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

Akshat Agarwal, Swaminathan Gurumurthy, Vasu Sharma, Katia Sycara*
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
aa7@cmu.edu,sgurumur@andrew.cmu.edu,vasus@andrew.cmu.edu,katia@cs.cmu.edu

ABSTRACT

The task of visually grounded dialog involves learning goal-oriented cooperative dialog between autonomous agents who exchange information about a scene through several rounds of questions and answers. We posit that requiring agents to adhere to rules of human language while also maximizing information exchange is an ill-posed problem, and observe that humans do not stray from a common language because they are social creatures and have to communicate with many people everyday, and it is far easier to stick to a common language even at the cost of some efficiency loss. Using this as inspiration, we propose and evaluate a multi-agent dialog framework where each agent interacts with, and learns from, multiple agents, and show that this results in more relevant and coherent dialog (as judged by human evaluators) without sacrificing task performance (as judged by quantitative metrics).

KEYWORDS

Visual Dialog; Multi Agent Reinforcement Learning; Curriculum Learning; Emergent Communication

1 INTRODUCTION

AI is increasingly becoming an important part of our daily lives, be it in the household, the workplace or in public places. In order for humans to be able to interact with and understand the AI system, it needs to learn how to communicate with us about our environment using the languages that we speak. This requires the AI system to visually interpret the world, and communicate descriptions of the physical world. While such a task would have been considered impossible a few years ago, the recent progress in the fields of Computer Vision and Natural Language Processing, which are important building blocks for this task, have reinvigorated interest in the community. Several problems like image captioning ([10], [33], [27], [9], [17], [34]), image classification ([12], [26], [7], [31]), object detection ([14], [20], [21]), image segmentation ([15], [8], [19]), dialog ([25], [28], [5]), question answering ([35], [24], [32]) etc. have received immense amounts of attention from the research community. The paradigm of reinforcement learning has also shown promising results in several problems including learning to play Go [23] and Atari games [18], among others, at superhuman levels.

Capitalizing on the growth in all these different domains, it now seems plausible to build more advanced dialog systems capable of reasoning over multiple modalities while also learning from one another. Such systems will allow humans to have a meaningful dialog with intelligent systems containing visual as well as textual content. Use cases include assistive systems for the visually impaired, smart multimodal dialog agents (unlike current versions of Siri and Alexa which are primarily audio based and cannot make effective use of multimodal data) and even large scale visual retrieval systems. However, as these systems become more advanced, it will become increasingly common to have two agents interact with each other to achieve a particular goal [13]. We want these conversations to be interpretable to humans for the sake of transparency and ease of debugging. This motivates our work on goal-driven agents which interact in coherent language understandable to humans.

This paper presents work on Goal driven Visual Dialog Agents. Most prior work on visual dialog [[2], [4]] has approached the problem using supervised learning where, conditioned on the question - answer pair dialog history, a caption c and the image I, the agent is required to answer a given question q. The model is trained in a supervised learning framework using ground truth supervision from a human-human dialog dataset.

Some recent work [3] has tried to approach the problem using reinforcement learning, with two agents, namely the Question (Q-) Bot and the Answer (A-) Bot. While the A-Bot still has the image, caption and the dialog history to answer any question, the Question Bot only has access to the caption and the dialog history. The two agents are initially trained with supervision using the VisDial v0.9 dataset [2], which consists of 80k images, each with a caption and 10 question-answer pairs discussing the image. Under supervision, the agents are trained in an isolated manner to maximize the likelihood of generating the ground truth answers. The agents are then made to interact and talk to each other, with a common goal of trying to improve the Q-Bot's understanding of the image. The agents learn from their conversation with each other via reinforcement learning. While the supervised training *in isolation* helps the agents to learn to interpret the images and communicate information, it is the *interactive* training phase which leads to richer dialog with more informative questions and answers as the agents learn to adapt to each others' strengths and weaknesses. However, it is important to note that the optimization problem in our conversational setting does not make the agents stick to the domain of grammatically correct and coherent natural language. Indeed, if the two agents are made to communicate with and learn from each other for too long, they quickly start generating non-grammatical and semantically meaningless sentences. It's important to note here that while the generated sentences stop making sense to human observers, the two agents are able to understand each other much better, and the Q-Bot's understanding of the image improves. This is similar to how twins often develop a private language [22], an idiosyncratic and exclusive form of communication understandable only to them.

