Carnegie Mellon

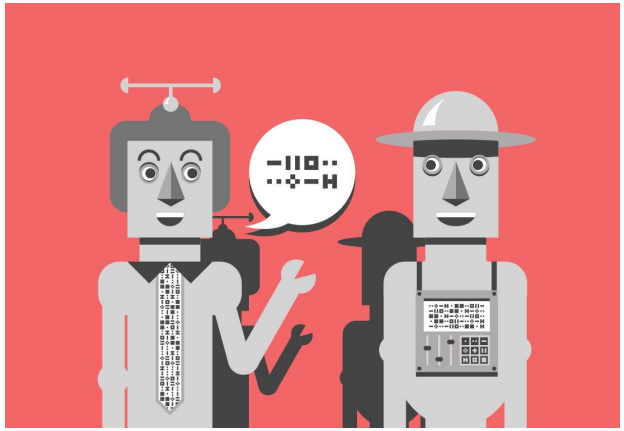
THE ROBOTICS INSTITUTE

Mind Your Language
Learning Visually Grounded Dialog in a Multi-Agent Setting

Akshat Agarwal*, Swaminathan Gurumurthy*, Vasu Sharma*, Katia Sycara

Adaptive Learning Agents Workshop
International Conference on Autonomous Agents and Multiagent Systems (AAMAS) 2018

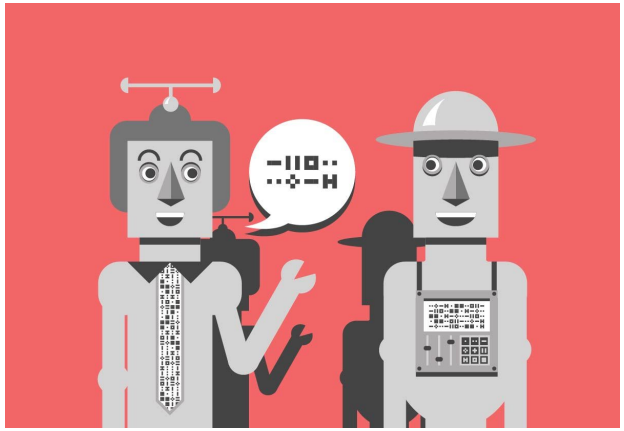
Motivation



Interpretable,
goal-oriented dialog
between artificial agents

- AI needs to communicate with humans conveying visual and textual info
  - use cases: assistive systems for visually impaired, assistants (like Siri/Alexa)
- It will become common to have two agents communicate with each other towards a goal, say, reserving a table (like Google Duplex)
- We want these conversations to be interpretable to humans for sake of transparency and ease of debugging

Motivation



Interpretable, goal-oriented dialog between artificial agents

- Humans adhere to natural language because they have to interact with an entire community
- Having a private language for each person would be inefficient
- Previous work on visual dialog showed a pair of agents adapting to each other start communicating in a private language to maximize the flow of information

We propose a multi-agent dialog framework (MADF) where each agent **interacts with and learns from multiple agents;** and show that it results in more coherent and human-interpretable dialog between agents, without compromising on task performance

Problem Statement

- Formulated as a conversation between two collaborative agents, a Question (Q-) Bot and an Answer (A-) Bot
- A-Bot given an image and a caption, while Q-Bot is given only a caption - both agents share a common objective, which is for Q-Bot to form an accurate mental representation of the unseen image
- Facilitated by exchange of 10 pairs of questions and answers between the two agents, using a shared common vocabulary

Carnegie Mellon

VisDial Dataset

THE ROBOTICS INSTITUTE

1. The VisDial dataset[1] contains ~80k images each with 10 pairs of questions and answers, collected from humans
2. To elicit temporal continuity, grounding in the image and naturalistic conversations, workers were paired on AMT to chat in real time
3. One worker (Q) saw only the caption to a hidden image, and had to ask questions about the image to 'imagine the scene' better
4. The 2nd worker (A) saw image and caption, and had to answer the questions
5. 10 pairs of questions and answers exchanged for each image

[1] Das, Abhishek, et al. "Visual dialog." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. 2017.

