

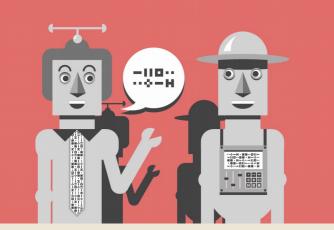
# Mind Your Language: Learning Visually **Grounded Dialog in a Multi-Agent Setting**

## Abstract

Humans adhere to natural language because they have to interact with an entire community

Having a private language for each

person would be inefficient



Interpretable, goaloriented dialog between artificial agents

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 humaninterpretable dialog between agents, without compromising on task performance

18:

19:

22:

23:

24:

end while

25: end procedure

## **Problem Statement**

- Formulated as a conversation between two collaborative agents, a Question (Q-) Bot and an Answer (A-) Bot
- A-Bot given an image, while Q-Bot is given only a caption to the image - both agents share a common objective, which is for Q-Bot to form an accurate mental representation of the unseen image

• Reinforcement Learning uses the change in distance between the predicted image embedding and the ground truth embedding as the reward function which is shared by both the Q and A bot. We REINFORCE the the using algorithm. system train

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

- No explicit incentive to maintain natural language and hence prone to deviate from it to optimize transfer of information between bots
- We solve the problem using our Multi Agent setup where we arbitrarily pick a Q and A bot pair and carry out a round of training for them and keep repeating the process
- Much harder for the bots to deviate from natural language in this setting as coming up with a new language pair for each pair of bots is highly inefficient

Algo	orithm 1 Multi-Agent Dialog Framework (MADF)	
1: ]	procedure MultiBotTrain	
2:	while train_iter < max_train_iter do	Main Training loop over batches
3:	$Qbot \leftarrow random\_select (Q_1, Q_2, Q_3, Q_q)$	
4:	$Abot \leftarrow random\_select (A_1, A_2, A_3A_a)$	$\triangleright$ Either q or a is equal to 1
5:	$history \leftarrow (0, 0, 0)$	History initialized with zeros
6:	$fact \leftarrow (0, 0,0)$	Fact encoding initialized with zeros
7:	$\Delta image\_pred \leftarrow 0$	Tracks change in Image Embedding
8:	$Qz_1 \leftarrow Ques\_enc(Qbot, fact, history, capt$	ion)
9:	<b>for</b> t in 1:10 <b>do</b>	▷ Have 10 rounds of dialog
10:	$ques_t \leftarrow Ques\_gen(Qbot, Qz_t)$	
11:	$Az_t \leftarrow Ans\_enc(Abot, fact, history, im)$	$aage, ques_t, caption)$
12:	$ans_t \leftarrow Ans\_gen(Abot, Az_t)$	
13:	$fact \leftarrow [ques_t, ans_t]$	▷ Fact encoder stores the last dialog pair
14:	$history \leftarrow concat(history, fact)$	▷ History stores all previous dialog pairs
15:	$Qz_t \leftarrow Ques\_enc(Qbot, fact, history, c)$	aption)
16:	$image\_pred \leftarrow image\_regress(Qbot, feetbook)$	act, history, caption)
17.	$R_t \leftarrow (target image - image pred)^2 -$	- Aimage pred

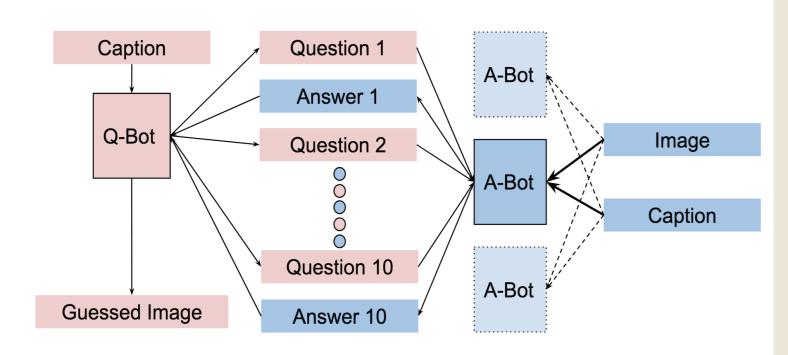
	Metric	Ν	Supervised	RL 1Q,1A	RL 1Q,3A	RL 3Q,1A
1	Q-Bot Relevance	8	2.5	2.75	2	2.75
2	Q-Bot Grammar	8	2.25	2.875	2.5	2.375
3	A-Bot Relevance	12	2.5	2.583	2.25	1.67
4	A-Bot Grammar	12	1.92	3.5	1.83	2.25
5	Overall Coherence	20	2.8	3.05	2.3	1.85

Table 2: Human Evaluation Results - Mean Rank (Lower is better). N refers to the number of human evaluators involved in the ranking.