* AA, SG and VS contributed equally to this paper

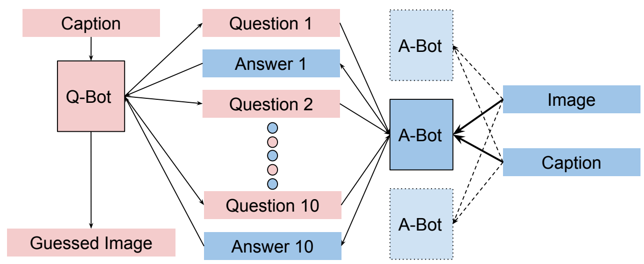


Figure 1: Multi-Agent (with 1 Q-Bot, 3 A-Bots) Dialog Framework

In our case, however, this reduces transparency of the agents' dialog to any observer (human or AI), and is hence undesirable. Prior work [2, 3] which has focused on improving performance as measured by the Q-Bot's image retrieval rank has suffered from incoherent dialog. We address this problem of improving the agents' performance while increasing dialog quality by taking inspiration from humans. We observe that humans continue to speak in commonly spoken languages simply, and hypothesize that this is because they need to communicate with an entire community, and having a private language for each person would be extremely inefficient. With this idea, we let our agents learn in a similar setting, by making them talk to (ask questions of, get answers from) multiple agents, one by one.

In the subsequent sections we describe the Visual Dialog task and the neural network architectures of our Q-Bots and A-Bots in detail. We then describe the training process of the agents sequentially: (a) in isolation (via supervision), (b) while interacting with one partner agent (via reinforcement), and (c) our proposed multi-agent setup where each agent interacts with multiple other agents (via reinforcement). We compare the performance of the agents trained in these different settings, both quantitatively using image retrieval ranks, and qualitatively evaluating the coherence and relevance of the dialog generated, as judged by impartial human evaluators. We make the following contributions: we show that our multi-agent dialog setup ensures that the interactions between the agents remain grounded in the rules and grammar of natural language, are coherent and human-interpretable without compromising on task performance. The grounded dialog makes the agents more helpful, transparent and trustworthy. We make our code available as open-source[1] .

2 PROBLEM STATEMENT

We begin by defining the problem of Visually Grounded Dialog for the co-operative image guessing game on the VisDial dataset.

Players and Roles: The game involves two collaborative agents – a question bot (Q-bot) and an answer bot (A-bot). The A-bot has access to an image, and the Q-bot has access to the image's caption, but not the image itself. Both the agents share a common objective, which is for the Q-bot to form a good mental representation of the unseen image using which it can retrieve, rank or generate that image. This is facilitated by the exchange of 10 pairs of questions and answers between the two agents, using a shared vocabulary, where the Q-bot asks the A-bot a question about the image, and the A-bot answers the question, hence improving the Q-Bot's understanding of the image scene.

General Game Objective: At each round, in addition to communicating, Q-bot must provide a 'description' $\hat{y}$ of the unknown image I based only on the dialog history. Both agents receive a common reward from the environment which is inversely proportional to the error in this description under some metric $L(\hat{y}, y_{gt})$. We note that this is a general setting where the 'description' $\hat{y}$ can take on varying levels of specificity – from image feature embeddings extracted by deep neural networks to textual descriptions and pixel-level image re-generations.

Specific Instantiation: In our experiments, we focus on the setting where the Q-bot is tasked with estimating a vector embedding of the image I, which is later used to retrieve a similar image from the dataset. Given a feature extractor (say, a pretrained CNN model like VGG [26]), the target 'description' y_{gt} of the image, can be obtained by simply forward propagating through the VGG model, without the requirement of any human annotation. Reward/error can be measured by simple Euclidean distance between the target description y_{gt} and the predicted description $\hat{y}$, and any image may be used as the visual grounding for a dialog. Thus, an unlimited number of games may be simulated without human supervision, motivating the use of reinforcement learning in this framework.

Our primary focus for this work is to ensure that the agents' dialog remains coherent and understandable while also being informative and improving task performance. For concreteness, an example of dialog that is informative yet incoherent: question: "do you recognize the guy and age is the adult?", answered with: "you couldn't be late teens, his". The example shows that the bots try to extract and convey as much information as possible in a single question/answer (sometimes by incorporating multiple questions or answers into a single statement). But in doing so they lose basic semantic and syntactic structure.

3 RELEVANT WORK

Most of the major work which combine vision and language have traditionally been focusing on the problem of image captioning [33] and visual question answering [1]. The problem of visual dialog is relatively new and was first introduced by Das et al. [2] who also created the VisDial dataset to advance the research on visually grounded dialog. The dataset was collected by pairing two annotators on Amazon Mechanical Turk to chat about an image. They formulated the task as a 'multi-round' VQA task and evaluated individual responses at each round in an image guessing setup. In a subsequent work by Das et al. [3] they proposed a Reinforcement Learning based setup where they allowed the Question bot and the Answer bot to have a dialog with each other with the goal of correctly predicting the image unseen to the Question bot. However, in their work they noticed that the Reinforcement Learning based training quickly lead the bots to diverge from Natural Language. In fact [11] recently showed that language emerging from two agents interacting with each other might not even be interpretable or compositional. Our multi-agent framework aims to alleviate this problem and prevent the bots from developing a

[1]https://github.com/agakshat/visualdialog-pytorch

specialized language between them. Interleaving supervised training with reinforcement learning also helps prevent the bots from becoming incoherent and generating non-sensical dialog.

