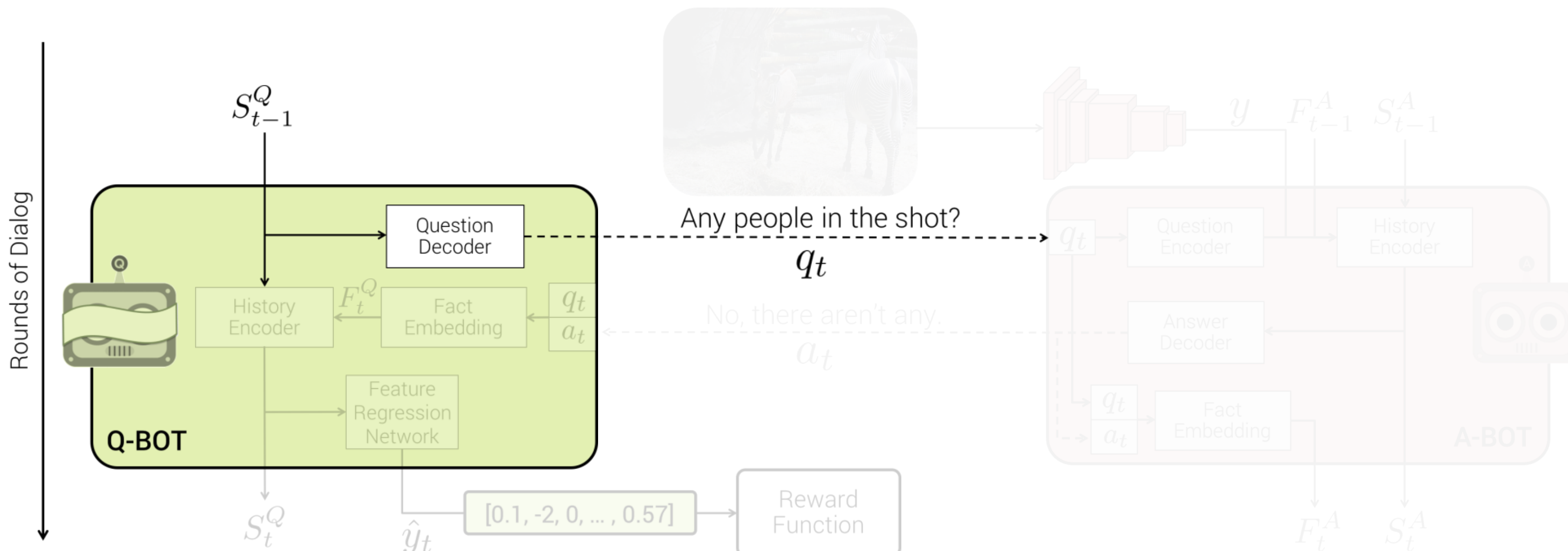


Slide Credit: Abhishek Das

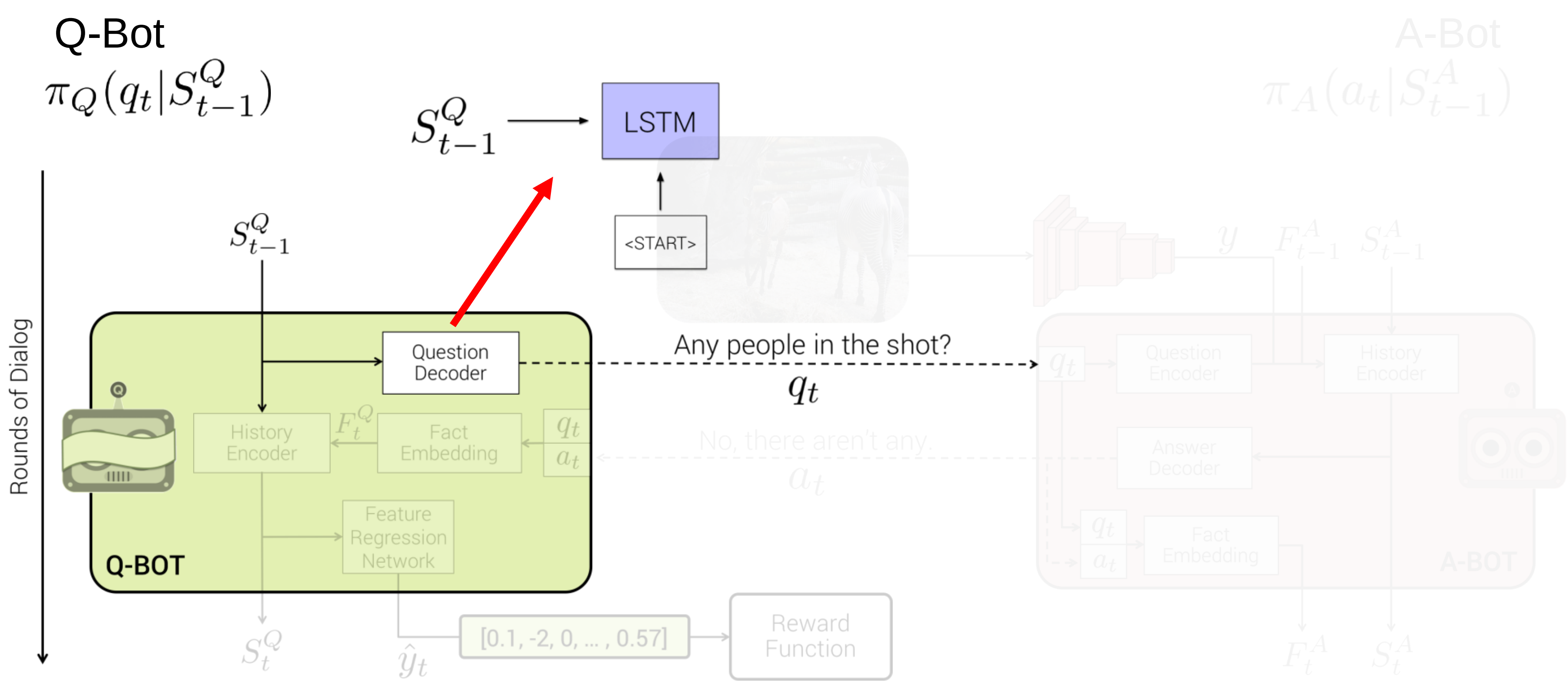
Policy Networks

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

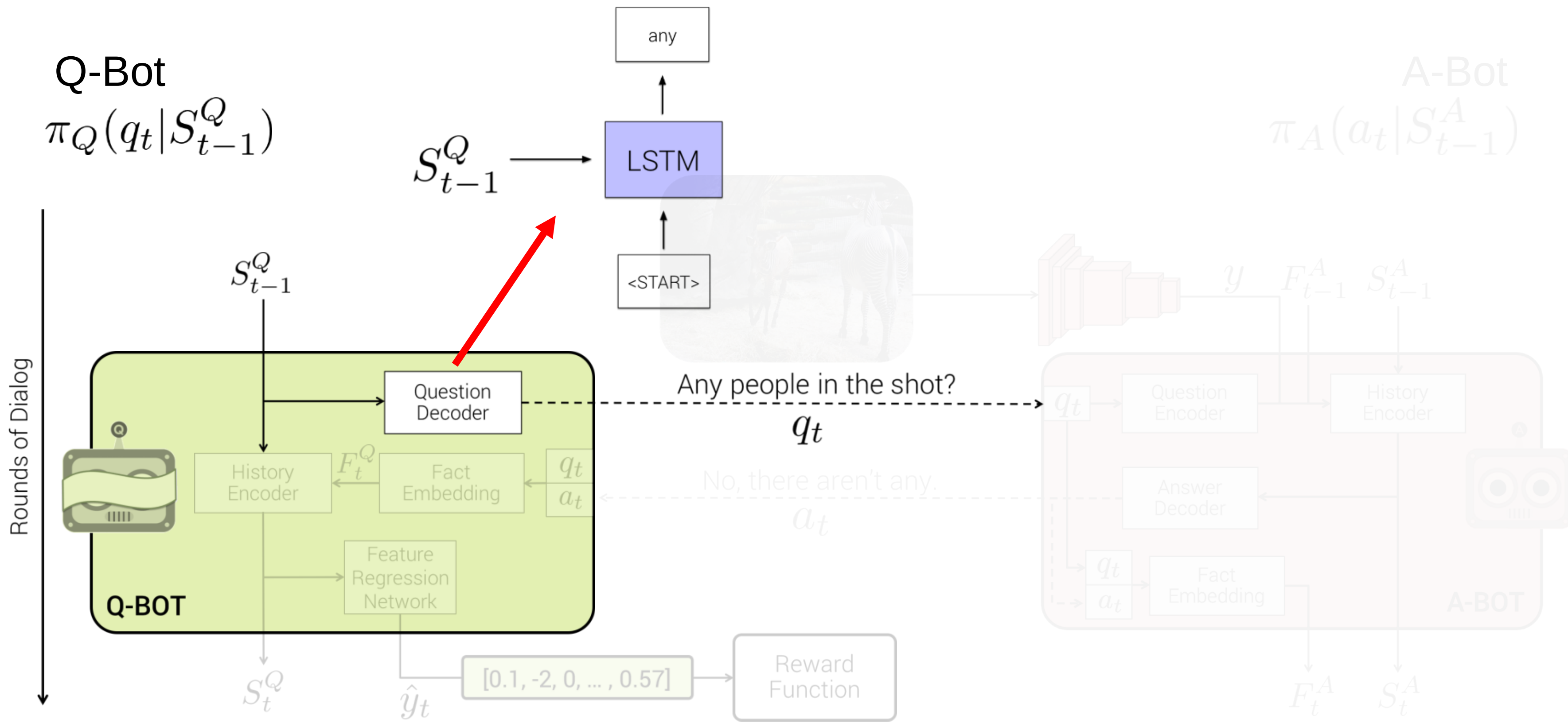
A-Bot
$\pi_A(a_t|S^A_{t-1})$



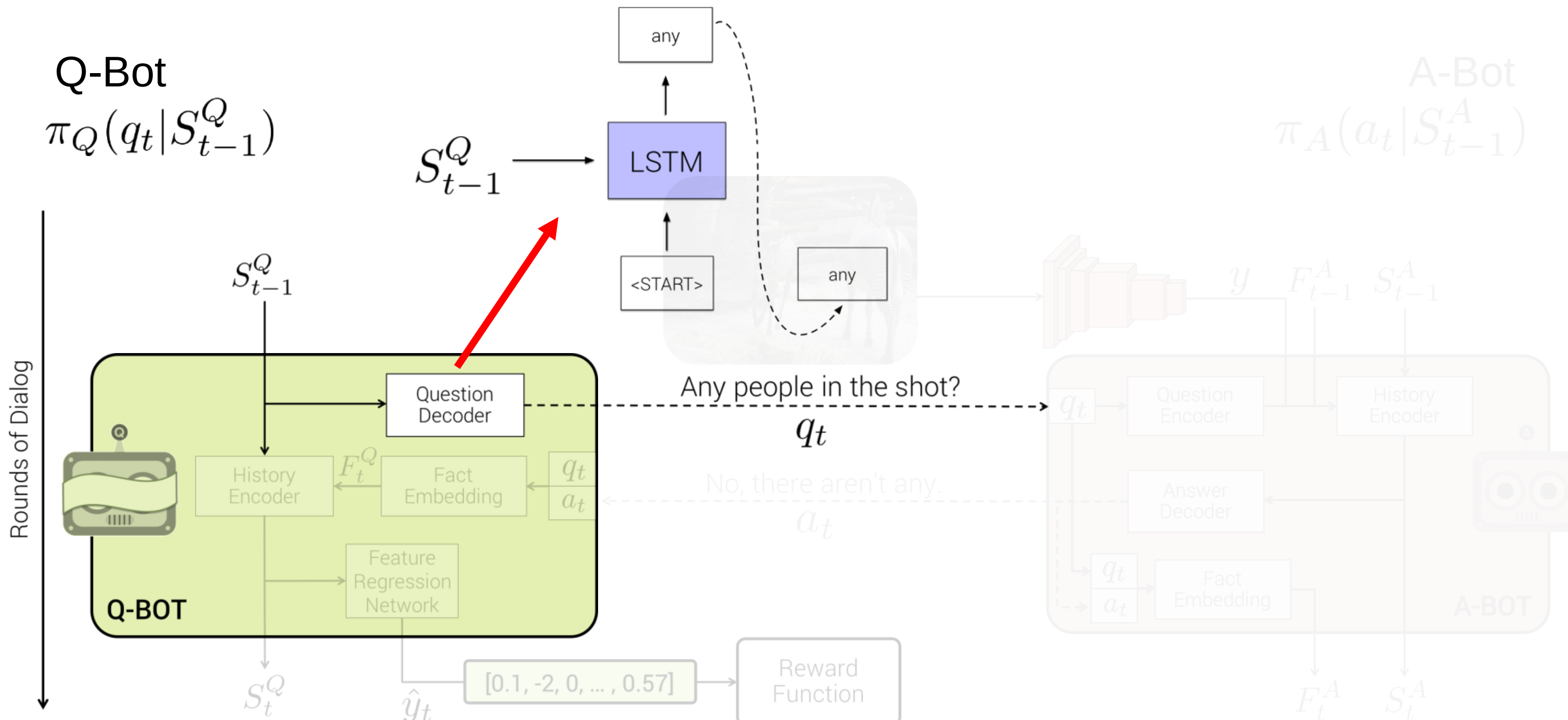
Policy Networks



Policy Networks



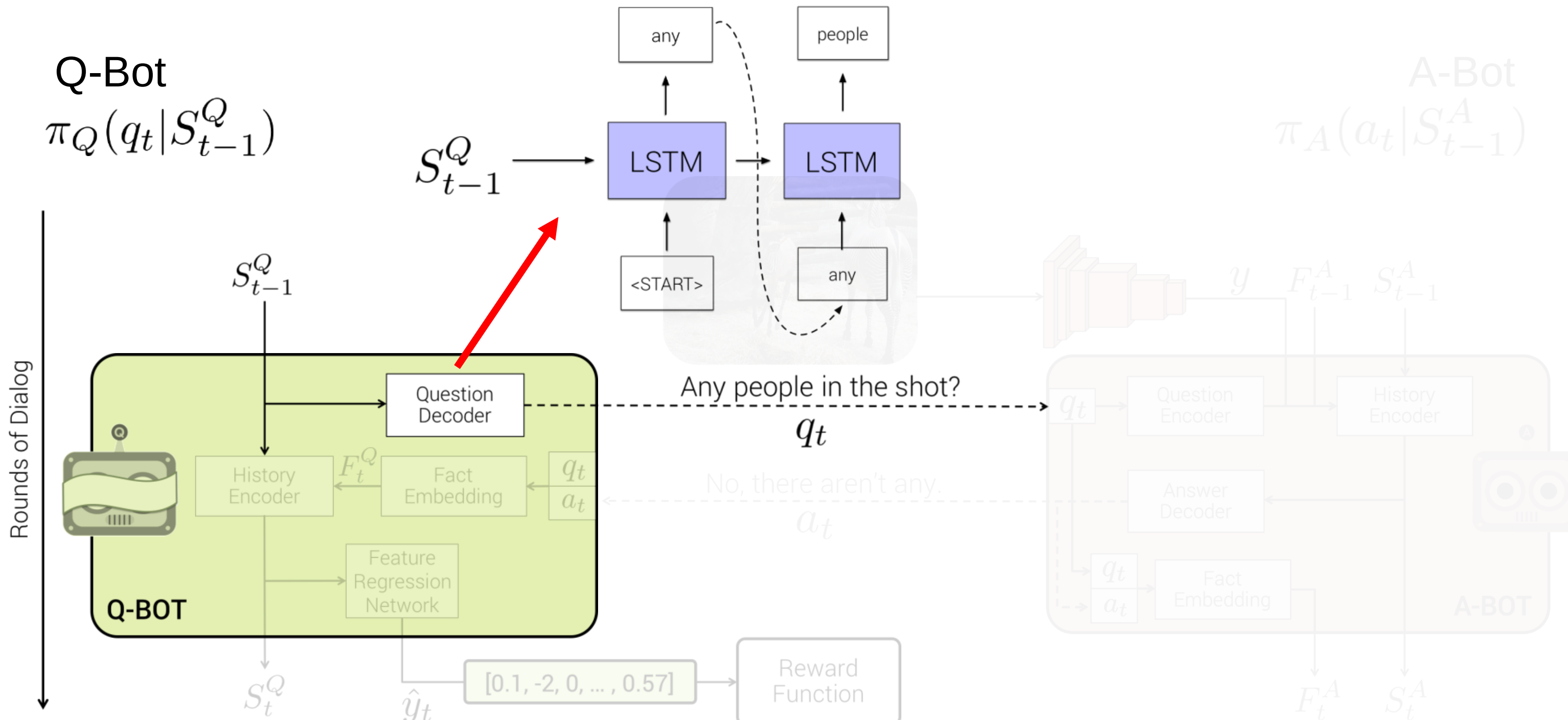
Policy Networks



Policy Networks

Q-Bot

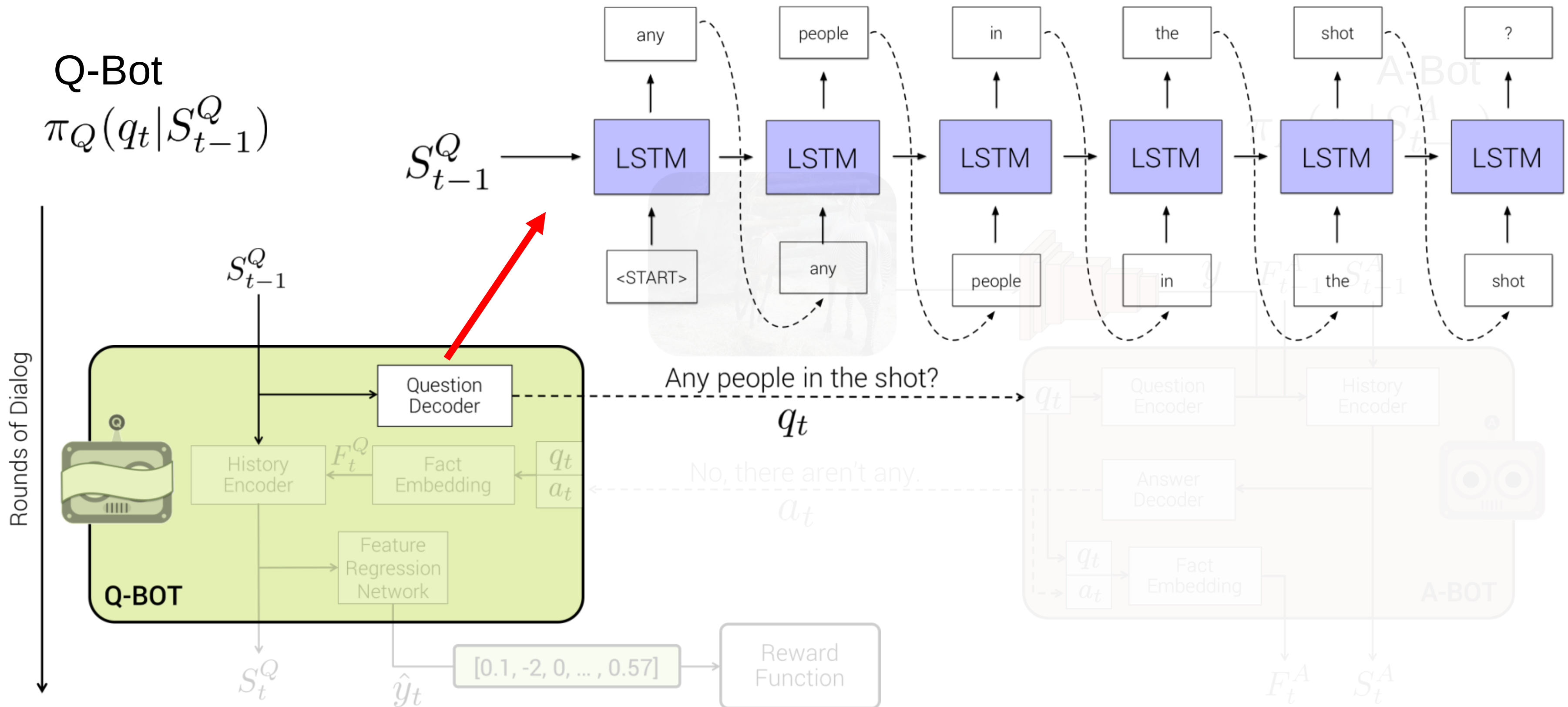
$\pi_Q(q_t | S^Q_{t-1})$



Policy Networks

Q-Bot

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



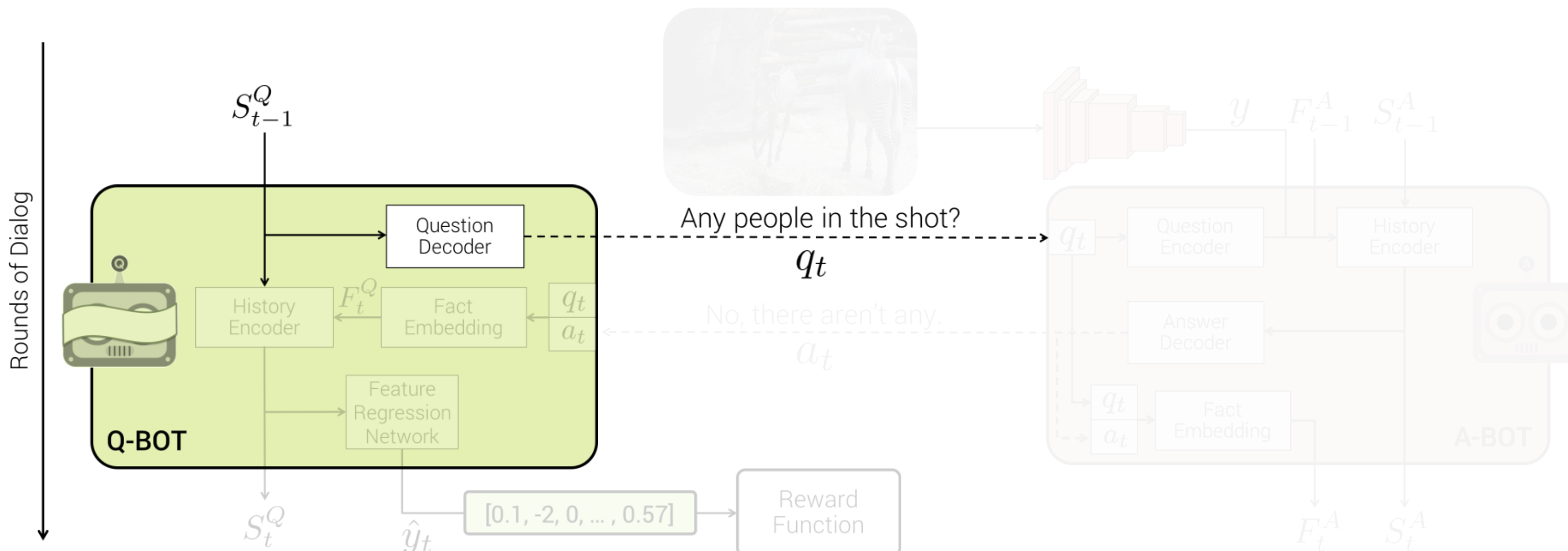
Policy Networks

Q-Bot

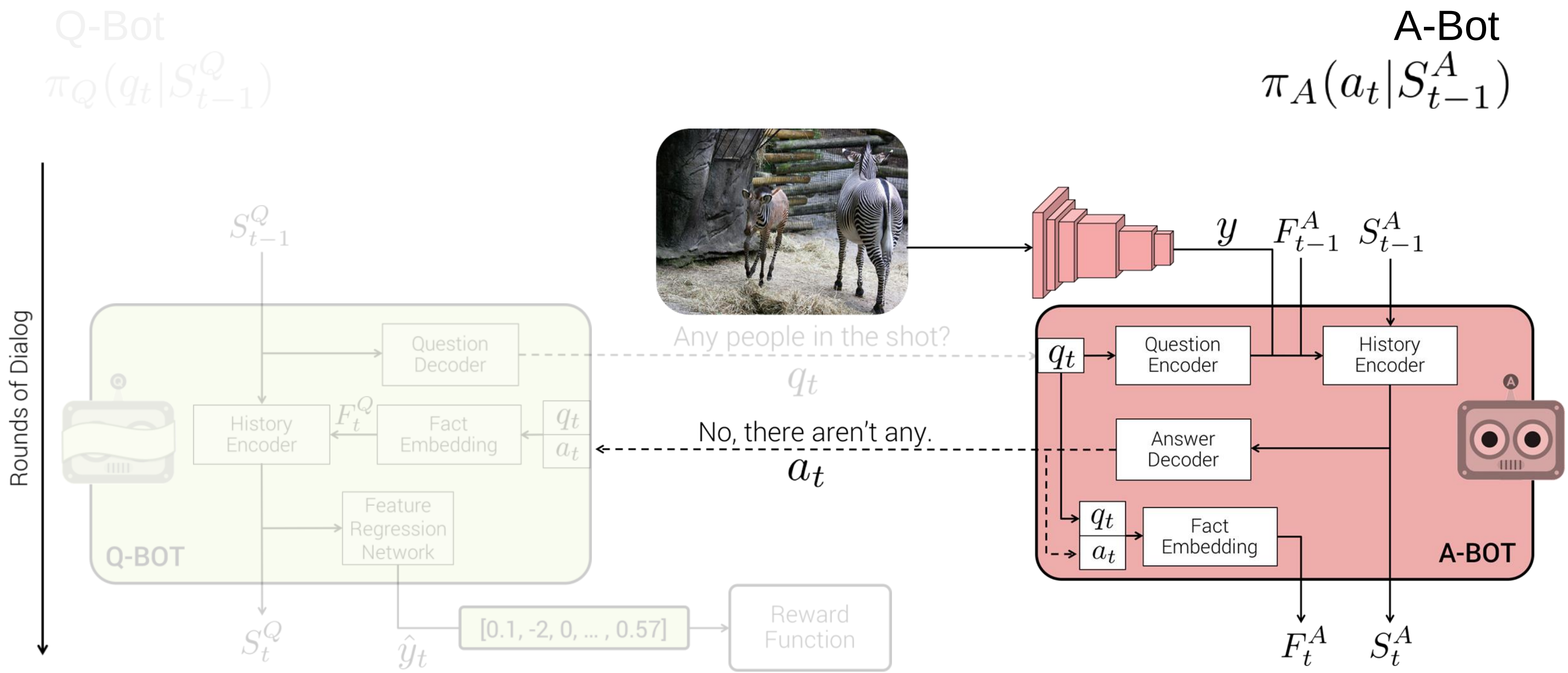
$\pi_Q(q_t | S^Q_{t-1})$

A-Bot

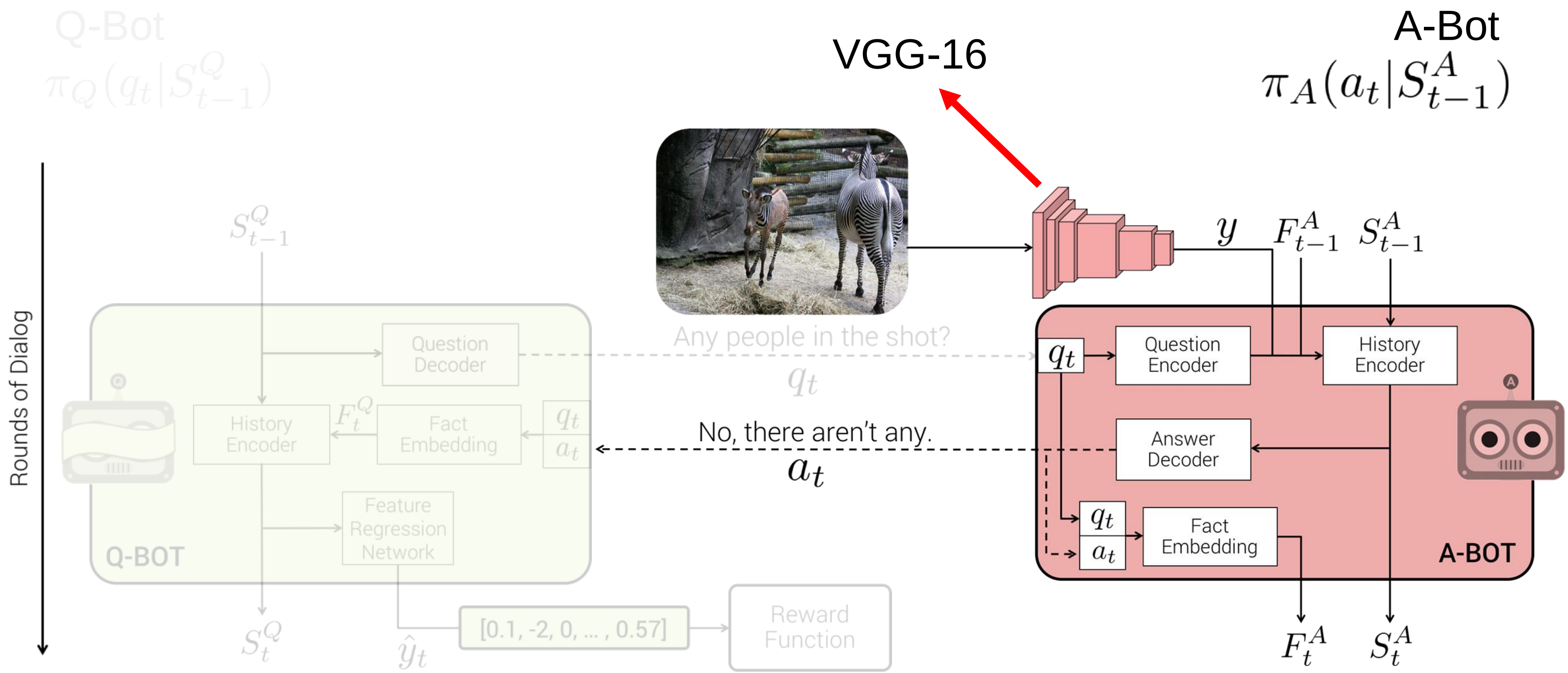
$\pi_A(a_t | S^A_{t-1})$



Policy Networks

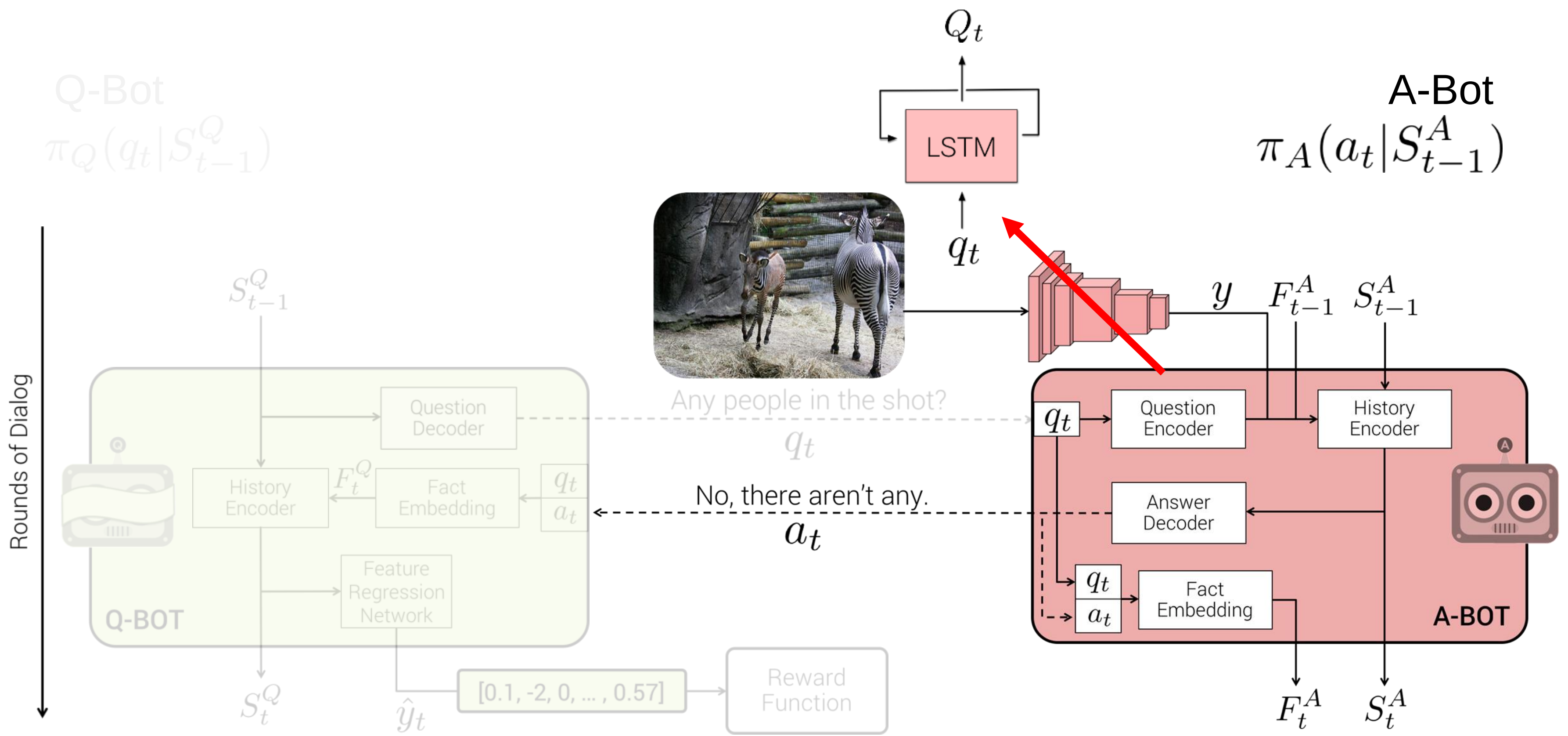


Policy Networks



Slide Credit: Abhishek Das

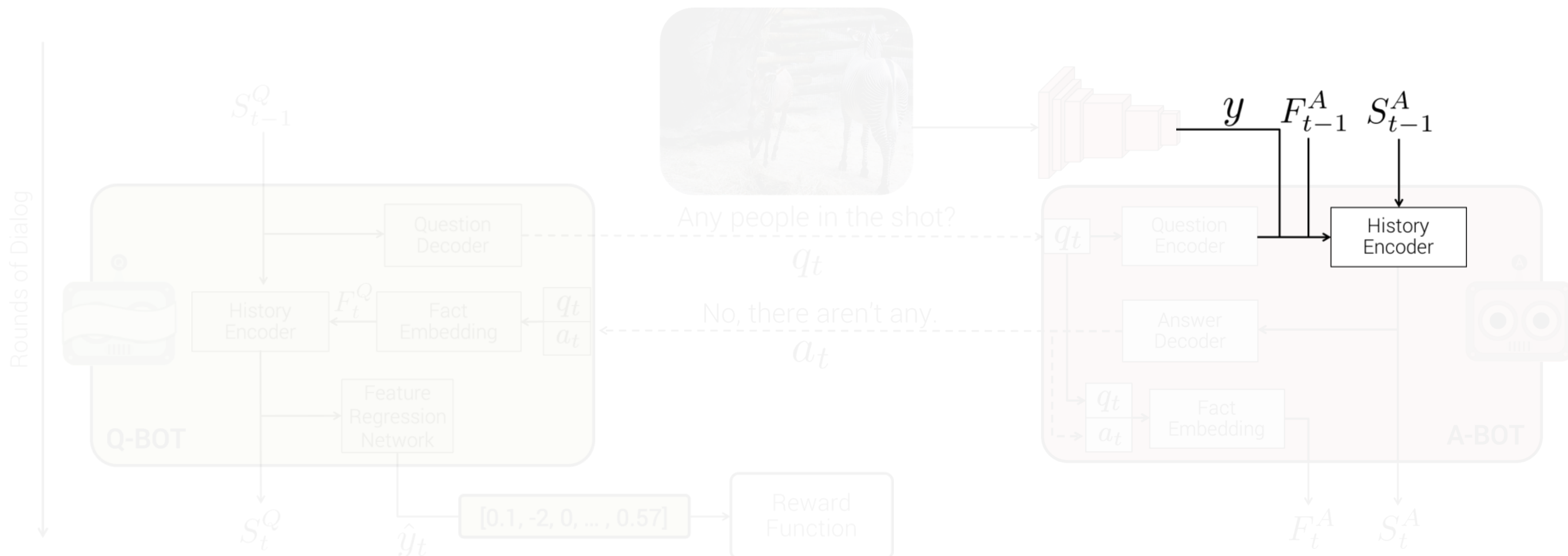
Policy Networks



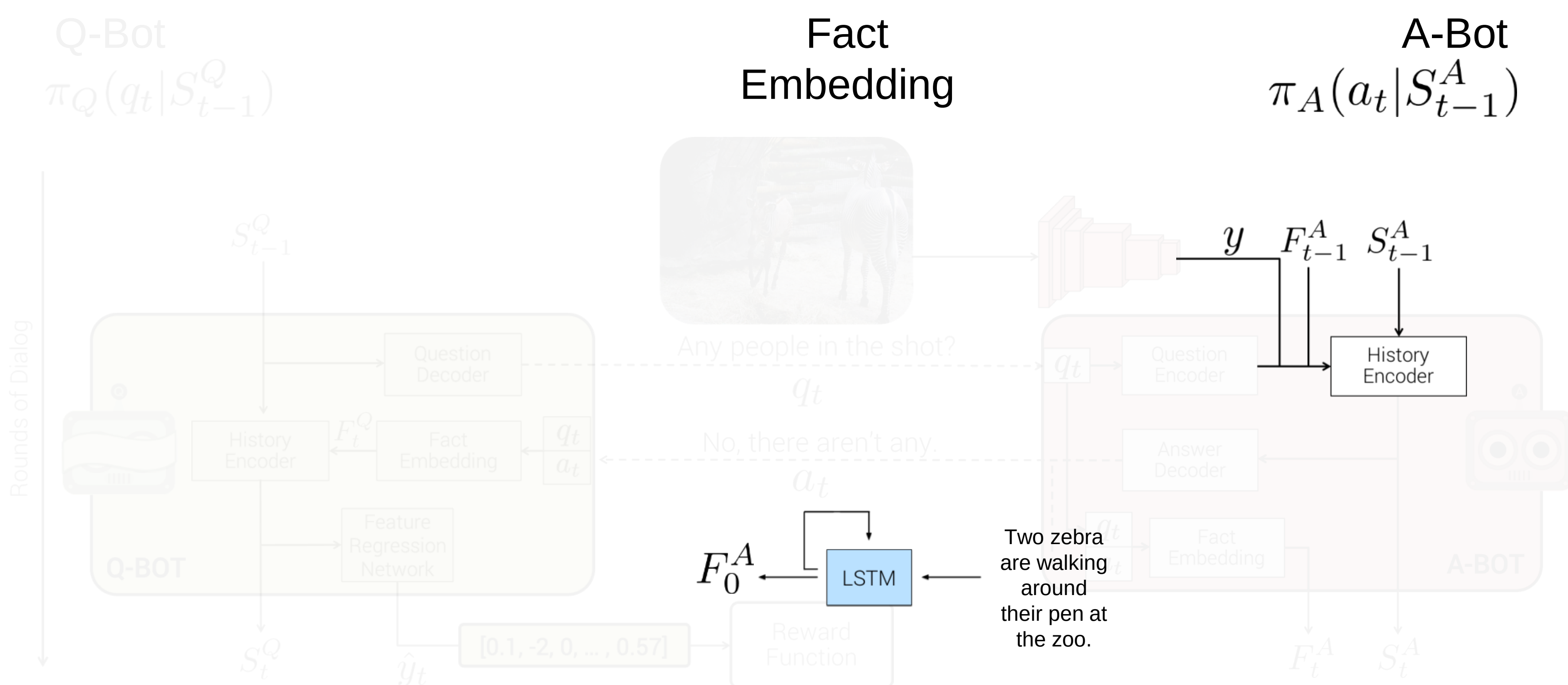
Policy Networks

A-Bot

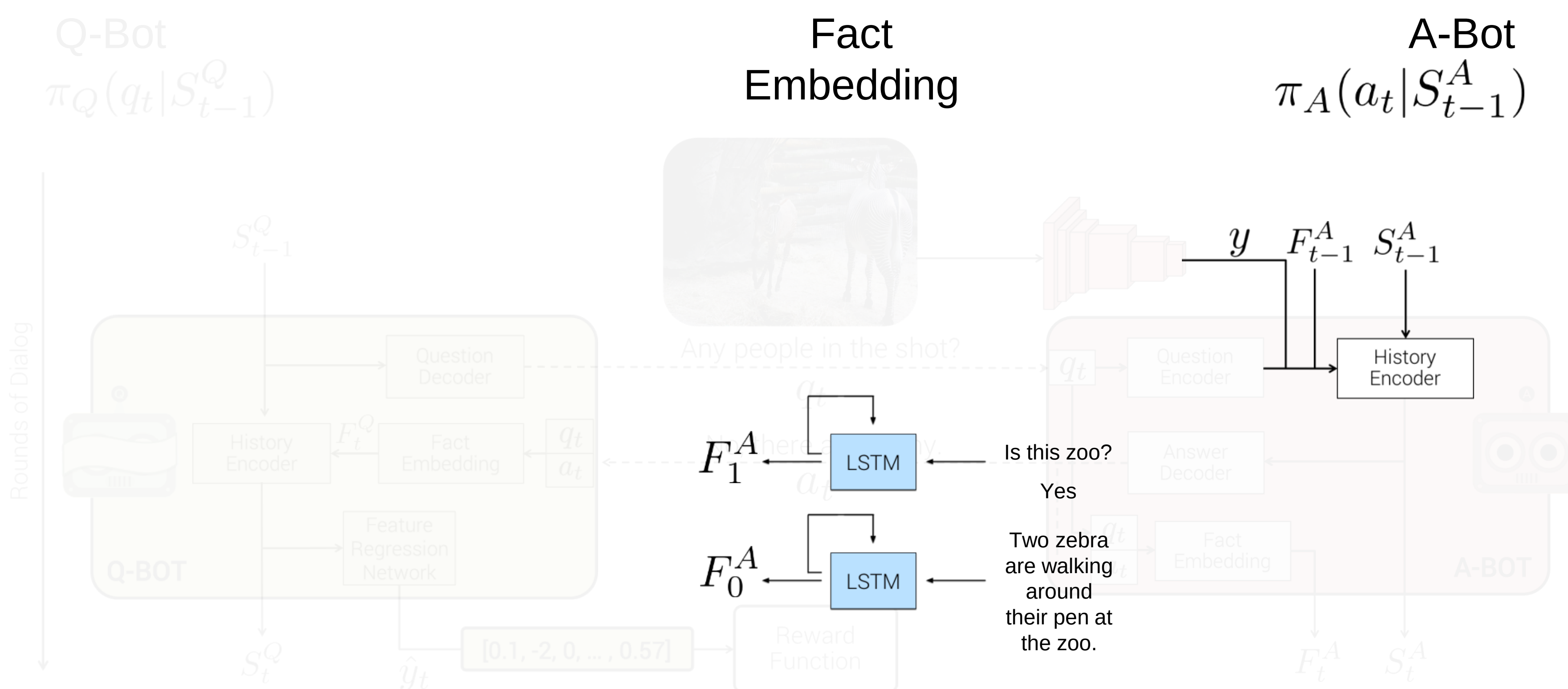
$\pi_A(a_t|S^A_{t-1})$



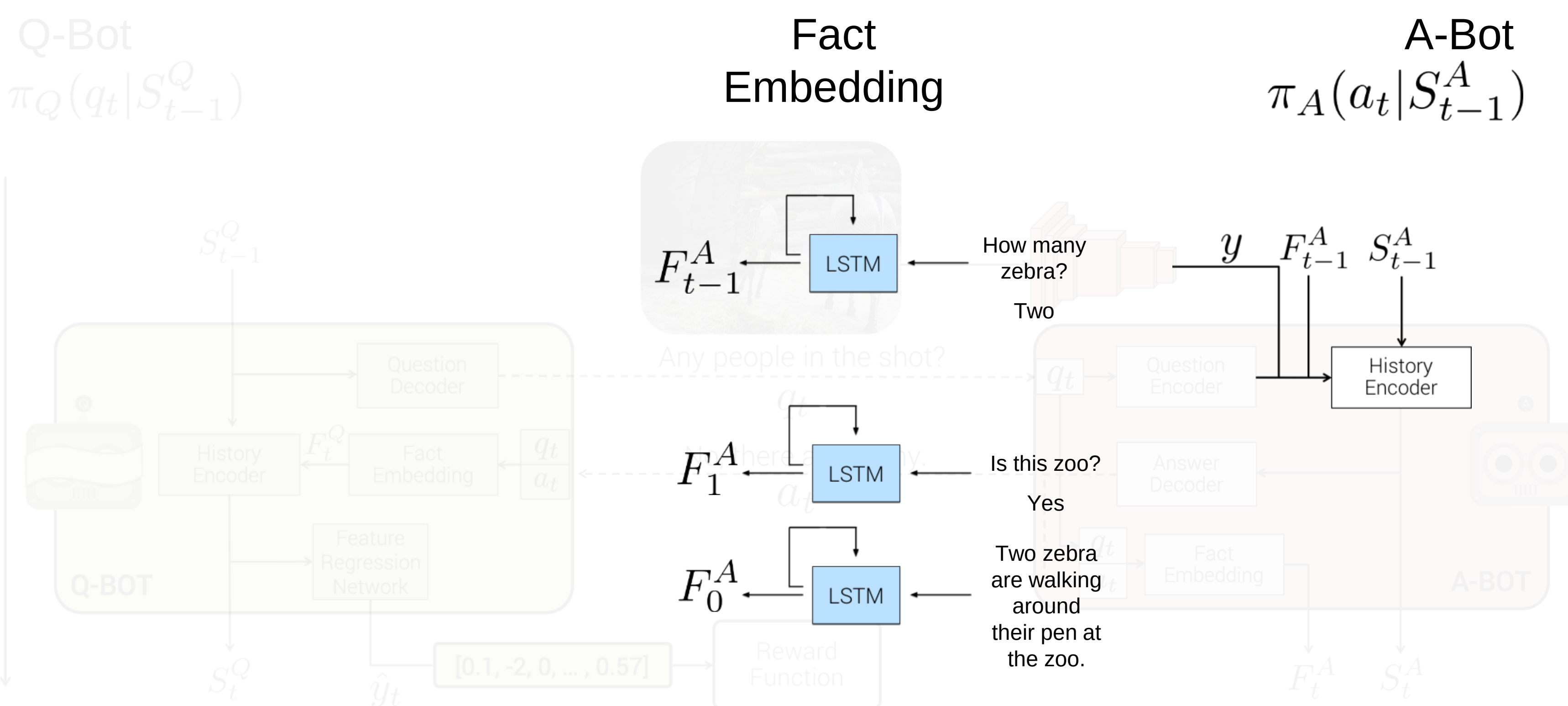
Policy Networks



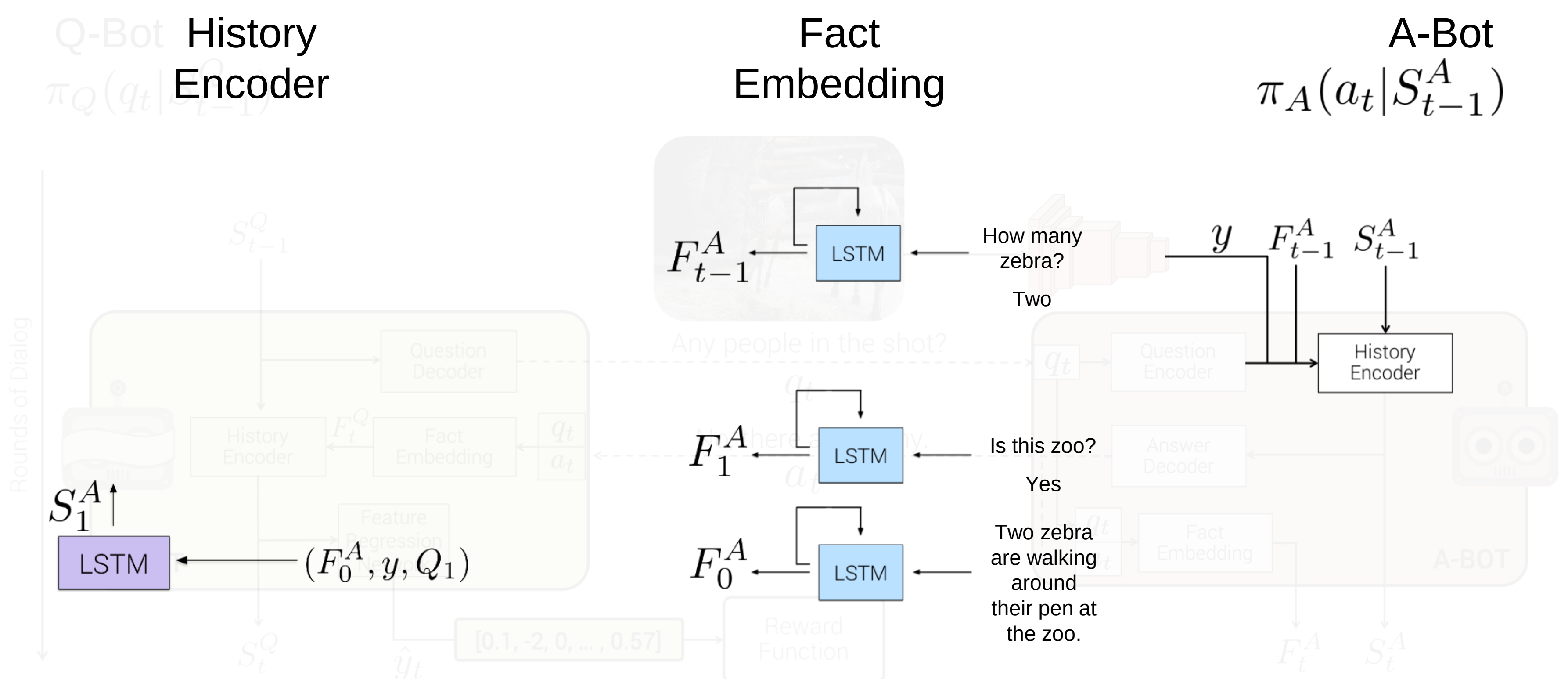
Policy Networks



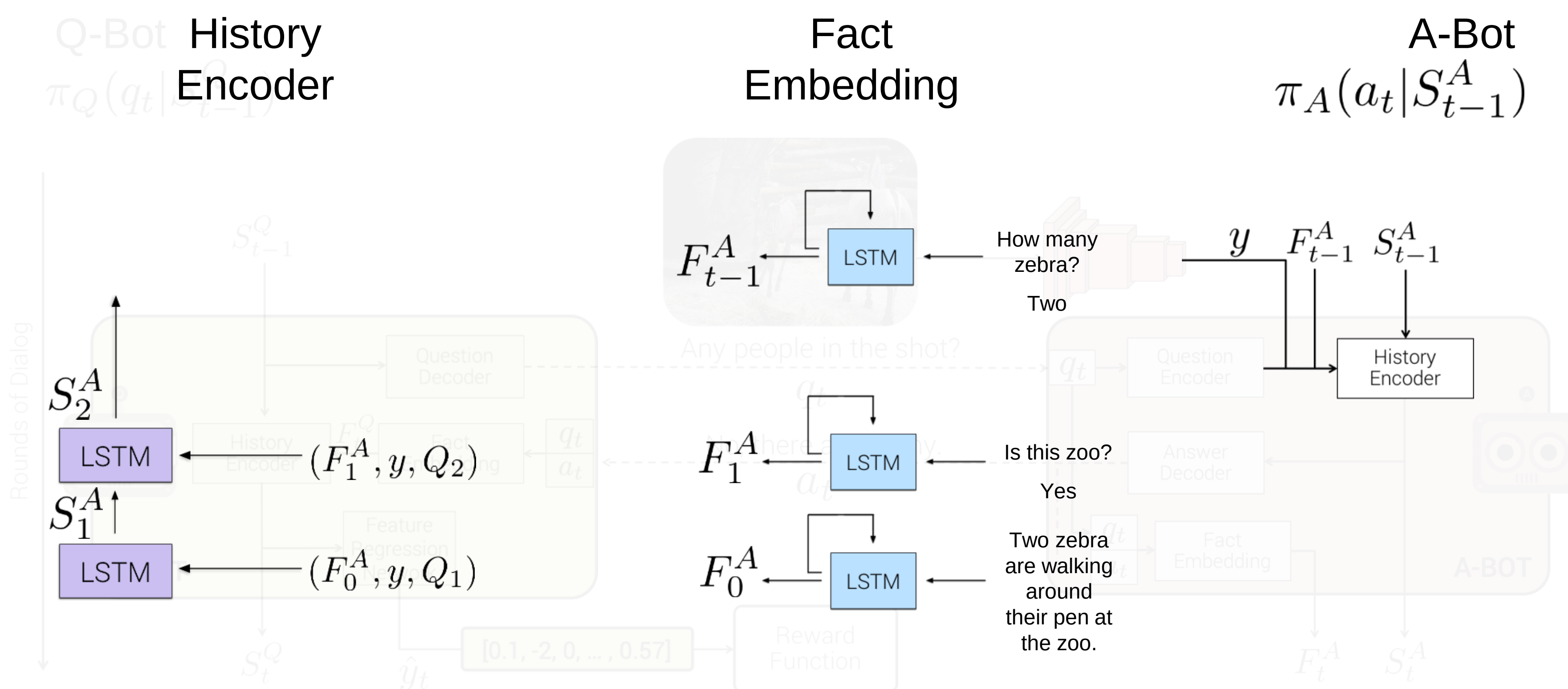