Carnegie Mellon

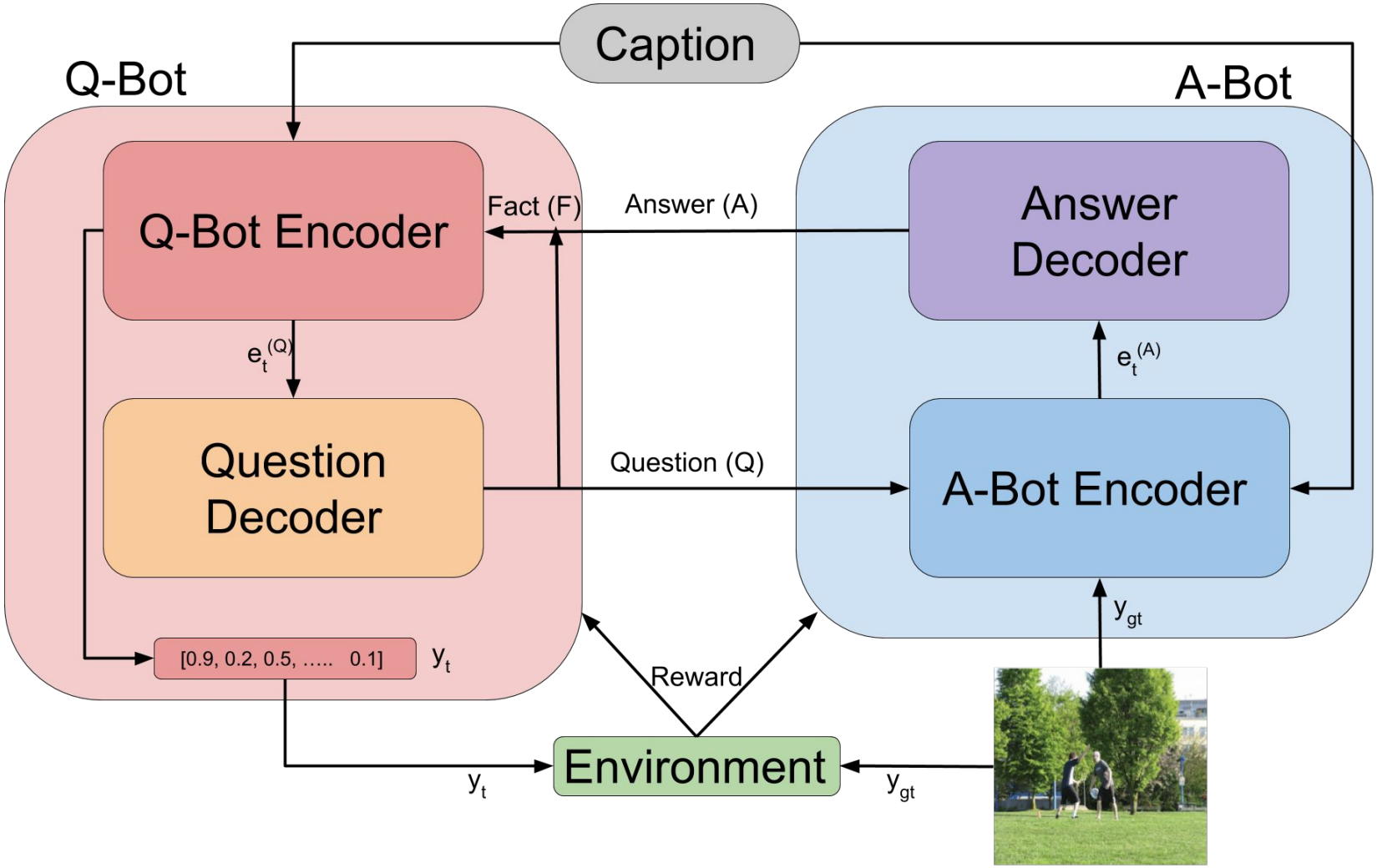
Overview

THE ROBOTICS INSTITUTE

1. The agents (Q-Bot and A-Bot) are pre-trained on the VisDial dataset using supervision[1]. They do not interact with each other in this phase
2. This is followed by making them interact and adapt to each other by reinforcement learning[2]
    a. They are rewarded *by the environment* to maximize transfer of information with each QA pair
    b. The transition from supervised to reinforcement learning is handled smoothly via a curriculum.