Lu et al. [16] proposed a generative-discriminative framework for visual dialog where they train only an answer bot to generate informative answers for ground truth questions. These answers were then fed to a discriminator, which was trained to rank the generated answer among a set of candidate answers. This is a major restriction of their model as it can only be trained when this additional information of candidate answers is available, which restricts it to a supervised setting. Furthermore, since they train only the answer bot and have no question bot, they cannot simulate an entire dialog which also prevents them from learning by self-play via reinforcement. Wu et al. [30] further improved upon this generative-discriminative framework by formulating the discriminator as a more traditional Generative Adversarial network (GAN) [6], where the adversarial discriminator is tasked to distinguish between human generated and machine generated dialogs. Furthermore, unlike [16] they modeled the discriminator using an attention network which also utilized the dialog history in predicting the next answer allowing it to maintain coherence and consistency across dialog turns.

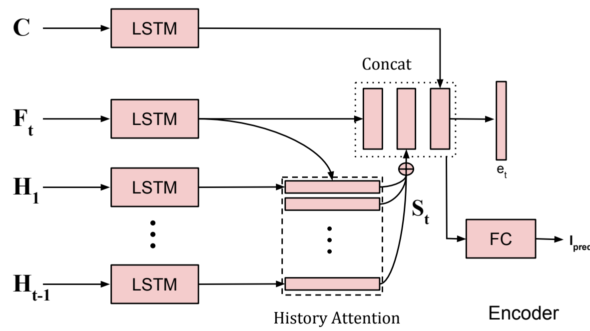
4 AGENT ARCHITECTURES

In this section we describe our model architectures in detail.

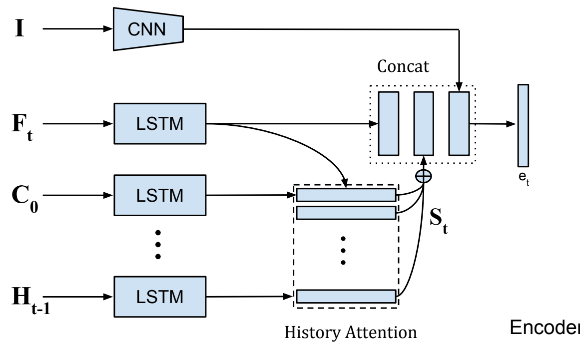
4.1 Question Bot Architecture

The question bot architecture we use is inspired by the answer bot architecture in [3] and [16] but the individual units have been modified to provide more useful representations. Similar to the original architecture, our Q-Bot, shown in Fig. 2a, also consists of 4 parts, (a) fact encoder, (b) state-history encoder, (c) question decoder and (d) image regression network. The fact encoder is modelled using a Long-Short Term Memory (LSTM) network, which encodes the previous question-answer pair into a fact embedding F_t^Q. We modify the state-history encoder to incorporate a two-level hierarchical encoding of the dialog. It uses the fact embedding F_t^Q at each time step to compute attention over the history of dialog, $(F_1^Q, F_2^Q, F_3^Q \ldots F_{t-1}^Q)$ and produce a history encoding S_t^Q. The key modification (compared to [16]) in our framework is the addition of a separate LSTM to compute a caption embedding C. This is key to ensuring that the hierarchical encoding does not exclusively attend on the caption while generating the history embedding. The caption embedding is then concatenated with F_t^Q and S_t^Q, and passed through multiple fully connected layers to compute the state-history encoder embedding e_t and the predicted image feature embedding $\hat{y}_t = f(S_t^Q)$. The encoder embedding, e_t^Q is fed to the question decoder, another LSTM, which generates the question, q_t. For all LSTMs and fully connected layers in the model we use a hidden layer size of 512. The image feature vector is 4096 dimensional. The word embeddings and the encoder embeddings are 300 dimensional.

4.2 Answer Bot Architecture



(a) Encoder architecture for Q-Bot



(b) Encoder architecture for A-Bot

Figure 2

The architecture for A-Bot, also inspired from [16], shown in Fig. 2b, is similar to that of the Q-Bot. It has 3 components: (a) question encoder, (b) state-history encoder and (c) answer decoder. The question encoder computes an embedding, Q_t for the question to be answered, q_t. The history encoding $(F_1^A, F_2^A, F_3^A \ldots F_t^A) \rightarrow S_t^A$ uses a similar two-level hierarchical encoder, where the attention is computed using the question embedding Q_t. The caption is passed on to the A-Bot as the first element of the history, which is why do not use a separate caption encoder. Instead, we use the fc7 feature embedding of a pretrained VGG-16 [26] model to compute the image embedding I. The three embeddings S_t^A, Q_t, I are concatenated and passed through another fully connected layer to extract the encoder embedding e_t^A. The answer decoder, which is another LSTM, uses this embedding e_t^A to generate the answer a_t. Similar to the Q-Bot, we use a hidden layer size of 512 for all LSTMs and fully connected layers. The image feature vector coming from the CNN is 4096 dimensional (FC7 features from VGG16). The word embeddings and the encoder embeddings are 300 dimensional.