Policy Networks



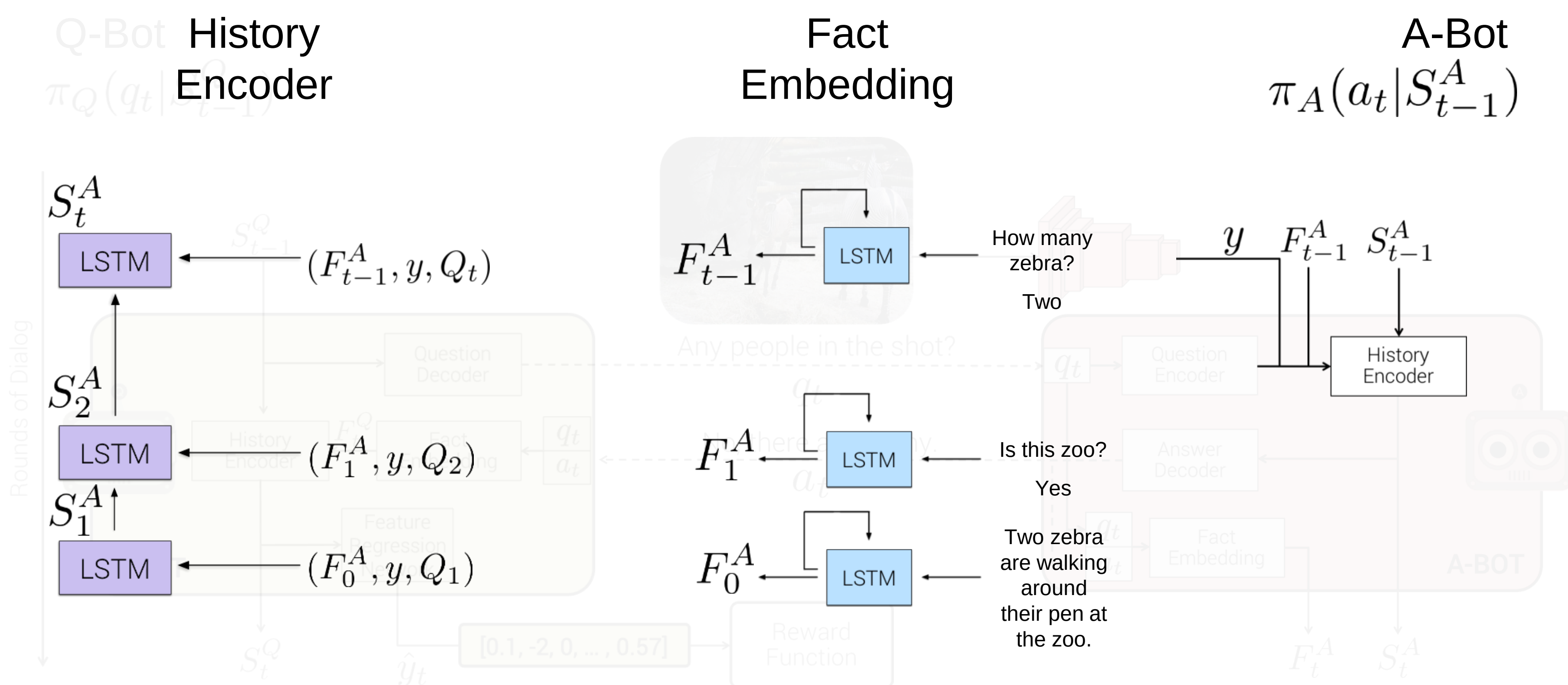
Policy Networks



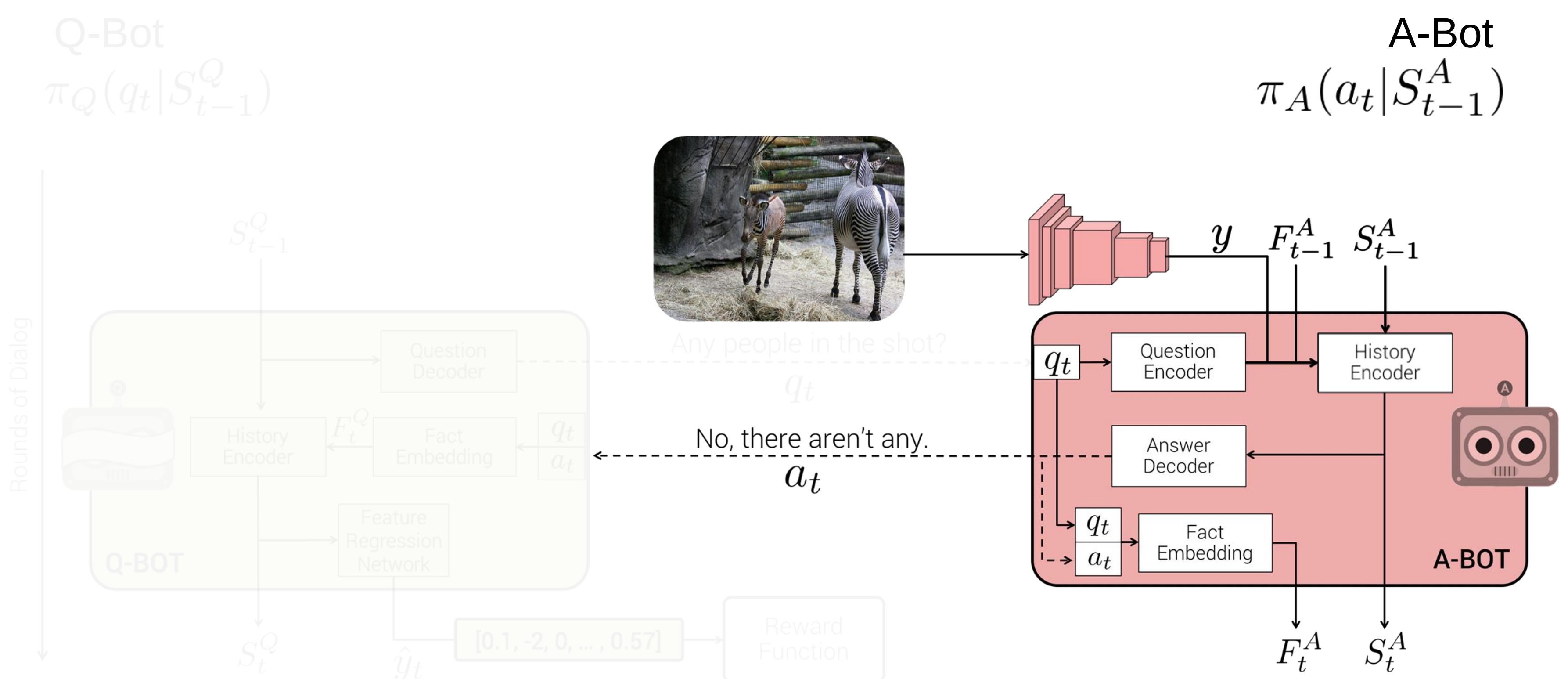
Policy Networks



Policy Networks



Policy Networks



Slide Credit: Abhishek Das

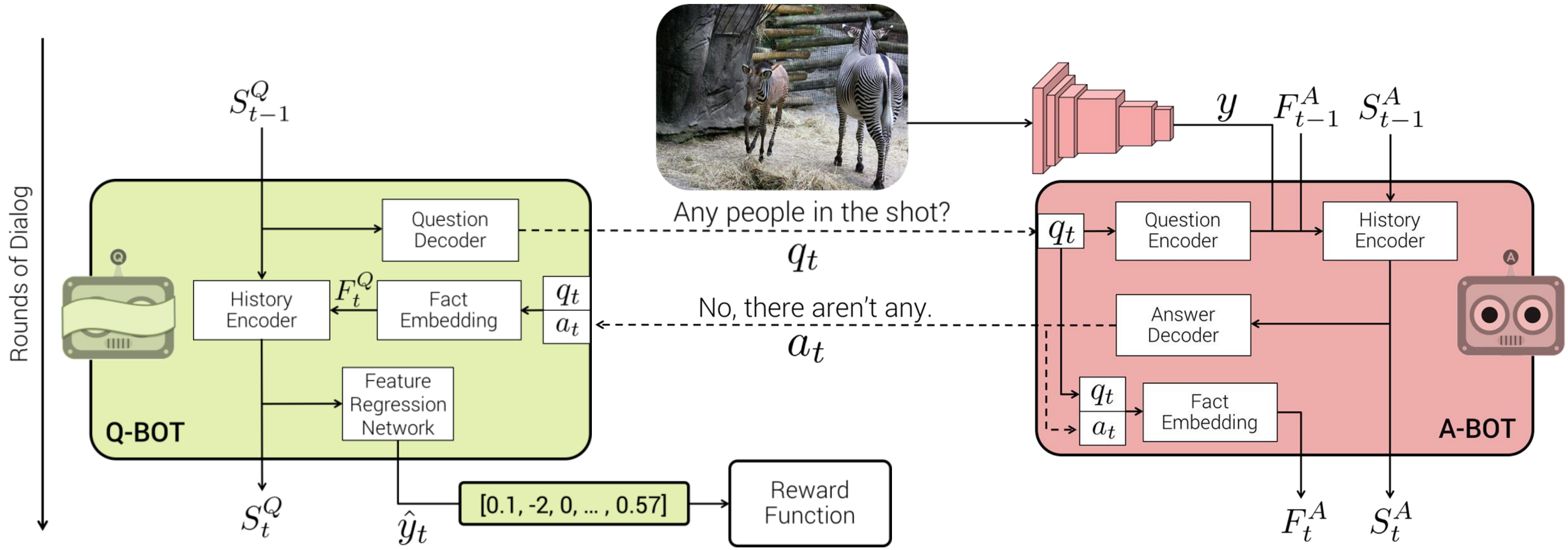
Policy Networks

Q-Bot

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

A-Bot

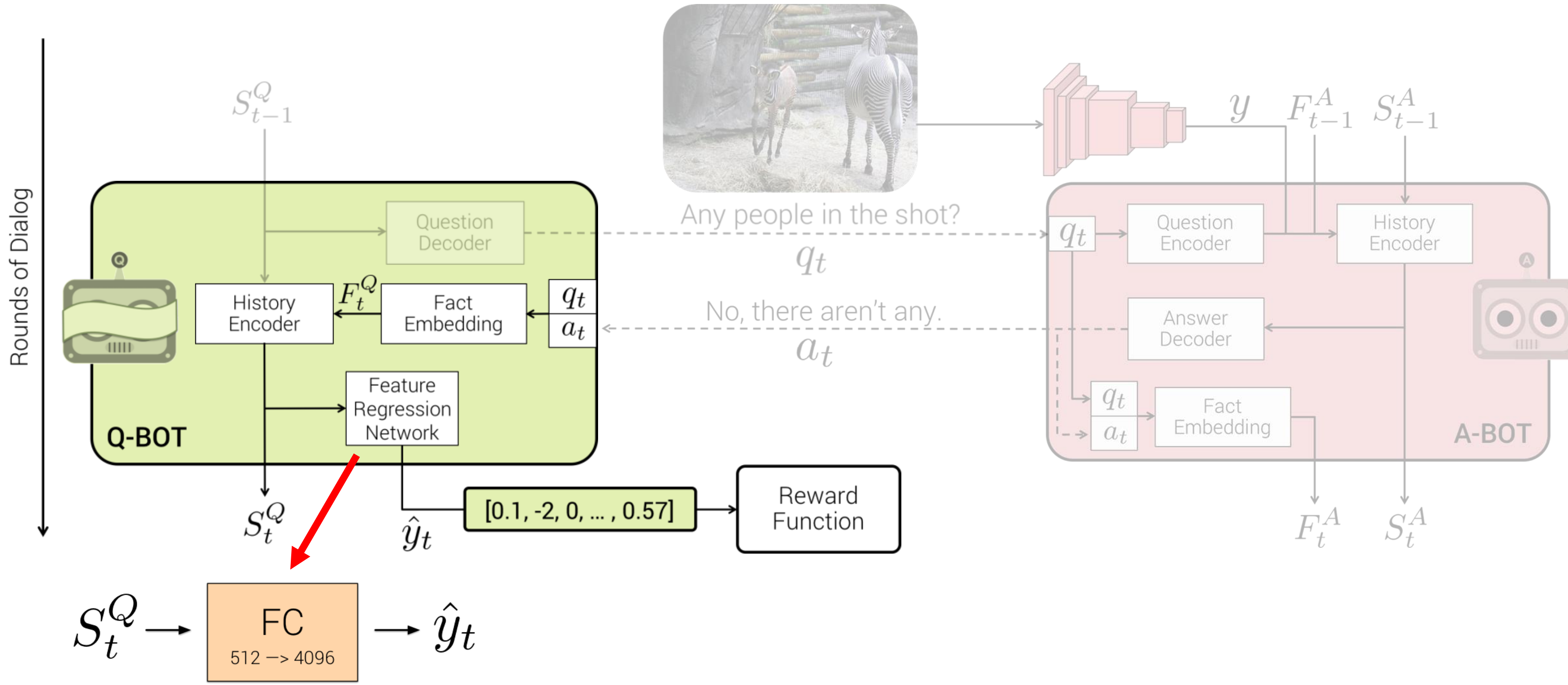
$\pi_A(a_t|S^A_{t-1})$



Policy Networks

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

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



Slide Credit: Abhishek Das

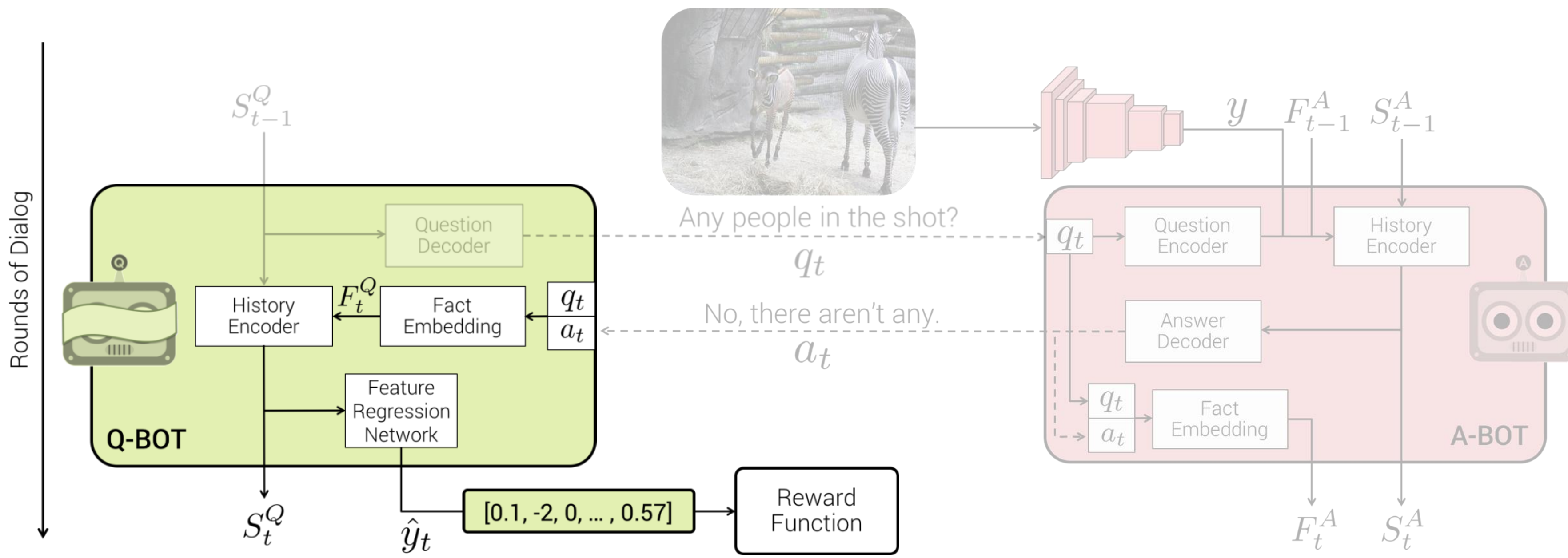
Policy Networks

Q-Bot

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

A-Bot

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



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

Policy Gradients

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

REINFORCE Gradients

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

Turing Test

How old is this man?	He looks like late teens
Does he have short hair or long hair?	Medium
What color is his hair?	Brown
Is he wearing glasses?	No
Does he have facial hair?	No
How is he dressed?	He's white dress shirt with red tie . . . school uniform