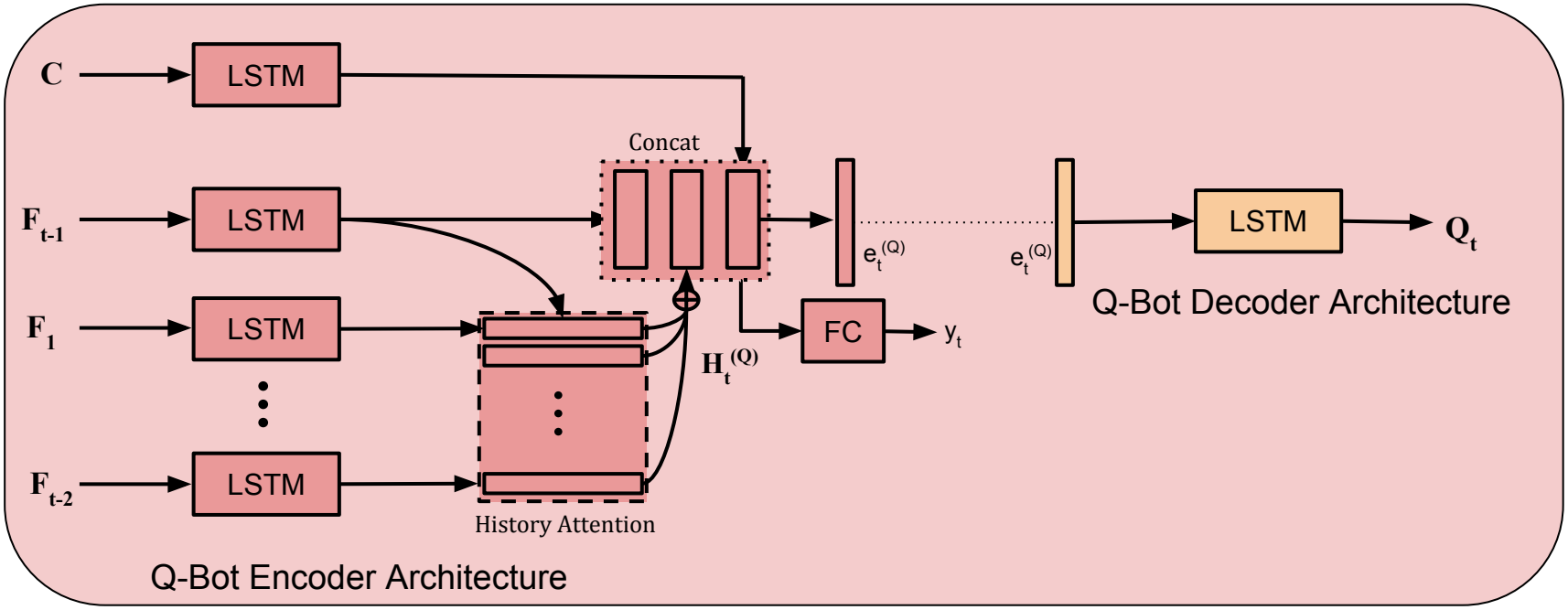
[1] Das, Abhishek, et al. "Visual dialog." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. 2017.
[2] Das, Abhishek, et al. "Learning cooperative visual dialog agents with deep reinforcement learning." arXiv preprint arXiv:1703.06585 (2017).