5 TRAINING

We follow the training process proposed in [3]. Two agents, a Q-Bot and an A-Bot are first trained in isolation, by supervision from the VisDial dataset. After this supervised pretraining for 15 epochs over the data, we smoothly transition the agents to learn by reinforcement via a curriculum. Specifically, for the first K rounds of dialog for each image, the agents are trained using supervision from the VisDial dataset. For the remaining 10-K rounds, however, they are trained via reinforcement learning. K starts at 9 and is linearly annealed to 0 over 10 epochs. The individual phases of training will be described in more detail below.

5.1 Supervised pre-training

In the supervised part of training, both the Q-Bot and A-Bot are trained separately. Both the Q-Bot and A-Bot are trained with a Maximum Likelihood Estimation (MLE) loss computed using the ground truth questions and answers, respectively, for every round of dialog. The Q-Bot simultaneously optimizes another objective: minimizing the Mean Squared Error (MSE) loss between the true and predicted image embeddings. The ground truth dialogs and image embeddings come from the VisDial dataset.

Given the true dialog history, image features and ground truth question, the A-Bot generates an answer to that question. Given the true dialog history and the previous question-answer pair, the Q-Bot is made to generate the next question to ask the A-Bot. However, there are multiple problems with this training scheme. First, MLE is known to result in models that generate repetitive dialogs and often produce generic responses. Second, the agents are never allowed to interact during training. Thus, when they interact during testing, they end up facing out of distribution questions and answers, and produce unexpected responses. Third, the sequentiality of dialog is lost when the agents are trained in an isolated, supervised manner.

5.2 Reinforcement Learning Setup

To alleviate the issues pointed out with supervised training, we let the two bots interact with each other via self-play (no ground-truth except images and captions). This interaction is also in the form of questions asked by the Q-Bot, and answered in turn by the A-Bot, using a shared vocabulary. The state space is partially observed and asymmetric, with the Q-Bot observing $\{c, q_1, a_1 \dots q_{10}, a_{10}\}$ and the A-Bot observing the same, plus the image I. Here, c is the caption, and q_i, a_i is the i^{th} dialog pair exchanged where $i = 1 \dots 10$. The action space for both bots consists of all possible output sequences of a specified maximum length (Q-Bot: 16, A-Bot: 9) under a fixed vocabulary (size 8645). Each action involves predicting words sequentially until a stop token is predicted, or the generated statement reaches the maximum length. Additionally, Q-Bot also produces a guess of the visual representation of the input image (VGG fc-7 embedding of size 4096). Both Q-Bot and A-Bot share the same objective and get the same reward to encourage cooperation. Information gain in each round of dialog is incentivized by setting the reward as the **change in distance** of the predicted image embedding to the ground-truth image representation. This means that a QA pair is of high quality only if it helps the Q-Bot make a better prediction of the image representation. Both policies are modeled by neural networks, as discussed in Section 4.

However, as noted above, this RL optimization problem is ill-posed, since rewarding the agents for information exchange does not motivate them to stick to the rules and conventions of the English language. Thus, we follow the elaborate curriculum scheme described above, despite which the bots are still observed to diverge from natural language and produce non-grammatical and incoherent dialog. Thus, we propose a multi bot architecture to help the agents interact in diverse and coherent, yet informative, dialog.

Learning Algorithm: A dialog round at time t consists of the following steps: 1) The Q-Bot, conditioned on the state encoding, generates a question q_t, 2) A-Bot updates its state encoding with q_t and then generates an answer a_t, 3) Both Q-Bot and A-Bot encode the completed exchange as a fact embedding, 4) Q-Bot updates its state encoding to incorporate this fact and finally 5) Q-Bot predicts the image representation for the unseen image conditioned on its updated state encoding.

Similar to Das et al. [2], we use the REINFORCE [29] algorithm that updates policy parameters in response to experienced rewards. The per-round rewards maximized are:

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

This reward is positive if the distance between image representation generated at time t is closer to the ground truth than the representation at time $t-1$, hence incentivizing information gain at each round of dialog. The REINFORCE update rule ensures that a (q_t, a_t) exchange that is informative has its probabilities pushed up. Do note that the image feature regression network f is trained directly via supervised gradient updates on the L-2 loss.