Is he larger man or is he skinny?	He's skinny
Is this picture taken outside?	Inside
Can you tell if it's daytime?	I'm not sure
Can you tell what room he is standing in?	He's in hallway

A man making the live long and prosper sign from star trek.

Where is man located?	Looks like classroom of some sort
What is man wearing?	Black t shirt and jeans
Are there any other people in picture?	Yes, quite few people in background
What color is table?	Picture is black and white, but it's wood table
Is there anything else on table?	Not that I can see
What are other people doing?	They are standing in front of him, but I don't . . .
Are there any windows?	Not that I can see
What color are walls?	I can't see walls
What time of day do you think it is?	I can't tell
What is man wearing?	Black t shirt and jeans

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.

Human-Human Q	A	SL-pretrained Q	A	RL-full-QAf Q	A
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

Two men giving thumbs up holding large slices of pizza.

Human-Human Q	A	SL-pretrained Q	A	RL-full-QAf Q	A
How old are men?	They look to be 30's	How old are men?	They look like they are in their 20's	How old are men?	They look like they are in their 20's
Are they in restaurant?	I can't tell, its close up	Are they male or female?	1 male and 1 female	Are they wearing chef's hat?	No
Do men look they might be related?	Maybe	What are they wearing?	T shirt and shorts	Do they have long hair?	No
Are they each holding slice of pizza in 1 hand and giving thumbs up with other?	1 holding slice, other holding box...	What kind of pizza is it?	Looks like pepperoni	What kind of pizza is it?	Looks like pepperoni
What kind of pizza do they have?	cheese	Does it look good?	Yes it does	Does it look good?	Yes
Does box have pizza company name on it?	No	Can you tell what kind of pizza it is?	Looks like pepperoni	Are there any other people around?	No
Are there any pictures or graphics on box?	No	Are there any toppings on pizza?	Pepperoni	Are they inside or outside?	Inside
Are men slender or heavy set?	Slender	Are there any toppings on pizza?	Pepperoni	Can you tell what time of day it is?	No
Do either of men wear glasses?	Yes	Are there any toppings on pizza?	Looks like cheese, but I can't tell...	Are there any other people around?	No
1 or both?	1	Are there any other toppings on pizza?	No	What color are walls?	White

A man riding a skateboard down the side of a ramp.