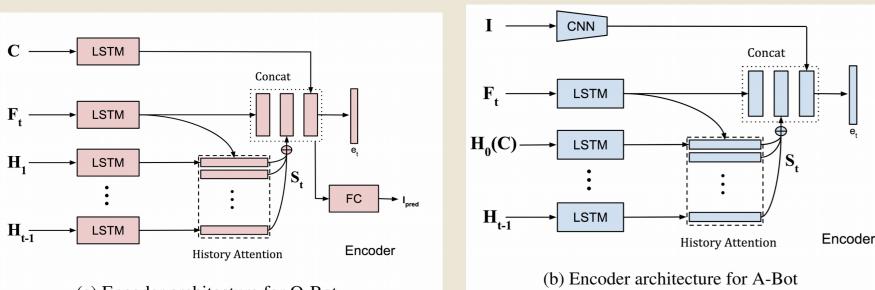
	The little girl is standing with skis on her feet Human-Human				n Dialogue			SL			
5			what color are the skis ? Are there any other people? Is this outdoors? Do you see snow? Is it currently snowing? Is she on a slope or hill? Do you see trees? Do you see the sky? Is she wearing gloves? Is she wearing a hat?		A UNK color Not that i can see Yes Yes No I don't think so No i don't think so Yes No Yep yes		'is sl is sh wha 'is sl 'is sl wha is sh wha	is the girl close to a race ? 'is she wearing a helmet ? is she wearing a hat ? what color is it ? 'is she wearing goggles ? 'is she wearing goggles ? what color is her skis ? is she wearing a hat ? what color does she have ? and what is the girl wearing ?		no , woman is riding no She is wearing a headband I would say it's black Can't see her face Not sure Dark, hard to tell i would guess No White A white striped long sleeve shirt and blue	
	RL - 10	Q, 1A			RL -	1Q, 3A			RL -	3Q, 1A	
es	is there more than 1 skier ? oh how many UNK do you see ? is the person holding a pipe of so me sort of basket ? is lady going fast ? how many is there ? is there snow ? how many animals do you see ? is there a lot of snow ? are they all together ? is there a lift ?	Y A fr Yes, there is a bow I can not f It's not silver people are in grocc Is covered by snow covered by snow Of the 1, and a Yes tons of snow everywhere cover Yes 1 is a darke Just a few skiing out y	rtie tell ery ow a 2 red r 1	are there any other people a is it snowing ? is it a sunny day ? is the woman wearing a dres what color is it ? is she by herself of snow ? what color is her jacket ? does it look like a competitio is it day or night ? is it sunny ?	ss ?	It has a white meat a Just lady, woma		how old is the worn is she in a ski resc can you see the sk is she alone ? what color are the is there snow on tf is the photo very c is she wearing a h is it snowing ? is it sunny ?	rt ? sy ? gloves ? ne ground ? lear ?	About 16 I don't think so No Yes Silver No snow Yes No it looks like it is stopped No really	
1						JL				JL	
OS	2 men in a grassy par	k playing inspee		Human-Human	Dialogu	]		S	<u> </u>		
os 1g			How many men? Are they playing frisbee? What color frisbee? Just white? Do men look skilled? Why not? They don't seem good? Is frisbee in air? Who has it? Is he bigger than other?		Yes h White d Yes a No c Playing c Guess not, I said no a No a Person 1 w no w		how old do do you see are they m can you se can you te are sitting are they c what color what color	can you see any cars ? no		yes, a professional boy maybe in their late 20 's no, i don't see buildings yes, both men and women can only see the bunch of ere are more people than 3 no, they 're wood women in they are black 'the snowboarder is white yes	
og		and the second second									
	RL - 1	RL - 1Q, 1A			RL - 1Q, 3A			RL - :		A	
uir rs	do you recognize 1 ball ? that you could you are seeing controllers ? do you recognize the guy and age is the adult ? do you recognize them ? how many people are there ? do you do they fence famous ? do you do they fence famous ? do you see fire ? is the fire dog UNK ? do they have long manes ? can you see other people watching them ?	looks like a around that i can see the fair you couldn't be late teens , his motorcycle but UNK tint and UNK 35 It looks black and white I'm not sure because it's is Yes Black and white, it looks Yes there is a people		How old do the men appear? Is this at a beach? Do they have on bathing suits? How old are they? What color frisbee? Do they have a regular ball shirt of With how old are they? Is there other people in the pic? How many of them are playing? What is the woman doing?			White nem do 1id 30s	What color is umbrella? What are they wearing? What color is frisbee? What are they doing? Are they all holding racke Are there any other peopl What color is the frisbee? Are there any other peopl Are the people tall?	ts? e?	Black with a blue stripe T shirts and jeans White ting on the beach, talking Yes Yes Creamy green Yes a lot Looks very tall no	

- Facilitated by exchange of 10 pairs of questions and answers between the two agents, using a shared common vocabulary
- Pretraining the agents with supervision from the VisDial 20: dataset, followed by making them interact and adapt to each 21: other via reinforcement learning maximizes task performance, but the agents learn to communicate in nongrammatical and semantically meaningless sentences, hence motivating our multi-agent setup