5.3 Multi-Agent Dialog Framework (MADF)

In this section we describe our proposed Multi-Agent Dialog architecture in detail. The motivation behind this is the observation that allowing a pair of agents to interact with each other and learn via reinforcement for too long leads to them developing an idiosyncratic private language which does not adhere to the rules of human language, and are hence not understandable by human observers. We claim that if, instead of allowing a single pair of agents to interact, we were to make the agents more social, and make them *interact and learn from multiple other agents*, they would be disincentivized to develop a private language, and would have to conform to the common language.

In particular, we create either multiple Q-bots to interact with a single A-bot, or multiple A-bots to interact with a single Q-bot. All these agents are initialized with the learned parameters from the supervised pretraining as described in Section 5.1. Then, for each batch of images from the VisDial dataset, we randomly choose a Q-bot to interact with the A-bot, or randomly choose an A-bot to interact with the Q-bot, as the case may be. The two chosen agents then have a complete dialog consisting of 10 question-answer pairs about each of those images, and update their respective weights based on the rewards received (as per Equation 1) during the conversation, using the REINFORCE algorithm. This process is repeated for each batch of images, over the entire VisDial dataset. It is important to note that histories are *not shared* across batches. MADF can be understood in detail using the pseudocode in Algorithm 1.

6 EXPERIMENTS AND RESULTS

6.1 Dataset description

We use the VisDial 0.9 dataset for our task introduced by Das et al. [2]. The data is collected using Amazon Mechanical Turk by pairing 2 annotators and asking them to chat about the image as a multi

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
    $\Delta(w_{Qbot}) \leftarrow \frac{1}{10} \sum_{t=1}^{10} \nabla_{\theta_{Qbot}} \left[G_t \log p(ques_t, \theta_{Qbot}) - \Delta image\_pred\right]$
    $\Delta(w_{Abot}) \leftarrow \frac{1}{10} \sum_{t=1}^{10} G_t \nabla_{\theta_{Abot}} \log p(ans_t, \theta_{Abot})$
    $w_{Qbot} \leftarrow w_{Qbot} + \Delta(w_{Qbot})$ ▷ REINFORCE and Image Loss update for Qbot
    $w_{Abot} \leftarrow w_{Abot} + \Delta(w_{Abot})$ ▷ REINFORCE update for Abot
  end while
end procedure

Table 1: Comparison of Metrics with Literature

Model	MRR	Mean Rank	R@1	R@5	R@10
Answer Prior [2]	0.3735	26.50	23.55	48.52	53.23
MN-QIH-G [2]	0.5259	17.06	42.29	62.85	68.88
HCIAE-G-DIS [16]	0.547	14.23	44.35	65.28	71.55
Frozen-Q-Multi [3]	0.437	21.13	N/A	53.67	60.48
CoAtt-GAN [30]	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

round VQA setup. One of the annotators acts as the questioner and has access to only the caption of the image and has to ask questions from the other annotator who acts as the 'answerer' and must answer the questions based on the visual information from the actual image. This dialog repeats for 10 rounds at the end of which the questioner has to guess what the image was. We perform our experiments on VisDial v0.9 (the latest available release) containing 83k dialogs on COCO-train and 40k on COCO-val images, for a total of 1.2M dialog question-answer pairs. We split the 83k into 82k for train, 1k for validation, and use the 40k as test, in a manner consistent with [2]. The caption is considered to be the first round in the dialog history.

6.2 Evaluation Metrics

We evaluate the performance of our model using 6 metrics, proposed by [3]: **1) Mean Reciprocal Rank (MRR)**, **2) Mean Rank**, **3) Recall@1**, **4) Recall@5**, **5) Recall@10** and **6) Image Retrieval Percentile**. Mean Rank and MRR compute the average rank (and its reciprocal, respectively) assigned to the ground truth answer, over a set of 100 candidate answers for each question (also averaged over all the 10 rounds). Recall@k computes the percentage of answers with rank less than k. Image Retrieval percentile is a measure of how close the image prediction generated by the Q-bot is to the ground truth. All the images in the test set are ranked according to their distance from the predicted image embedding, and the rank of the ground truth embedding is used to calculate the image retrieval percentile. All results for RL-1Q,1A, RL-1Q,3A and RL-3Q,1A are reported after 15 epochs of supervised learning and 10 epochs of

The little girl is standing with skis on her feet

Human-Human Dialogue

what color are the skis ?	A UNK color
Are there any other people?	Not that i can see
Is this outdoors?	Yes
Do you see snow?	Yes
Is it currently snowing?	No I don't think so
Is she on a slope or hill?	No i don't think so
Do you see trees?	Yes
Do you see the sky?	No
Is she wearing gloves?	Yep
Is she wearing a hat?	yes