Human-Human Q	A	SL-pretrained Q	A	RL-full-QAf Q	A
How old is man?	In his teens	How old is man?	I can't see his face, I don't know	What color is skateboard?	It is hard to tell but I think it's black
Is he wearing helmet?	No	What is he wearing?	shorts and t shirt	Is he wearing helmet?	Yes
Is he wearing any protective gear?	No	What color is his shirt?	white	Is he wearing knee pads?	No, no knee pads
What color is his hair?	Dirty blonde or dark brown	what color are his pants?	black	Is it day or night?	It is daytime
What color is his shirt?	Black with red letters	what color is his shirt?	blue	Is it sunny?	It is bright out, but I see no shadows, so I can not
Is he wearing shorts?	Yes	What color are his shoes?	Black	Can you see any other people?	There are 2 other people in background
What color are his shorts?	Denim	What color are his shoes?	Black	Is it day or night?	It is daytime
Are there other people visible?	No	What color are his shoes?	Black	Is it sunny?	Yes, it is sunny
What color are his shoes?	I can't see them	What color are his shoes?	Black	What color is skateboard?	It is hard to tell but I think it's black
Is this at skate park?	Yes	What color are his shoes?	Black	Is he wearing helmet?	Yes, he is wearing helmet

Human-Human Q	A	SL-pretrained Q	A	RL-full-QAf Q	A
What color is plane?	White and blue	What color is plane?	White	What color is airplane?	White with red and blue stripes
Is there writing on plane?	Yes it says lufthansa	Is it large or small?	Large	Can you tell what airline it is?	No, it's too far away
What color is ground?	Grey tarmac	Any people?	No	What time of day is it?	It is daytime
Any people?	No people visible	Any writing on plane?	Yes	Can you see any people?	No
Can you see airport?	Yes	Any writing on plane?	Yes	Can you see any buildings?	Yes, in background
What color is it?	Grey	Any writing on plane?	Yes	Is it day or night time?	It is daytime
Is there sunshine?	No it's kind of overcast	Any writing on plane?	Yes	Can you see sky?	Yes
Any other planes?	No other planes	What does it say?	UNK	Are there clouds in sky?	Yes, few
Any grass?	No grass	What color is plane?	White	Does it look like it's going to rain?	Yes it does
Any trees?	No tree	Any writing on it?	Yes	Does it look like it's going to rain?	Yes it does

An airplane sitting on an asphalt landing strip

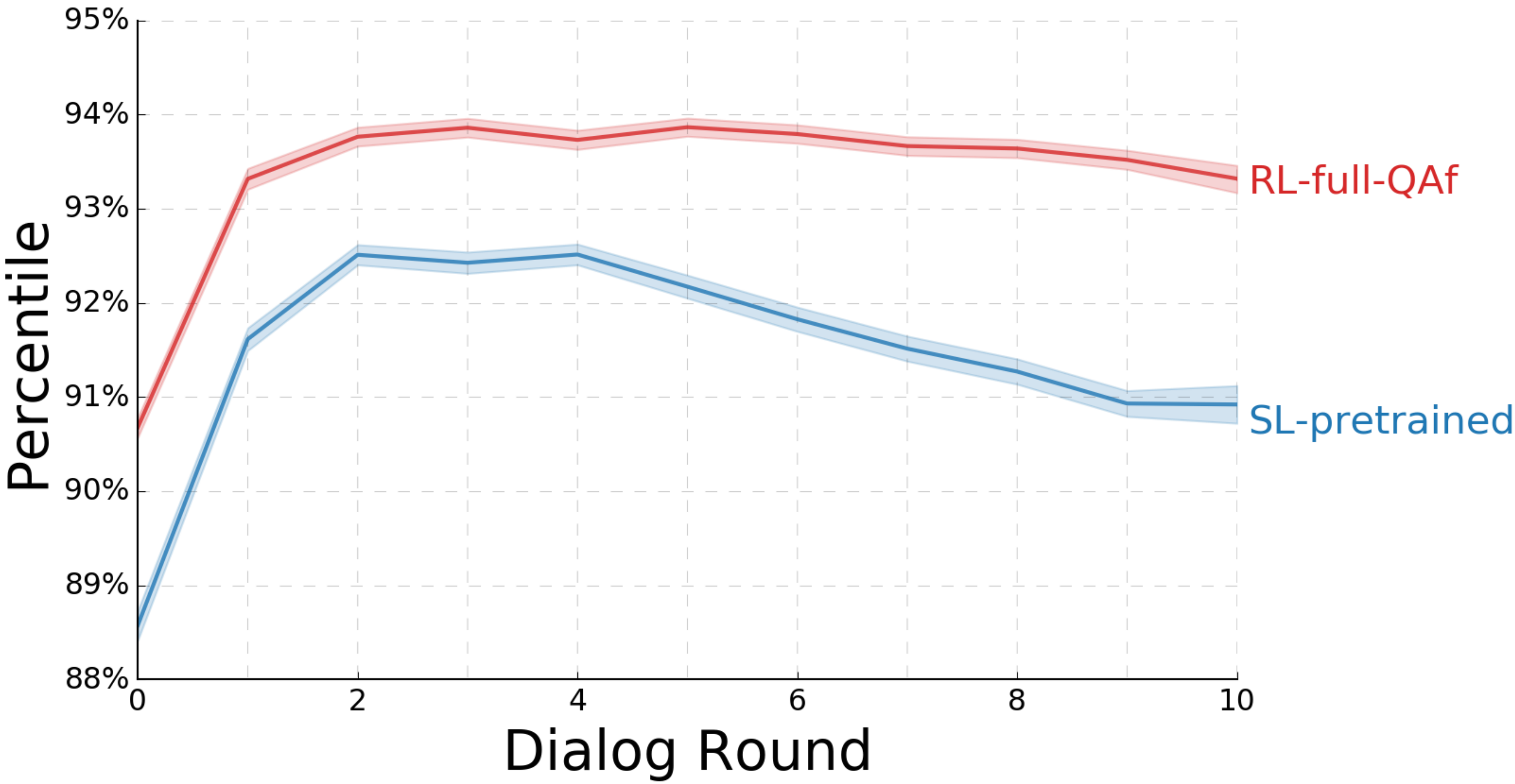
How old is man?	I can't see his face, I don't know
What is he wearing?	shorts and t shirt
What color is his shirt?	white
what color are his pants?	black
what color is his shirt?	blue
What color are his shoes?	Black
What color are his shoes?	Black
What color are his shoes?	Black
What color are his shoes?	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

SL Agents

RL Agents

Image Guessing



Concurrent Work

Is it a person? *No*
Is it an item being worn or held? *Yes*
Is it a snowboard? *Yes*
Is it the red one? *No*
Is it the one being held by the person in blue? *Yes*

Is it a cow? *Yes*
Is it the big cow in the middle? *No*
Is the cow on the left? *No*
On the right ? *Yes*
First cow near us? *Yes*

GuessWhat?! Visual object discovery through multi-modal dialogue

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, Aaron Courville

(Submitted on 23 Nov 2016 (v1), last revised 6 Feb 2017 (this version, v2))

We introduce GuessWhat?!, a two-player guessing game as a testbed for research on the
the game is to locate an unknown object in a rich image scene by asking a sequence of qu
reasoning and language grounding, is required to solve the proposed task. Our key contri
150K human-played games with a total of 800K visual question-answer pairs on 66K ima
and introduce the oracle and questioner tasks that are associated with the two players of
initial baselines of the introduced tasks.

End-to-end optimization of goal-driven and visually grounded dialogue systems

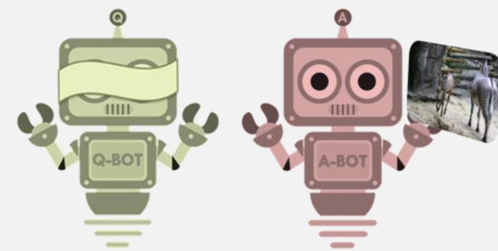
Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, Olivier Pietquin

(Submitted on 15 Mar 2017)

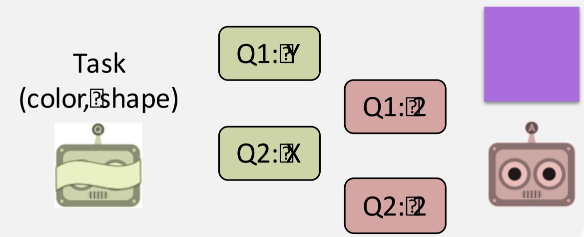
End-to-end design of dialogue systems has recently become a popular research topic thanks to powerful tools such as encoder-decoder architectures for sequence-to-sequence learning. Yet, most current approaches cast human-machine dialogue management as a supervised learning problem, aiming at predicting the next utterance of a participant given the full history of the dialogue. This vision is too simplistic to render the intrinsic planning problem inherent to dialogue as well as its grounded nature, making the context of a dialogue larger than the sole history. This is why only chit-chat and question answering tasks have been addressed so far using end-to-end architectures. In this paper, we introduce a Deep Reinforcement Learning method to optimize visually grounded task-oriented dialogues, based on the policy gradient algorithm. This approach is tested on a dataset of 120k dialogues collected through Mechanical Turk and provides encouraging results at solving both the problem of generating natural dialogues and the task of discovering a specific object in a complex picture.

Outline

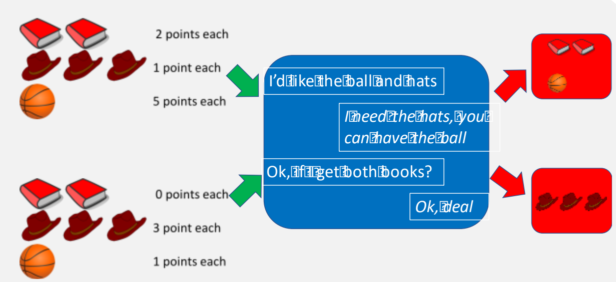
Cooperative Visual Dialog Agents



Emergence of Grounded Dialog



Negotiation Dialog Agents



Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog
[EMNLP ‘17]


Satwik Kottur*
(CMU)

José Moura
(CMU)

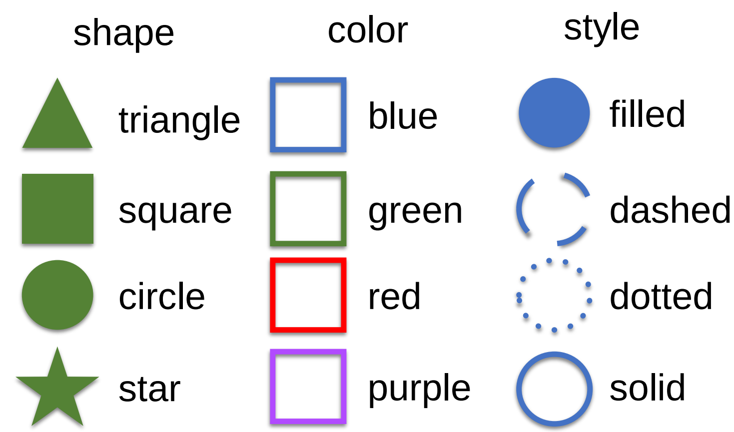

Stefan Lee
(Virginia Tech)

Dhruv Batra
(Georgia Tech)

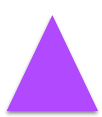
Toy World

- Sanity check

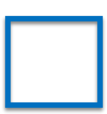
- Simple, synthetic world
  - Instances - (shape, color, style)
  - Total of $4^3(64)$ instances



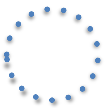
  - Example instances:



(triangle, purple, filled)



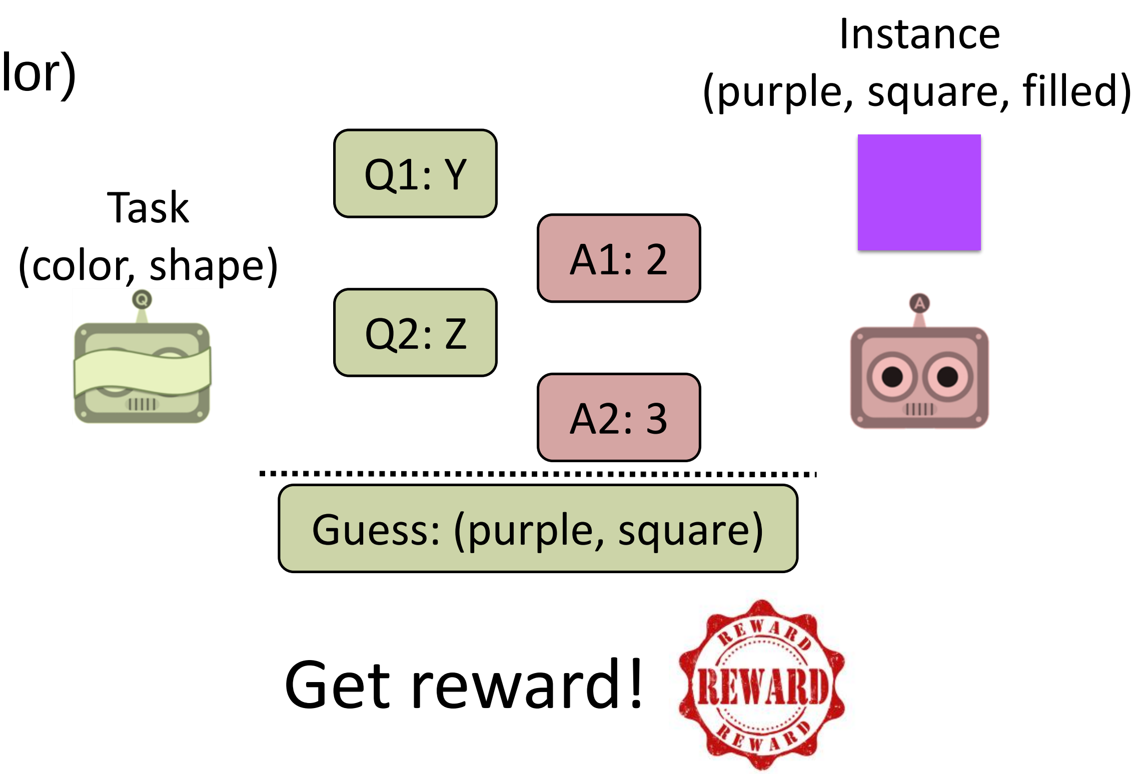
(square, blue, solid)



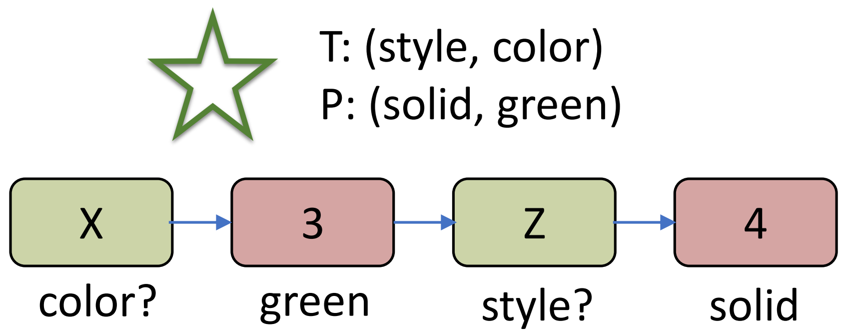
(circle, blue, dotted)

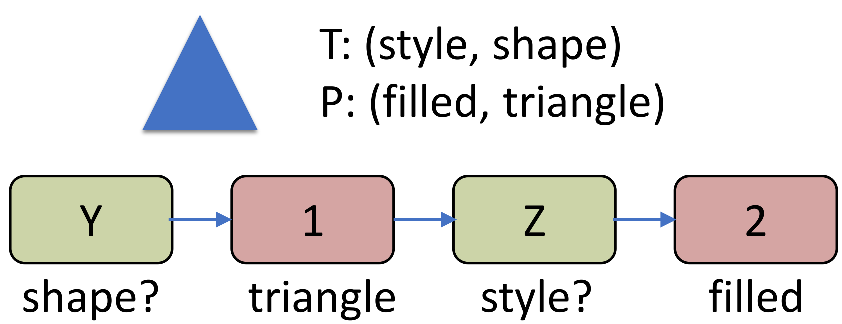
Task & Talk

- Task (G)
  - Inquire pair of attributes
  - (color, shape), (shape, color)
- Talk
  - Single token per round
  - Two rounds
- Q-bot guesses a pair
  - Reward : +1 / -1
  - Prediction order matters!



Emergence of Grounded Dialog





Emergence of Grounded Dialog

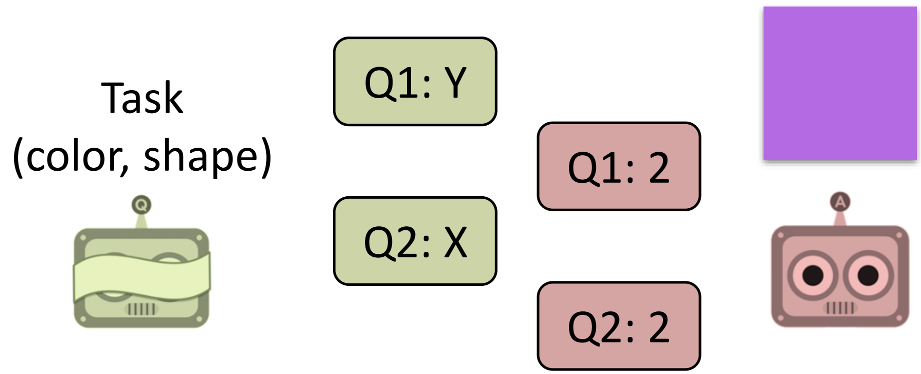
- Compositional grounding
- Predict dialog for unseen instances

	Attributes		
	color	*shape*	*style*
V_A	*X*	*Y*	*Z*
1	*blue*	*triangle*	*dotted*
2	*purple*	*square*	*filled*
3	*green*	*circle*	*dashed*
4	*red*	*start*	*solid*

(a) A-BOT

Task	q_1, q_2
(color, shape) *(shape, color)*	*Y, X*
(shape, style) *(style, shape)*	*Y, Z*
(color, style)	*Z, X*
(style, color)	*X, Z*

(b) Q-BOT



Summary of findings

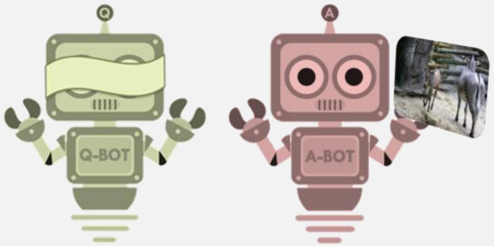
Setting	Vocabulary		Memory		Generalization	Characteristics
	$\lvert V_Q \rvert$	$\lvert V_A \rvert$	Q-bot	A-bot		
A. Over-complete	64	64	Yes	Yes	25.6 %	• Non-compositional language • Q-bot insignificant • Inconsistent A-bot grounding • Poor generalization
B. Attribute	3	12	Yes	Yes	38.5 %	• Non-compositional language • Q-bot uses one round to convey task • Inconsistent A-bot grounding • Poor generalization
C. Minimal	3	4	Yes	No	74.4 %	• Compositional language • Q-bot uses both rounds • Consistent A-bot grounding • Good generalization

Deep Multi-Agent Communication

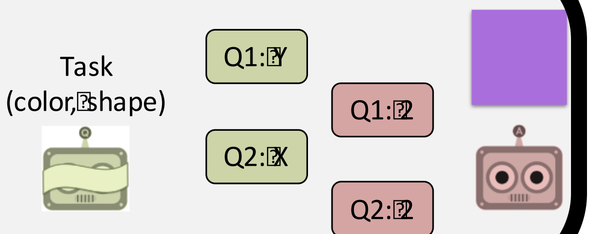
- NIPS ‘16
  - [DeepMind] Learning to Communicate with Deep Multi-Agent Reinforcement Learning. Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, Shimon Whiteson. NIPS ‘16.
  - [NYU / FAIR] Learning Multiagent Communication with Backpropagation. Sainbayar Sukhbaatar, Arthur Szlam, Rob Fergus. NIPS ‘16.
- Arxiv ‘17
  - [OpenAI] Emergence of Grounded Compositional Language in Multi-Agent Populations. Igor Mordatch, Pieter Abbeel.
  - [FAIR] Multi-Agent Cooperation and the Emergence of (Natural) Language. Angeliki Lazaridou, Alexander Peysakhovich, Marco Baroni.
  - Learning to play guess who? and inventing a grounded language as a consequence. Emilio Jorge, Mikael Kågebäck, and Emil Gustavsson.
  - Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. Serhii Havrylov and Ivan Titov.
  - [Berkeley] Translating neuralese. Jacob Andreas, Anca Dragan and Dan Klein. ACL 2017.