Visual Dialog RL Framework



Note: The agents have been pretrained on the VisDial dataset before interacting as shown above

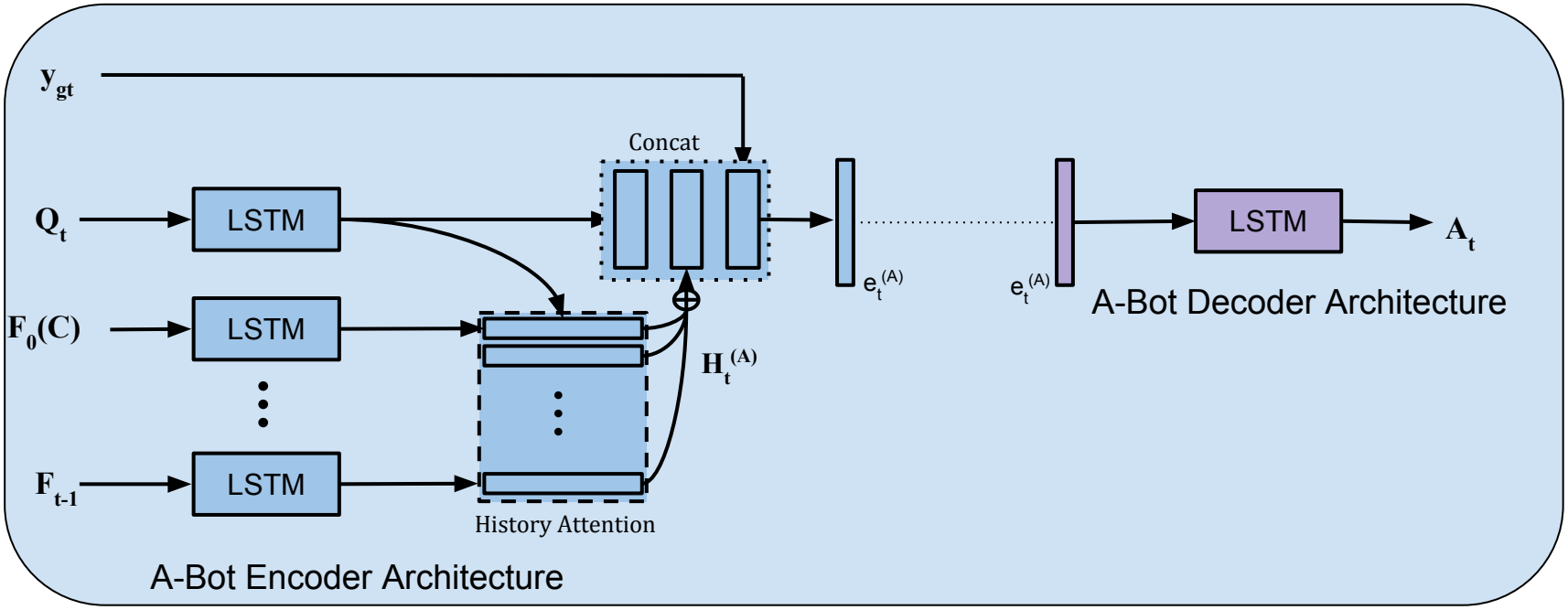
Agent Architectures



Fact (F_{t-1}) is concatenation of a question (Q_{t-1}) and its answer (A_{t-1})
C is the caption
History (F_1,F_2.. F_{t-2}) is a combination of all previous QA pairs for any particular image
Attend over history using fact, F_{t-1}, to produce $H_t^{(Q)}$
Concat F_{t-1}, $H_t^{(Q)}$ and C embeddings and pass through linear layers to get $e_t^{(Q)}$ and y_t

[1] Lu, Jiasen, et al. "Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model." *Advances in Neural Information Processing Systems*. 2017.

Agent Architectures



$F_0(C)$ is the caption

History (F_0,F_1,F_2.. F_{t-1}) is a combination of all previous QA pairs for any particular image

Attend over history using question, Q_t to produce $H_t^{(A)}$

Concat Q_t , $H_t^{(A)}$ and y_{gt} embeddings and pass through linear layer to get $e_t^{(A)}$

[1] Lu, Jiasen, et al. "Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model." *Advances in Neural Information Processing Systems*. 2017.