REINFORCE (1 Q-Bot - 1 A-bot)

is there more than 1 skier ?	Yes
oh how many UNK do you see ?	A few
is the person holding a pipe of	Yes, there is a bowtie
some sort of basket ?	
is lady going fast ?	I can not tell
how many is there ?	It's not silver people are in grocery
is there snow ?	Is covered by snow covered by snow
how many animals do you see ?	Of the 1, and a 2
is there a lot of snow ?	Yes tons of snow everywhere covered
are they all together ?	Yes 1 is a darker 1
is there a lift ?	Just a few skiing out yet

Supervised

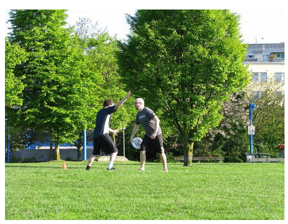
is the girl close to a race ?	no , woman is riding
'is she wearing a helmet ?	no
is she wearing a hat ?	She is wearing a headband
what color is it ?	I would say it's black
'is she wearing goggles ?	Can't see her face
'is she wearing goggles ?	Not sure
what color is her skis ?	Dark, hard to tell i would guess
is she wearing a hat ?	No
what color does she have ?	White
and what is the girl wearing ?	A white striped long sleeve shirt
	and blue

REINFORCE (1 Q-Bot - 3 A-bots)

are there any other people around ?	No
is it snowing ?	Can't tell
is it a sunny day ?	Yes
is the woman wearing a dress ?	Yes
what color is it ?	It has a white meat and girl on
is she by herself of snow ?	Just lady, woman and dog
what color is her jacket ?	Orange
does it look like a competition ?	Yes
is it day or night ?	Daytime
is it sunny ?	yes

REINFORCE (3 Q-Bots - 1 A-bot)

how old is the woman ?	About 16
is she in a ski resort ?	I don't think so
can you see the sky ?	No
is she alone ?	Yes
what color are the gloves ?	Silver
is there snow on the ground ?	No snow
is the photo very clear ?	Yes
is she wearing a helmet ?	No
is it snowing ?	it looks like it is stopped
is it sunny ?	No really



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

REINFORCE (1 Q-Bot - 1 A-bot)

do you recognize 1 ball ?	looks like a around
that you could you are seeing controllers?	that i can see the fair
do you recognize the guy and age is the	you couldn't be late teens , his
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 watching them?	Yes there is a people

Supervised

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

REINFORCE (3 Q-Bots - 1 A-bot)

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

REINFORCE (1 Q-Bot - 3 A-bots)

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

Figure 3: Randomly selected examples of dialog between Q-Bot A-Bot under different settings. These images were also used in the human evaluation results shown in Table 2

Table 2: Human Evaluation Results - Mean Rank (Lower is better)

	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**

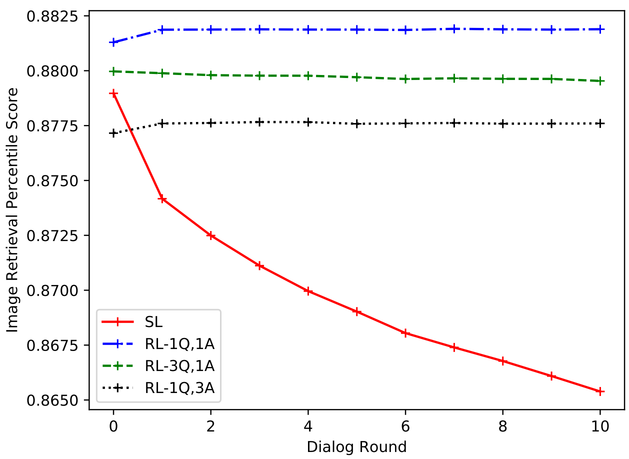


Figure 4: The percentile scores of the ground truth image compared to the entire test set of 40k images. The X-axis denotes the dialog round number (from 1 to 10), while the Y-axis denotes the image retrieval percentile score.

curriculum learning as described in Section 5. Consequently, the training time of all 3 systems are equal.

Table 1 compares the Mean Rank, MRR, Recall@1, Recall@5 and Recall@10 of our agent architecture and dialog framework (below the horizontal line) with previously proposed architectures (above the line). SL refers to the agents after only the isolated, supervised training of Section 5.1. RL-1Q,1A refers to a single, idiosyncratic pair of agents trained via reinforcement as in Section 5.2. RL-1Q,3A and RL-3Q,1A refer to social agents trained via our Multi-Agent Dialog framework in Section 5.3, with 1Q,3A referring to 1 Q-Bot and 3 A-Bots, and 3Q,1A referring to 3 Q-Bots and 1 A-Bot.