## Method



### Figure 1: Multi-Agent (with 1 Q-Bot, 3 A-Bots) Dialog Framework



$R_t \leftarrow (target\_image - image\_pred)$	$)^2 - \Delta i mage_pred$
$\Delta image\_pred \leftarrow (target\_image-i$	$mage\_pred)^2$
end for	
$\Delta(w_{Qbot}) \leftarrow \frac{1}{10} \sum_{t=1}^{10} \nabla_{\theta_{Qbot}} \left[ G_t \log p(qt) \right]$	$ues_t, \theta_{Qbot}) - \Delta image\_pred]$
$\Delta(w_{Abot}) \leftarrow \frac{1}{10} \sum_{t=1}^{10} G_t \nabla_{\theta_{Abot}} \log p(an)$	$s_t,  heta_{Abot})$
$w_{Qbot} \leftarrow w_{Qbot} + \Delta(w_{Qbot})$	REINFORCE and Image Loss update for Qbot
$w_{Abot} \leftarrow w_{Abot} + \Delta(w_{Abot})$	REINFORCE update for Abot

Model	MRR	Mean Rank	<b>R@1</b>	R@5	<b>R@10</b>
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

Table 1: Comparison of Metrics with Literature

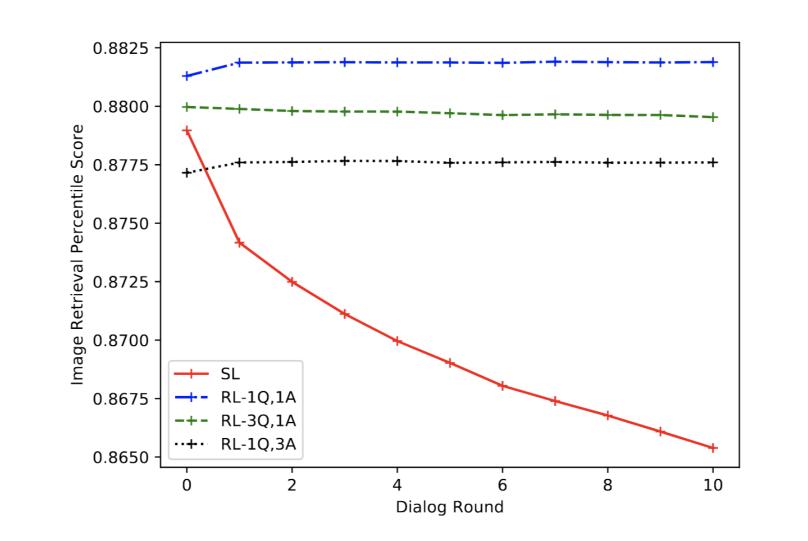


Figure 4: Two randomly selected images from the VisDial dataset followed by the ground truth (human) and generated dialog about that image for each of our 4 systems (SL, RL-1Q,1A, RL-1Q,3A, RL-3Q,1A). These images were also used in the human evaluation results shown in Table 2

## **Future Work**

- We plan to explore several other multi bot architectural settings and perform a more thorough human evaluation for qualitative analysis of our dialog.
- We also plan on incorporating other language priors in our reinforcement learning setup to further improve the dialog quality.
- We will also experiment with using a discriminative answer decoder which uses information of the possible answer candidates to rank the generated answer with respect to all the candidate answers and use the ranking performance to train the answer decoder.

## References

- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2016.Visual dialog.CoRR, abs/1611.08669
- Abhishek Das, Satwik Kottur, Jose M. F. Moura, StefanLee, and Dhruv Batra. 2017. Learning cooperativevisual dialog agents with deep reinforcement learn-ing.CoRR, abs/1703.06585.
- Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv

#### (a) Encoder architecture for Q-Bot

• 1 Q-Bot and 1 A-Bot are trained in isolation via supervision to optimize the MLE objective, leading to uninformative and repetitive dialog and inability to respond to out of distribution questions/answers

• We use Curriculum based learning to smoothly transition from supervised learning to Reinforcement Learning

Figure 3: The percentile scores of the ground truth image compared to the entire test set of 40k images. The X-axis denotes the dialog round number (from 1 to 10), while the Y-axis denotes the image retrieval percentile score.

Batra. 2017.Best of both worlds: Transferring knowledge from discriminative learn-ing to a generative visual dialog model. CoRR, abs/1706.01554 • Satwik Kottur, Jos'e M. F. Moura, Stefan Lee, and Dhruv Batra. 2017.Natural language does notemerge 'naturally' in multi-agent dialog.CoRR,abs/1706.08502. • Mike Lewis, Denis Yarats, Yann N Dauphin, DeviParikh, and Dhruv Batra. 2017. Deal or no deal?end-to-end learning for negotiation dialogues.arXivpreprint arXiv:1706.05125