Outline

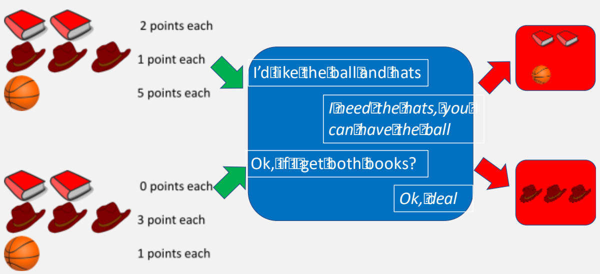
Cooperative Visual Dialog Agents



Emergence of Grounded Dialog



Negotiation Dialog Agents



Deal or No Deal? End-to-End Learning for Negotiation Dialogues
[EMNLP '17]

Mike Lewis
(FAIR)

Denis Yarats
(FAIR)

Yann Dauphin
(FAIR)

Devi Parikh
(Georgia Tech)

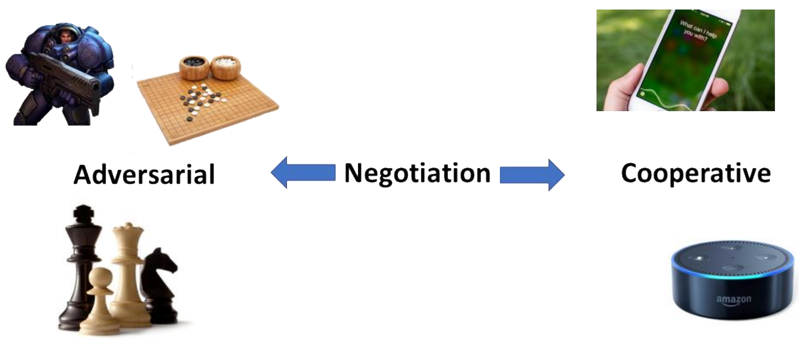
Dhruv Batra
(Georgia Tech)

Why Negotiation?

Adversarial ⟵ **Negotiation** ⟶ **Cooperative**

Slide Credit: Mike Lewis

Why Negotiation?



Negotiation useful when:

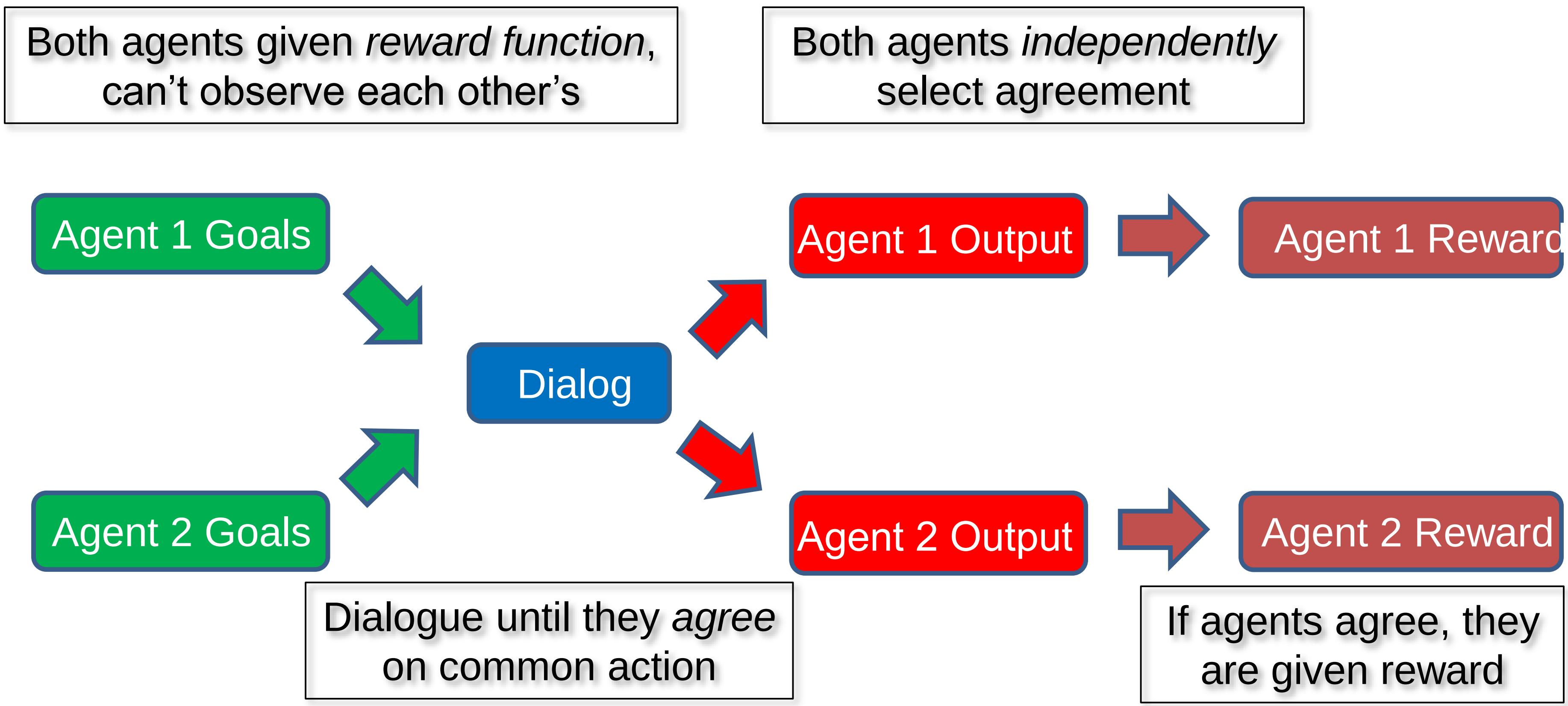
- Agents have different goals
- Not all can be achieved at once
- *(all the time)*

Why Negotiation?

- Both **linguistic** and **reasoning** problem
- *Interpret* multiple sentences, and *generate* new message
- Plan ahead, make proposals, counter-offers, bluffing, lying, compromising

Slide Credit: Mike Lewis

Framework

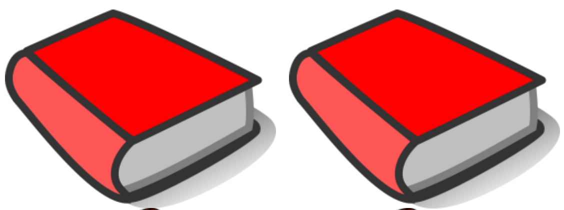


Slide Credit: Mike Lewis

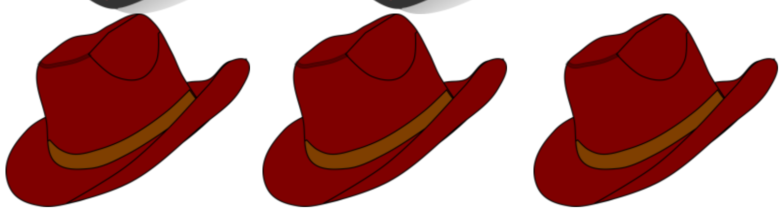
Object Division Task

Agents shown *same* set of object but *different* values for each

Asked to agree how to divide objects between them



2 points each

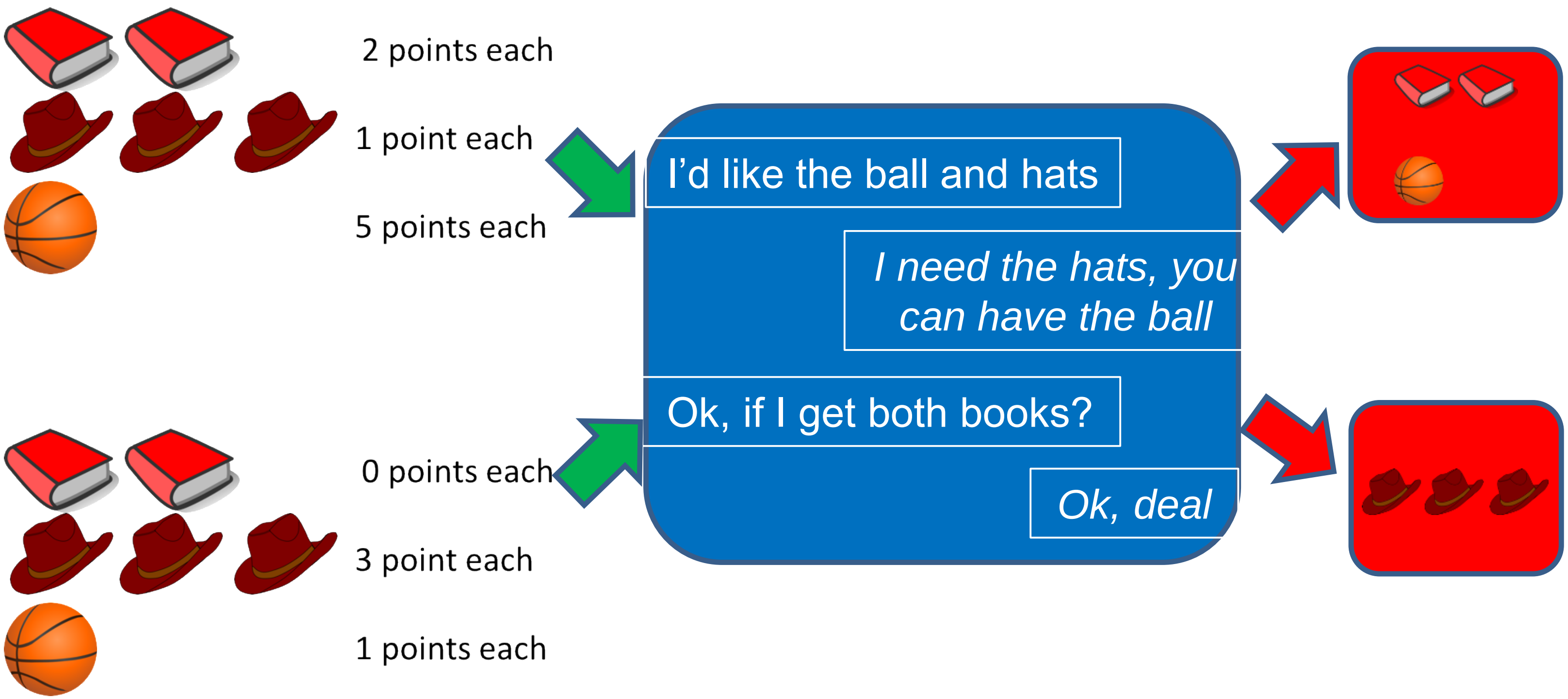


1 point each

5 points each

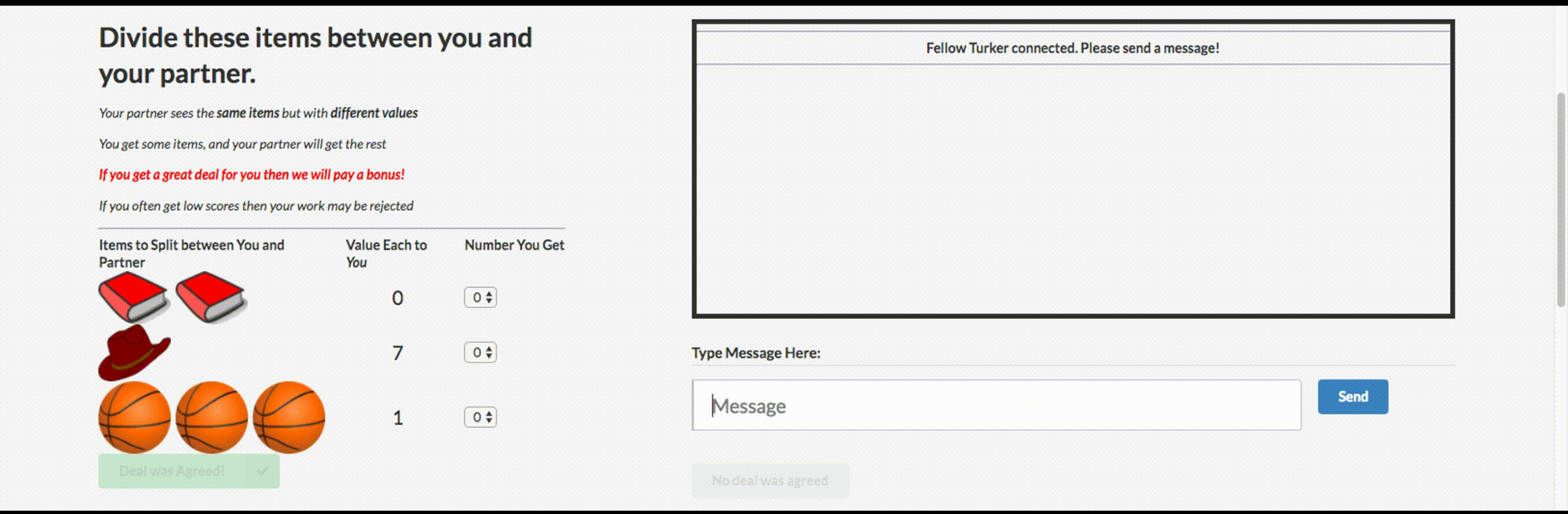
Slide Credit: Mike Lewis

Multi-Issue Bargaining



Slide Credit: Mike Lewis

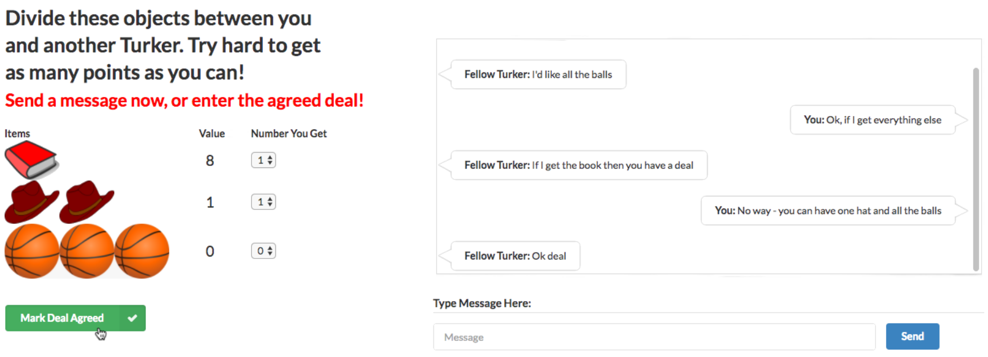
Data Collection on AMT



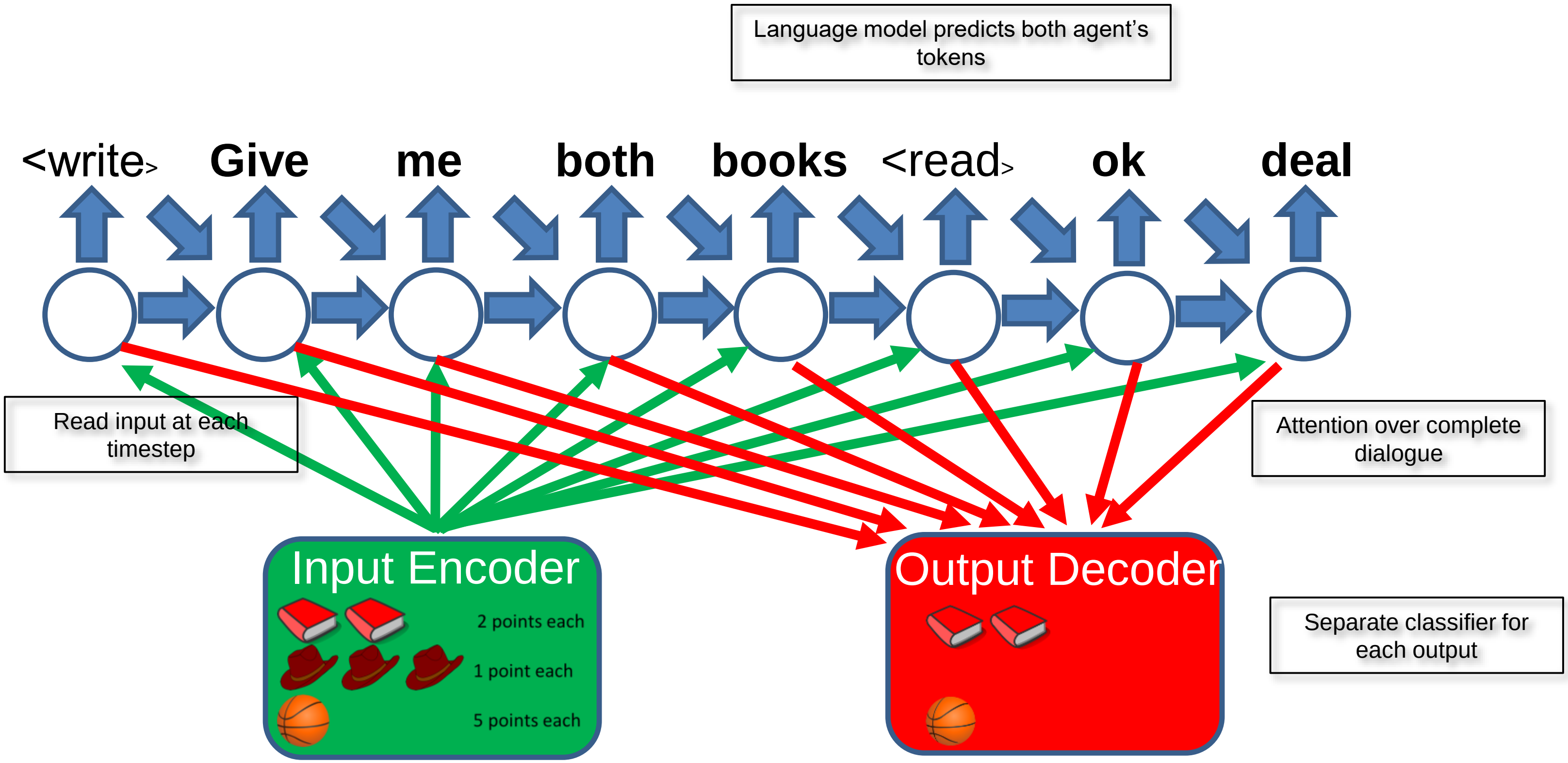
Slide Credit: Mike Lewis

Dataset

- ~6k **dialogs**
- Average 6.6 **turns/dialog**
- Average 7.6 **words/turn**
- 80% **agreed** solutions
- 77% **Pareto Optimal** solutions