It can be seen that our agent architectures clearly outperform all previously proposed generative architectures in MRR, Mean Rank and R@10, but not in R@1 and R@5. This indicates that our approach produces consistently good answers (as measured by MRR, Mean Rank and R@10), even though they might not be the best possible answers (as measured by R@1 and R@5). SL has the best task performance, which drops drastically in RL-1Q,1A. The agents trained by MADF recover task performance compared to SL, hence outperforming all previously proposed models in the literature - while vastly improving dialog coherence (see Section 6.3). Fig. 4 shows the change in image retrieval percentile scores over dialog rounds. The percentile score decreases monotonically for SL, but is stable for all versions using RL. The decrease in image retrieval score over dialog rounds for SL is because the test set questions and answers are not perfectly in-distribution (compared to the training set), and the SL system can't adapt to these samples as well as the systems trained with RL. As the dialog rounds increase, the out-of-distribution nature of dialog exchange increases, hence leading to a decrease in SL scores. Interestingly, despite having strictly more information in later rounds, the scores of RL agents do not increase - which we think is because of the nature of LSTMs to forget.

6.3 Human Evaluation

There are no quantitative metrics to comprehensively evaluate dialog quality, hence we do a human evaluation of the generated dialog. There are 5 metrics we evaluate on: 1) Q-Bot Relevance, 2) Q-Bot Grammar, 3)A-Bot Relevance, 4) A-Bot Grammar and 5) Overall Dialog Coherence. We evaluate 4 Visual Dialog systems, trained via: 1) **Supervised Learning (SL)**, 2) **Reinforce for 1 Q-Bot, 1 A-Bot (RL-1Q,1A)**, 3) **Reinforce for 1 Q-Bot, 3 A-Bots (RL-1Q,3A)** and 4) **Reinforce for 3 Q-Bots, 1 A-Bot (RL-3Q,1A)**. A total of 20 evaluators (randomly chosen students at CMU) were shown the caption and the 10 QA-pairs generated by each system for one of 4 randomly chosen images, and asked to give an ordinal ranking (from 1 to 4) for each metric. If the evaluator was also given access to the image, she was asked only to evaluate metrics 3, 4 and 5 above, while if the evaluator was not shown the image, she asked only to evaluate metrics 1, 2 and 5. Table 2 contains the average ranks obtained on each metric (lower is better).

The results convincingly **prove our hypothesis that having multiple A-Bots to interact and learn from will improve the Q-Bot relevance, and vice versa. The grammar for both bots improves in both MADF settings**. This is because having multiple A-Bots to interact with gives the Q-Bot access to a variety of diverse dialog, leading to more stable updates with lower bias. The results confirm this, with Q-Bot Relevance rank being lowest in RL-1Q,3A, and A-Bot Relevance rank being lowest in RL-3Q,1A - meaning that RL-1Q,3A produces better questions while RL-3Q,1A produces better answers. These two dialog systems, which were trained via MADF, also have the best overall dialog coherence by a significant margin over RL-1Q,1A and SL. We show two of the examples shown to the human evaluators in Figure 3. The trends observed in the scores given by human evaluators is also clearly visible in these examples. MADF agents are able to model the human responses much better than the other agents. It can also be seen that although the RL-1Q,1A system has greater diversity in its responses, the quality of those responses is greatly degraded, with the A-Bot's answers especially being both non-grammatical and irrelevant. In Section 5.1, we discussed how the MSE loss used in SL results in models which generate repetitive dialog, which can be seen in Fig. 3. Consider the first image, where in the SL QA-generations, the Q-Bot repeats the same questions multiple times, and gets inconsistent answers from the A-Bot for the same question. By contrast, all 10 QA-generations for RL-3Q,1A are grammatically correct. The Q-Bot's questions are very relevant to the image being considered, and the A-Bot's answers appropriate and correct.

7 DISCUSSION AND CONCLUSION

In this paper we propose a novel Multi-Agent Dialog Framework (MADF), inspired from human communities, to improve the dialog quality of AI agents. We show that training 2 agents with supervised learning can lead to uninformative and repetitive dialog. Furthermore, we observe that the task performance (measured by the image retrieval percentile scores) for the system trained via supervision only deteriorates as dialog round number increases. We hypothesize that this is because the agents were trained in isolation and never allowed to interact during supervised learning, which leads to failure during testing when they encounter out of distribution samples (generated by the other agent, instead of ground truth) for the first time. We show how allowing a single pair of agents to interact and learn from each other via reinforcement learning dramatically improve their percentile scores, which additionally does not deteriorate over multiple rounds of dialog, since the agents have interacted with one another and been exposed to the other's generated questions or answers. However, the agents, in an attempt to improve task performance end up developing their own private language which does not adhere to the rules and conventions of human languages, and generates non-grammatical and non-sensical statements. As a result, the dialog system loses interpretability and sociability. To alleviate this issue, we propose a multi-agent dialog framework based on self-play reinforcement learning, where a single A-Bot is allowed to interact with multiple Q-Bots and vice versa. Through a human evaluation study, we show that this leads to significant improvements in dialog quality measured by relevance, grammar and coherence, without compromising task performance. This is because interacting with multiple agents prevents any particular pair from maximizing performance by developing a private language, since it would harm performance with all the other agents.