Carnegie Mellon

Training Methodology

THE ROBOTICS INSTITUTE

1. Two agents, a Q-Bot and an A-Bot are first trained in isolation via supervision from the VisDial dataset for 15 epochs
2. Then smoothly transitioned to reinforcement learning via a curriculum
   a. For the first K rounds of dialog for each image, agents are trained by supervision, and for remaining 10-K rounds they are made to interact and train via RL.
   b. K starts at 9 and is reduced to 0 over 10 epochs

[1] Das, Abhishek, et al. "Visual dialog." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. 2017.
[2] Das, Abhishek, et al. "Learning cooperative visual dialog agents with deep reinforcement learning." arXiv preprint arXiv:1703.06585 (2017).

Supervised Pre-Training

1. Both agents trained using a Maximum Likelihood Estimation (MLE) loss against the ground truth QA for every round of dialog[1]
2. Q-Bot simultaneously minimizes Mean Squared Error (MSE) loss between the true and predicted image embeddings[2]
3. Problems:
    a. MLE results in repetitive, 'safe' responses (*e.g. I don't know, I can't see)*
    b. No interaction during training leads to unexpected responses during testing when they interact with each other and face out of distribution QA

[1] Das, Abhishek, et al. "Visual dialog." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. 2017.
[2] Das, Abhishek, et al. "Learning cooperative visual dialog agents with deep reinforcement learning." arXiv preprint arXiv:1703.06585 (2017).

Carnegie Mellon

Reinforcement Learning

THE ROBOTICS INSTITUTE

1. Both agents allowed to interact with each other and learn by self-play[1]
2. No ground-truth data except images and captions
3. Q-Bot observes $\{c, q_1, a_1, \ldots, q_{10}, a_{10}\}$, A-Bot observes $\{I, c, q_1, a_1, \ldots, q_{10}, a_{10}\}$

I : image, c : caption, q_i, a_i: i^{th} dialog pair exchanged where i = [1,..10]

4. Action: Predict words sequentially until a stop token is encountered (or max length reached)
5. Reward: Incentivizing information gain from each round of QA, measured using the predicted image embedding y_t

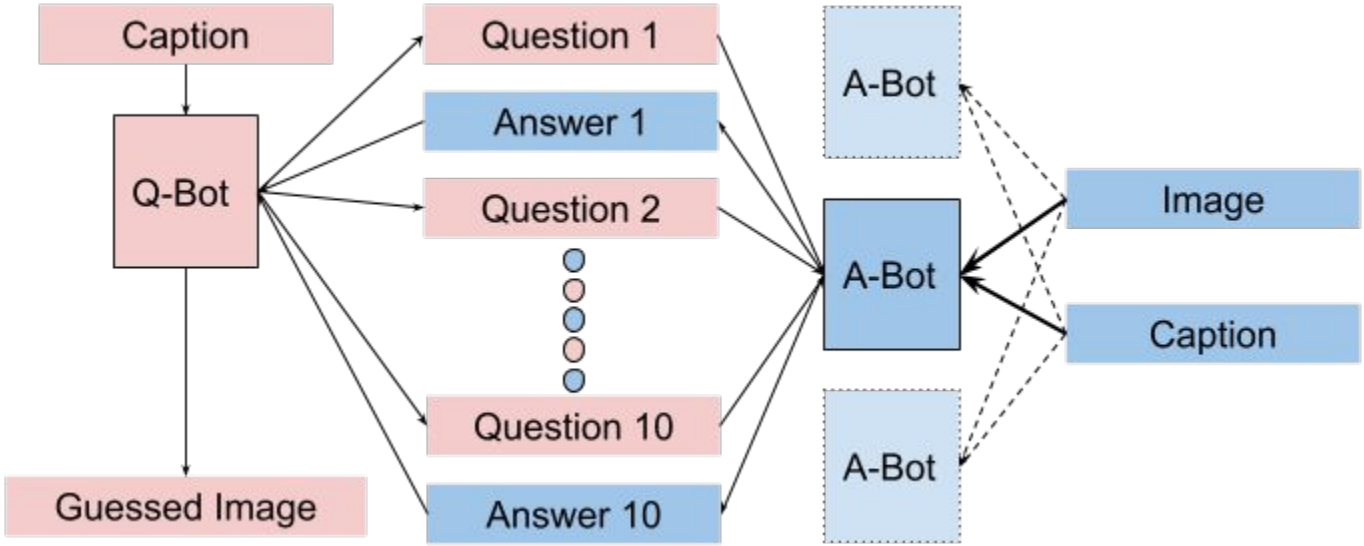
$$r_t(s_t^Q, (q_t, a_t, y_t)) = l(y_{t-1}, y^{gt}) - l(y_t, y^{gt})$$