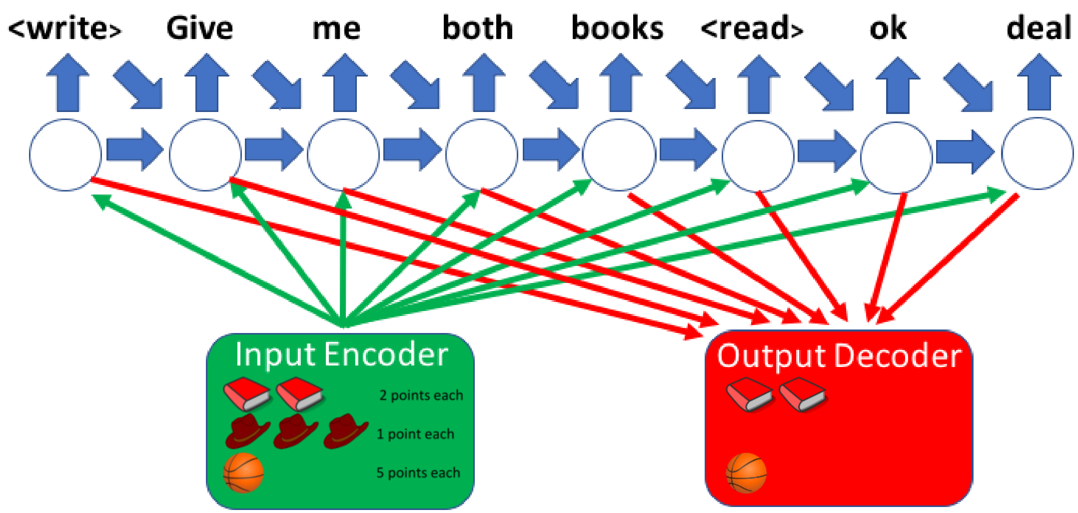


Baseline Model



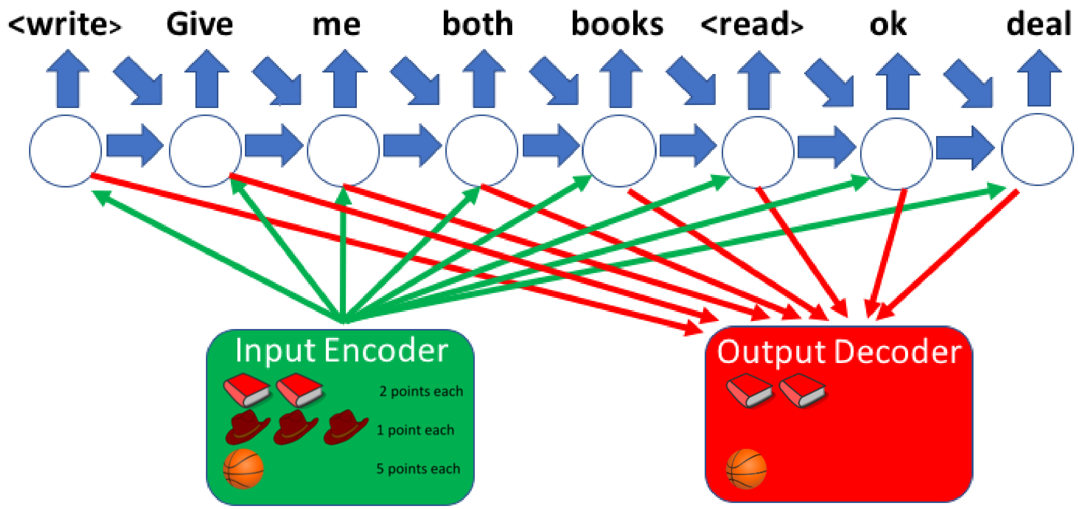
Slide Credit: Mike Lewis

SL-Pretraining



- Train to maximize likelihood of human-human dialogues
- Decode by sampling likely messages

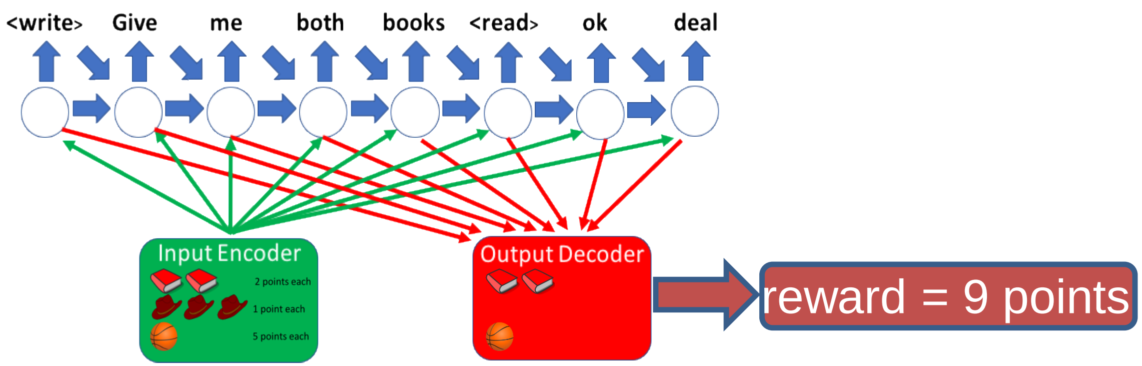
SL-Pretraining



- Model knows nothing about task, just tries to imitate human actions
- Agrees too easily
- Can’t go beyond human strategies

Slide Credit: Mike Lewis

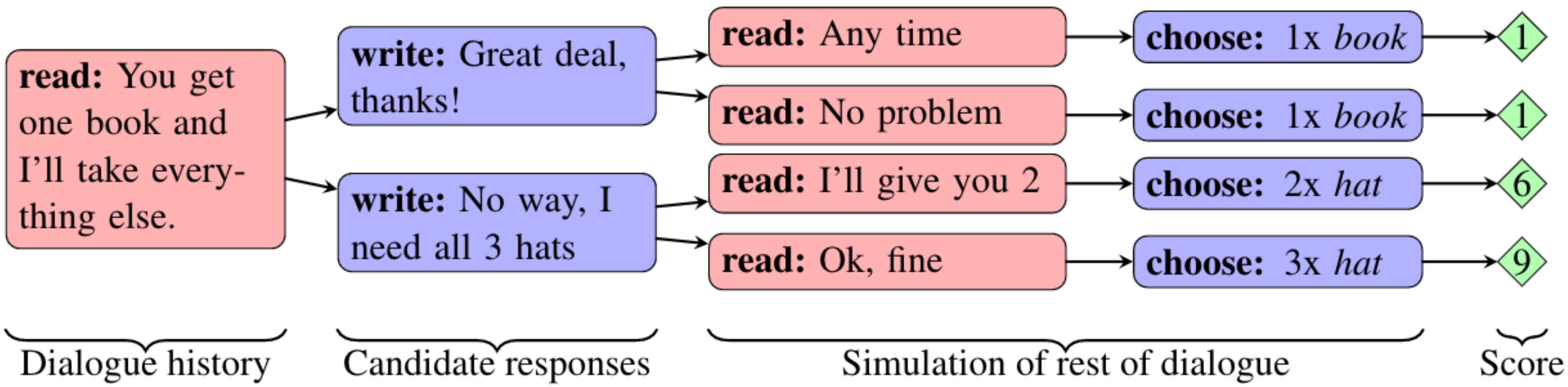
Goal-based RL-Finetuning



- Generate dialogues using self-play
- Backpropagate reward using REINFORCE
- Interleave with supervised updates

- Very sensitive to hyperparameters

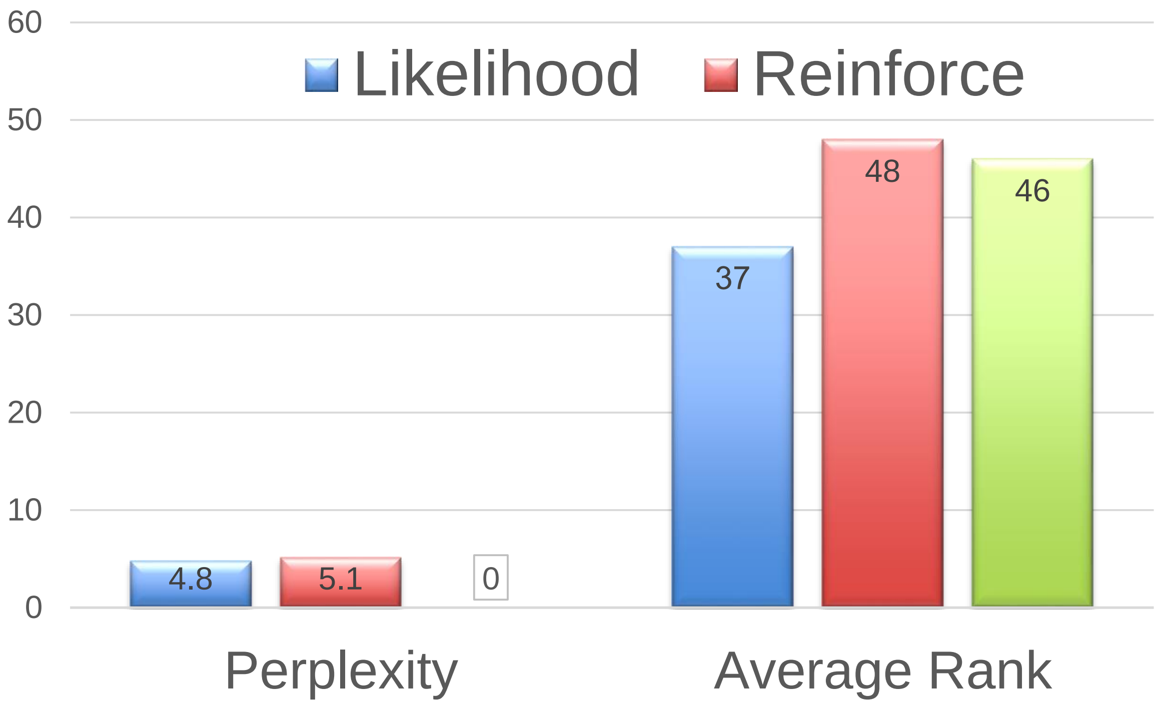
Dialog Rollouts: Goal-based Decoding



- Dialog rollouts use model to simulate remainder of conversation
- Average scores to estimate future reward

Slide Credit: Mike Lewis

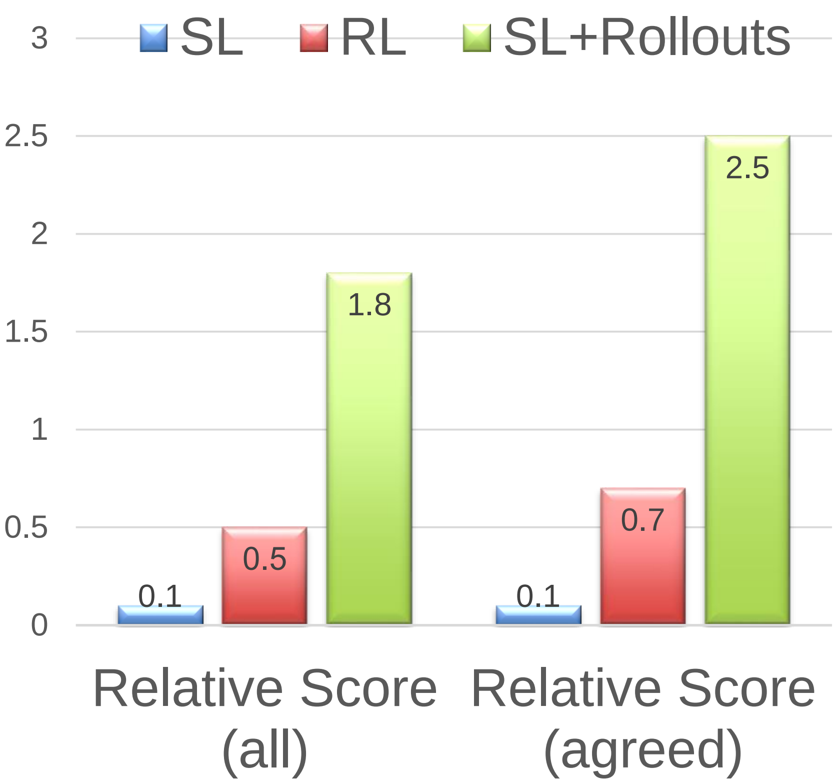
Intrinsic Evaluation

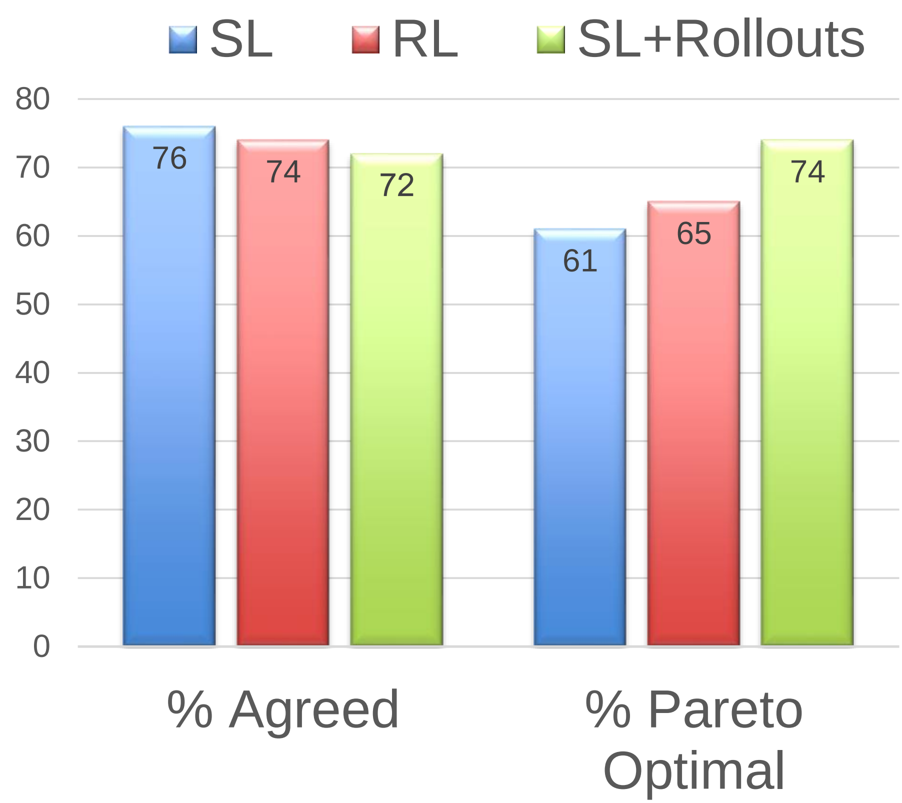


Supervised learning gives most "human like" dialog

Slide Credit: Mike Lewis

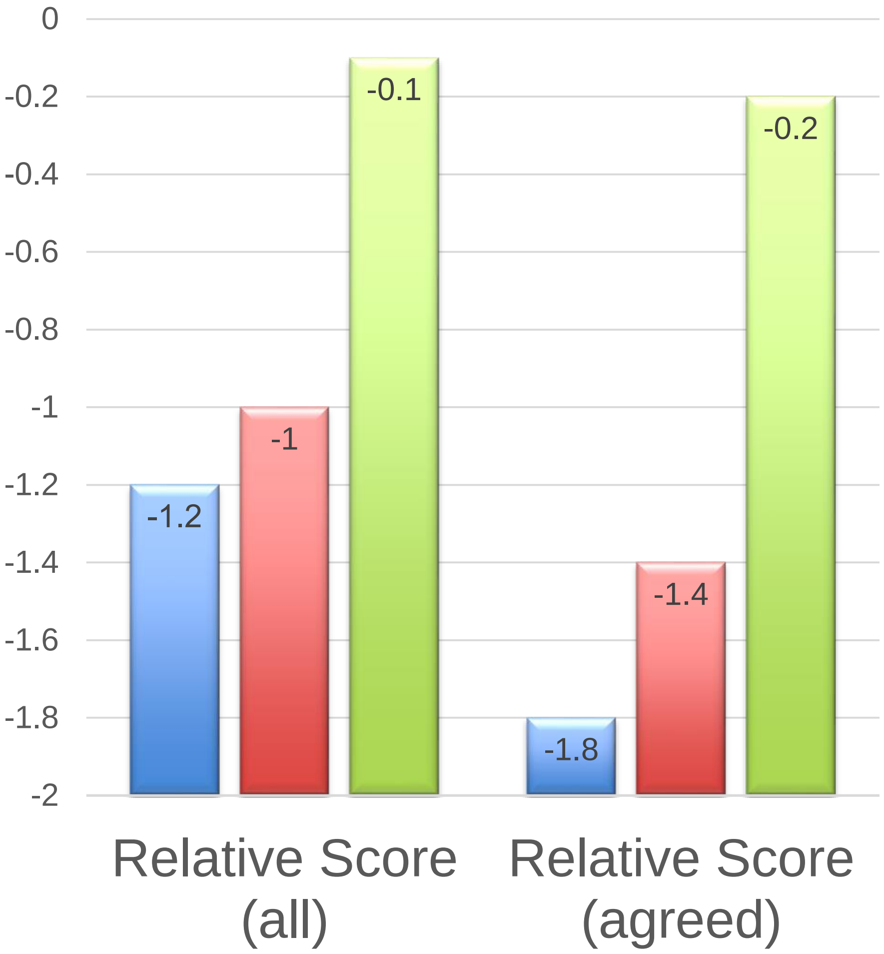
End-to-End Evaluation against SL negotiators