Learn by REINFORCE

No motivation to stick to rules and conventions of English language[2], making the RL optimization problem ill-posed

[1] Das, Abhishek, et al. "Learning cooperative visual dialog agents with deep reinforcement learning." arXiv preprint arXiv:1703.06585 (2017).
[2] Kottur, Satwik, et al. "Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog." arXiv preprint arXiv:1706.08502 (2017)

Multi-Agent Dialog Framework (MADF)



We create either multiple Q-Bots to interact with a single A-Bot, OR multiple A-Bots to interact with a Q-Bot - and randomly pick one pair of agents to interact for each batch of images, and learn via REINFORCE

Carnegie Mellon

MADF Algorithm

THE ROBOTICS INSTITUTE

Algorithm 1 Multi-Agent Dialog Framework (MADF)

procedure MULTIBOTTRAIN
  while train_iter $<$ max_train_iter **do** ▷ Main Training loop over batches
    $Qbot \leftarrow$ *random_select* $(Q_1, Q_2, Q_3 .... Q_q)$
    $Abot \leftarrow$ *random_select* $(A_1, A_2, A_3 .... A_a)$ ▷ Either q or a is equal to 1
    $history \leftarrow (0, 0, ...0)$ ▷ History initialized with zeros
    $fact \leftarrow (0, 0, ...0)$ ▷ Fact encoding initialized with zeros
    $\Delta image\_pred \leftarrow 0$ ▷ Tracks change in Image Embedding
    $Qz_1 \leftarrow Ques\_enc(Qbot, fact, history, caption)$
    for t in 1:10 **do** ▷ Have 10 rounds of dialog
      $ques_t \leftarrow Ques\_gen(Qbot, Qz_t)$
      $Az_t \leftarrow Ans\_enc(Abot, fact, history, image, ques_t, caption)$
      $ans_t \leftarrow Ans\_gen(Abot, Az_t)$
      $fact \leftarrow [ques_t, ans_t]$ ▷ Fact encoder stores the last dialog pair
      $history \leftarrow concat(history, fact)$ ▷ History stores all previous dialog pairs
      $Qz_t \leftarrow Ques\_enc(Qbot, fact, history, caption)$
      $image\_pred \leftarrow image\_regress(Qbot, fact, history, caption)$
      $R_t \leftarrow (target\_image - image\_pred)^2 - \Delta image\_pred$
      $\Delta image\_pred \leftarrow (target\_image - image\_pred)^2$
    end for
    $\Delta(w_{Qbot}) \leftarrow \frac{1}{10} \sum_{t=1}^{10} \nabla_{\theta_{Qbot}} [G_t \log p(ques_t, \theta_{Qbot}) - \Delta image\_pred]$
    $\Delta(w_{Abot}) \leftarrow \frac{1}{10} \sum_{t=1}^{10} G_t \nabla_{\theta_{Abot}} \log p(ans_t, \theta_{Abot})$
    $w_{Qbot} \leftarrow w_{Qbot} + \Delta(w_{Qbot})$ ▷ REINFORCE and Image Loss update for Qbot
    $w_{Abot} \leftarrow w_{Abot} + \Delta(w_{Abot})$ ▷ REINFORCE update for Abot
  end while
end procedure