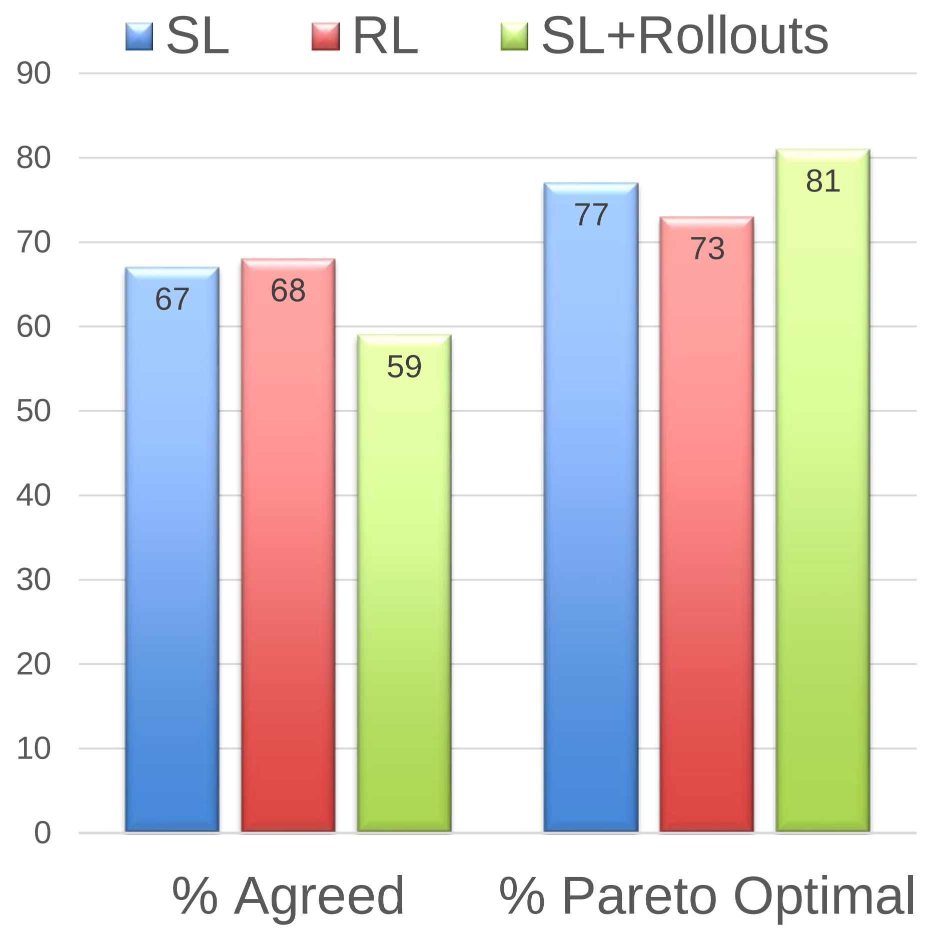




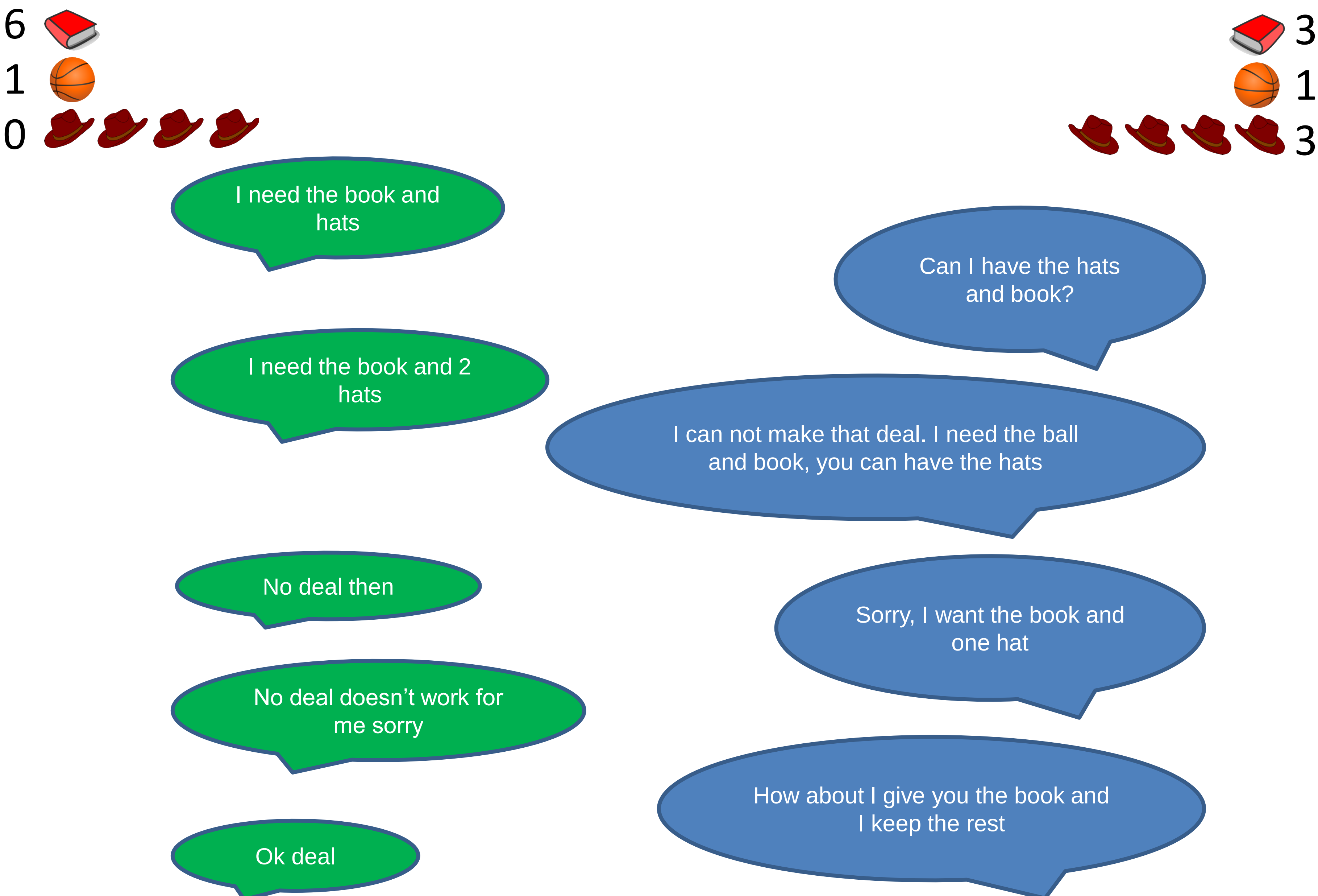
Slide Credit: Mike Lewis

End-to-End Evaluation against Turkers

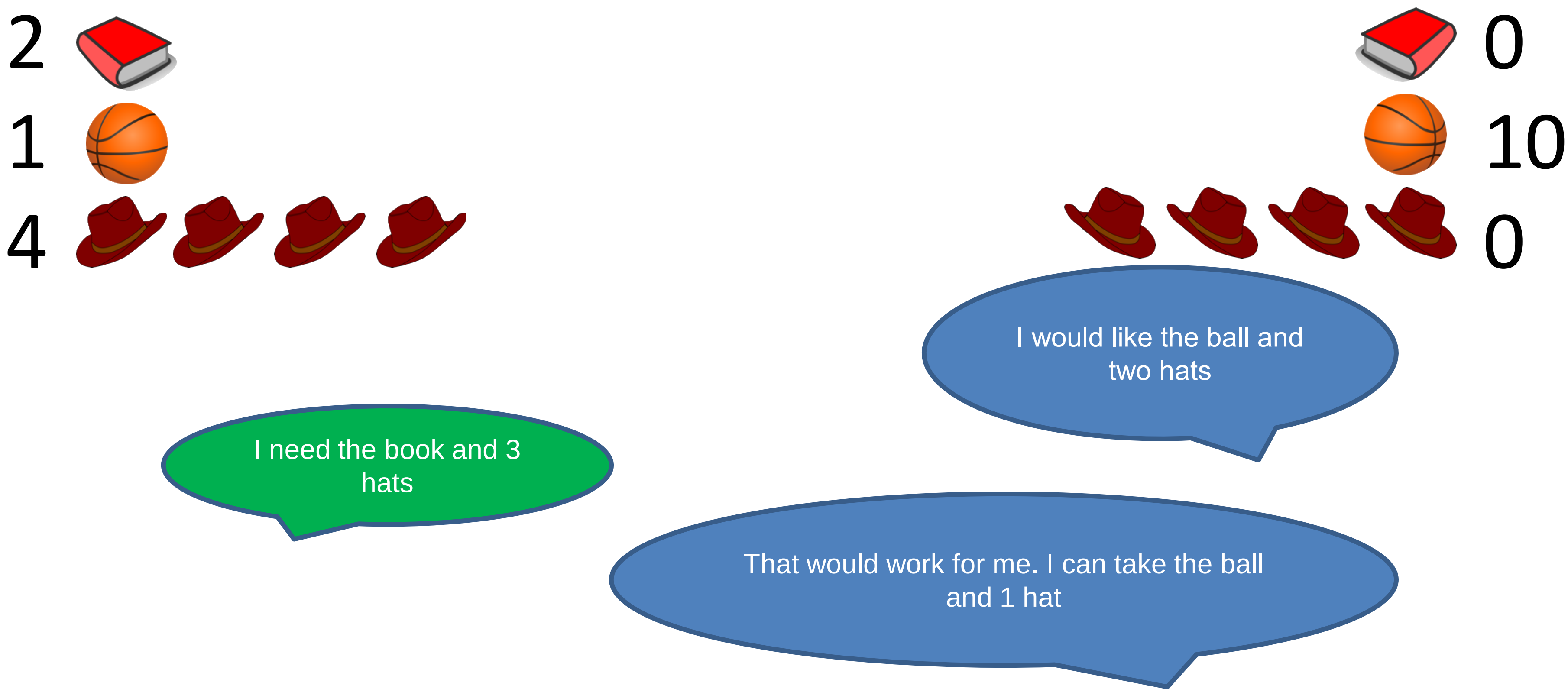




Slide Credit: Mike Lewis



Model generates meaningful novel language



Model can be deceptive to achieve its goals

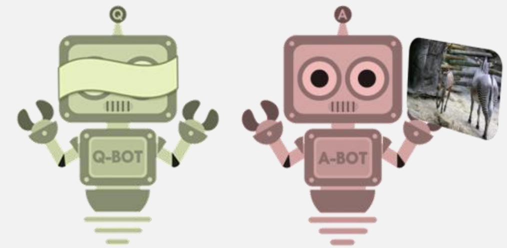
Conclusion

- Negotiation is **useful** and **challenging**
- End-to-End approach trades cheaper data for difficult modelling
- Goal-based training and decoding improves over likelihood
- Model can generate meaningful language be be deceptive to achieve their goals

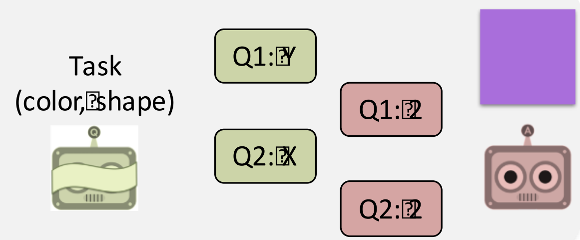
Slide Credit: Mike Lewis

Outline

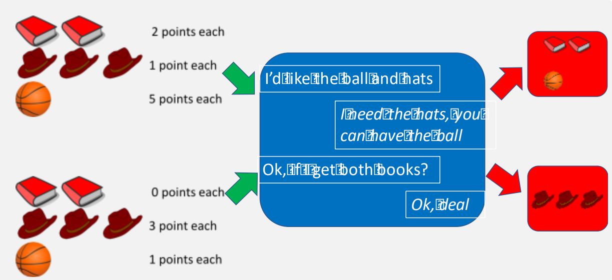
Cooperative Visual Dialog Agents



Emergence of Grounded Dialog



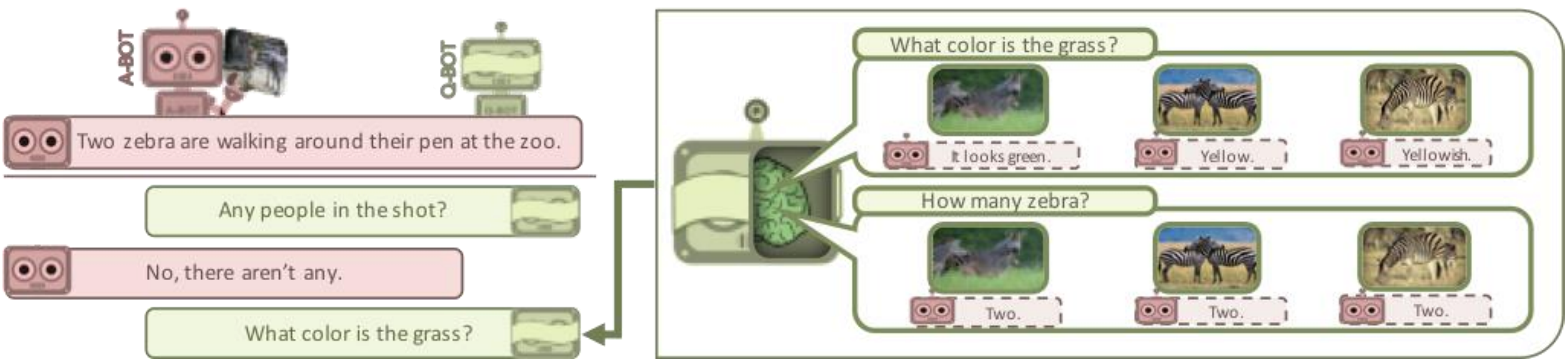
Negotiation Dialog Agents



Sneak Peek:

Inner Dialog:
Pragmatic Visual Dialog Agents that Rollout a Mental Model of their Interlocutors

Inner Dialog



What next?

- So far
  - Vision + Language
    - Captioning → VQA → Visual Dialog

- Interacting with an intelligent agent
  - Perceive + Communicate + Act
  - Vision + Language + Reinforcement Learning
  - Ok Google – can you find my picture where I was wearing this red shirt? And order me a new one?

Empty

Original Scene

There is a cloud on the left side of the frame all the way at the top. The rest of the sky is clear.

What is located in the grass?

There is a tent on the right side sitting on the grass with a bear behind it on the left side.

What else is on the grass?

There is a little boy sitting with his legs straight on the right side of the tent

Is there anything else?

Yes, there is a red grill to the right of the bear and a little girl sitting on the right side of the tent

If the tent is on the right side there is no room for the kids? Is the tent on the right side of the picture?

I am so sorry the grill is on the left of the bear. Yes, the little boy is on the bottom right of the picture

Where is the little girl?

He is in front of the tent. The tent is in the middle of the grass on the right side.

Is that all? Are we goo?

She is on the left side of the tent, she is crying with her legs straight

Great! Are we good?

Yes, we are good.

~ Fin. ~

Agents in Virtual Environments

AI2 Thor

SUNCG dataset

What next?

- So far
  - Vision + Language
    - Captioning → VQA → Visual Dialog

- Interacting with an intelligent agent
  - Perceive + Communicate + Act
  - Vision + Language + Reinforcement Learning
  - Ok Google – can you find my picture where I was wearing this red shirt? And order me a new one?

What next?

- So far
  - Vision + Language
    - Captioning → VQA → Visual Dialog
- Interacting with an intelligent agent
  - Perceive + Communicate + Act
  - Vision + Language + Reinforcement Learning
  - Ok Google – can you find my picture where I was wearing this red shirt? And order me a new one?
- Teaching with natural language
  - ”No, not that shirt. *This* one.”

ParlAI: A Dialog Research Software Platform

Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, Jason Weston

(Submitted on 18 May 2017)

We introduce ParlAI (pronounced "par-lay"), an open-source software platform for dialog research implemented in Python, available at this http URL Its goal is to provide a unified framework for training and testing of dialog models, including multitask training, and integration of Amazon Mechanical Turk for data collection, human evaluation, and online/reinforcement learning. Over 20 tasks are supported in the first release, including popular datasets such as SQuAD, bAbI tasks, MCTest, WikiQA, QACNN, QADailyMail, CBT, bAbI Dialog, Ubuntu, OpenSubtitles and VQA. Included are examples of training neural models with PyTorch and Lua Torch, including both batch and hogwild training of memory networks and attentive LSTMs.

ParlAI *(pronounced "parlay")*

A Dialogue Dataset "Universe"

QA datasets
bAbI tasks
MCTest
SquAD, NewsQA, MS MARCO
SimpleQuestions
WebQuestions, WikiQA
WikiMovies, MTurkWikiMovies
MovieDD (Movie-Recommendations)

Sentence Completion
QACNN
QADailyMail
CBT
BookTest

Dialogue Goal-Oriented
bAbI Dialog tasks
Camrest
Dialog-based Language Learning
MovieDD (QA,Recs dialogue)
CommAI-env

Dialogue Chit-Chat
Ubuntu multiple-choice
UbuntuGeneration
Movies SubReddit
Reddit
Twitter

VQA/Visual Dialog
VQA already in v1.0.

Add your own dataset!
Open source…

Machine Learning & Perception Group

Dhruv Batra
Assistant Professor

PhD

Qing Sun

Aishwarya Agrawal

Yash Goyal

Michael Cogswell

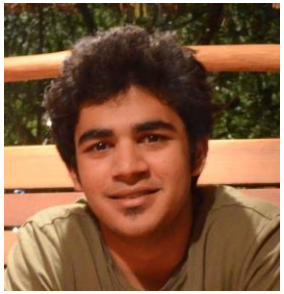
Abhishek Das

Ashwin Kalyan

MS

Aroma Mahendru

Akrit Mohapatra



Postdoc

Stefan Lee

Interns

Deshraj Yadav

Tejas Khot

Viraj Prabhu

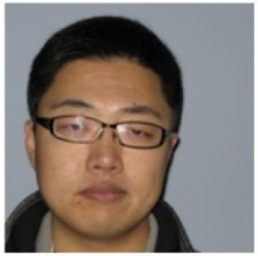
Computer Vision Lab

Devi Parikh
Assistant Professor

Xiao Lin
Ph.D. Student

Arjun Chandrasekaran
Ph.D. Student

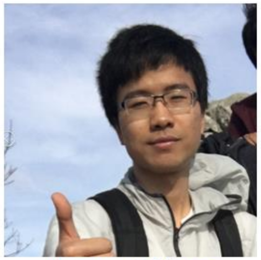
Ramakrishna Vedantam
Ph.D. Student



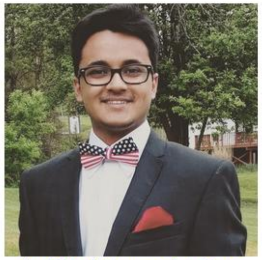
Peng Zhang
Ph.D. Student

Jiasen Lu
Ph.D. Student

Ram Prasaath Selvaraju
Ph.D. Student



Jianwei Yang
Ph.D. Student



Arijit (Arren) Ray
M.S. Student

Prithvijit Chattopadhyay
Intern

Thanks!