Quantitative Results for Answer Ranking

Model	MRR	Mean Rank	R@1	R@5	R@10
Answer Prior (Das et al., 2016)	0.3735	26.50	23.55	48.52	53.23
MN-QIH-G (Das et al., 2016)	0.5259	17.06	42.29	62.85	68.88
HCIAE-G-DIS (Lu et al., 2017)	0.547	14.23	44.35	65.28	71.55
Frozen-Q-Multi (Das et al., 2017)	0.437	21.13	N/A	53.67	60.48
CoAtt-GAN (Wu et al., 2017)	0.5578	14.4	**46.10**	**65.69**	71.74
SL(Ours)	**0.610**	**5.323**	34.74	57.67	**72.68**
RL - 1Q,1A(Ours)	0.459	7.097	16.04	54.69	72.34
RL - 1Q,3A(Ours)	0.601	5.495	34.83	57.47	72.48
RL - 3Q,1A(Ours)	0.590	5.56	33.59	57.73	72.61

About the Metrics: **Mean Rank and MRR** compute the average rank (and average of their reciprocals), respectively, assigned to the ground truth answer, over a set of 100 candidate answers for each question (provided in the VisDial dataset). **Recall@k** computes the percentage of answers with rank less than k

Used VisDial v0.9 dataset, with 83k train images + 40k test images. Results are reported on the test set

We outperform all previous architectures in MRR, Mean Rank and Recall @ 10, showing consistently good answers

While RL-1Q,1A performance drops (since it is being optimized for image estimation and not answer ranking, unlike SL), our multi-agent systems RL-1Q,3A and 3Q,1A **recover most of the performance gap**

Quantitative Results for Image Retrieval

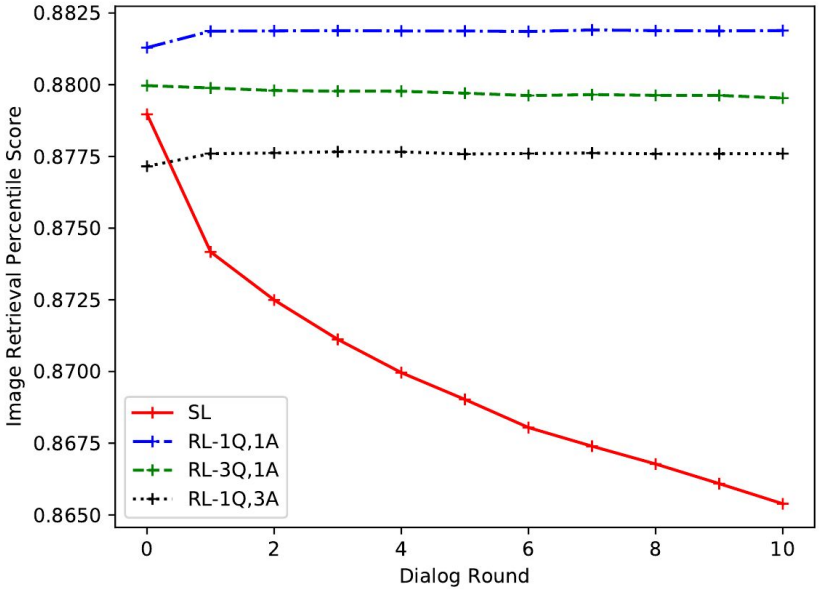


Fig. shows Image Retrieval Percentile Score (Y-axis) vs Dialog Round, from 1-10 (X-axis)

The score is calculated by ranking the Q-Bot's prediction of the image over all 40k images in the VisDial test set

1. While the performance of SL decreases (because of the nature of LSTMs to forget), the performance of our RL systems (RL-1Q,3A and RL-3Q,1A) remain constant
2. This, combined with previous result, validate our hypothesis that our multi-agent dialog framework results in more coherent dialog without compromising on task performance

Qualitative Evaluations of Generated Language by Humans

Carnegie Mellon

THE ROBOTICS INSTITUTE

	Metric	N	Supervised	RL 1Q,1A	RL 1Q,3A	RL 3Q,1A
1	Q-Bot Relevance	8	2.5	2.75	**2**	2.75
2	Q-Bot Grammar	8	**2.25**	2.875	2.5	2.375
3	A-Bot Relevance	12	2.5	2.583	2.25	**1.67**
4	A-Bot Grammar	12	1.92	3.5	**1.83**	2.25
5	Overall Coherence	20	2.8	3.05	**2.3**	**1.85**

1. 20 evaluators shown one image and the 10 QA pairs generated for 4 systems, asked to rank on each metric
2. Overall Coherence of 1Q,3A and 3Q,1A ranked much better
3. Q-Bot more relevant in 1Q,3A and A-Bot more relevant in 3Q,1A which proves our hypothesis that exposure to multiple agents to interact with improves relevance

Sample Generations

2 men in a grassy park playing frisbee

Human-Human Dialogue

How many men?	2
Are they playing frisbee?	Yes
What color frisbee?	White
Just white?	Yes
Do men look skilled?	No
Why not?	Playing
They don't seem good?	Guess not, I said no
Is frisbee in air?	No
Who has it?	Person 1
Is he bigger than other?	no

SL

are the boys playing volleyball ?	yes , a professional boy
how old do they look ?	maybe in their late 20 's
do you see any buildings ?	no , i don't see buildings
are they male ?	yes , both men and women
can you see any cars ?	no , can only see the bunch of
can you tell how many women	there are more people than 3
are sitting there or playing ?	no , they 're wood
are they combing leather ?	women in
what color are the men 's shoes?	they are black
what color are the women ?	'the snowboarder is white
are the boys well groomed ?	yes

RL - 1Q, 1A

do you recognize 1 ball ?	looks like a around
that you could you are seeing	that i can see the fair
controllers ?	
do you recognize the guy and	you couldn't be late teens , his
age is the adult ?	
do you recognize them ?	motorcycle but UNK tint and UNK
how many people are there ?	35
do you do they fence famous ?	It looks black and white
do you see fire ?	I'm not sure because it's is
is the fire dog UNK ?	Yes
do they have long manes ?	Black and white, it looks
can you see other people	Yes there is a people
watching them ?	

RL - 1Q, 3A

How old do the men appear?	30s
Is this at a beach?	No
Do they have on bathing suits?	No
How old are they?	Young adults
What color frisbee?	White
Do they have a regular ball shirt on?	1 of them do
With how old are they?	Mid 30s
Is there other people in the pic?	Yes, there is a man behind the him
How many of them are playing?	2
What is the woman doing?	sitting

RL - 3Q, 1A

What color is umbrella?	Black with a blue stripe
What are they wearing?	T shirts and jeans
What color is frisbee?	White
What are they doing?	Sitting on the beach, talking
Are they all holding rackets?	Yes
Are there any other people?	Yes
What color is the frisbee?	Creamy green
Are there any other people?	Yes a lot
Are the people tall?	Looks very tall
Are they in a field?	no

Carnegie Mellon

Future Work

THE ROBOTICS INSTITUTE

1. Understanding why task performance (image retrieval score) does not improve with dialog rounds (though intuitively, with more information having been exchanged, it should!)
2. Improving the image embeddings used as ground truth to have richer information

Carnegie Mellon

Conclusion

THE ROBOTICS INSTITUTE

1. Through quantitative evaluations of answer ranking and image retrieval task performance, we show that our multi-agent systems generate interpretable dialog without compromising on task performance
2. Through qualitative evaluations of dialog relevance and coherence by humans, we show that our multi-agent systems produce more coherent dialog
3. This approach has also been validated in related works published in parallel:
    a. Cao, Kris, et al. "Emergent Communication through Negotiation." arXiv preprint arXiv:1804.03980 (2018).
    b. Lee, Jason, et al. "Emergent translation in multi-agent communication." arXiv preprint arXiv:1710.06922 (2017).

Github: https://goo.gl/gc6dGZ

Thank You!

Questions?