8 FUTURE WORK

There are several possible extensions to this 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 an explicit perplexity based reward term 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. Another avenue for future exploration is to use a richer image feature embedding to regress on. Currently, we use a regression network to compute the estimated image embedding which represents the Q-Bot's understanding of the image. This, however, does not allow us to visualize the image to see what the Q-bot is actually able to see. We plan to implement an image generation GAN which can use this embedding as a latent code to generate an image which can be visualized. While the MADF in its current form only works if we have multiple Q-Bots or multiple A-Bots but not both, future work could possibly look at incorporating that into the framework, while ensuring that the updates do not become too unstable.

Acknowledgments. This research was sponsored in part by AFOSR Grant FA9550-15-1-0442. We would like to thank William Guss, Abhishek Das, Satwik Kottur and Fei Fang for their insightful and fruitful discussions and feedback.

REFERENCES

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: Visual Question Answering. *Int. J. Comput. Vision* 123, 1 (May 2017), 4–31. https://doi.org/10.1007/s11263-016-0966-6
[2] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual Dialog. *CoRR* abs/1611.08669 (2016). arXiv:1611.08669 http://arxiv.org/abs/1611.08669
[3] Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. *CoRR* abs/1703.06585 (2017). arXiv:1703.06585 http://arxiv.org/abs/1703.06585
[4] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2016. GuessWhat?! Visual object discovery through multi-modal dialogue. *CoRR* abs/1611.08481 (2016). arXiv:1611.08481 http://arxiv.org/abs/1611.08481
[5] Mihail Eric and Christopher D. Manning. 2017. A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented Dialogue. *CoRR* abs/1701.04024 (2017). http://arxiv.org/abs/1701.04024
[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385
[8] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. 2018. Adversarial Learning for Semi-Supervised Semantic Segmentation. *CoRR* abs/1802.07934 (2018).
[9] Justin Johnson, Andrej Karpathy, and Fei-Fei Li. 2015. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. *CoRR* abs/1511.07571 (2015). arXiv:1511.07571 http://arxiv.org/abs/1511.07571
[10] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *CoRR* abs/1411.2539 (2014). arXiv:1411.2539 http://arxiv.org/abs/1411.2539
[11] Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. *CoRR* abs/1706.08502 (2017). arXiv:1706.08502 http://arxiv.org/abs/1706.08502
[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*. Curran Associates Inc., USA, 1097–1105. http://dl.acm.org/citation.cfm?id=2999134.2999257
[13] Yaniv Leviathan. 2018. Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. (May 2018). https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html
[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2015. SSD: Single Shot MultiBox Detector. *CoRR* abs/1512.02325 (2015). http://arxiv.org/abs/1512.02325
[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation. *CVPR* (2015).
[16] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model. *CoRR* abs/1706.01554 (2017). arXiv:1706.01554 http://arxiv.org/abs/1706.01554
[17] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. *CoRR* abs/1612.01887 (2016). arXiv:1612.01887 http://arxiv.org/abs/1612.01887

[18] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. *CoRR* abs/1312.5602 (2013). http://arxiv.org/abs/1312.5602

[19] Geraldin Nanfack, Azeddine Elhassouny, and Rachid Oulad Haj Thami. [n. d.]. Squeeze-SegNet: A new fast Deep Convolutional Neural Network for Semantic Segmentation. *CoRR* abs/1711.05491 ([n. d.]).

[20] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2015. You Only Look Once: Unified, Real-Time Object Detection. *CoRR* abs/1506.02640 (2015). http://arxiv.org/abs/1506.02640

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)*. MIT Press, Cambridge, MA, USA, 91–99. http://dl.acm.org/citation.cfm?id=2969239.2969250

[22] Michael Rutter, Karen Thorpe, Rosemary Greenwood, Kate Northstone, and Jean Golding. 2003. Twins as a natural experiment to study the causes of mild language delay: I: Design; twin–singleton differences in language, and obstetric risks. *Journal of Child Psychology and Psychiatry* 44, 3 (2003), 326–341.

[23] Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of Go without human knowledge. *Nature* 550 (2017), 354–359. https://doi.org/10.1038/nature24270

[24] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional Attention Flow for Machine Comprehension. *CoRR* abs/1611.01603 (2016). http://arxiv.org/abs/1611.01603

[25] Iulian Vlad Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Mudumba, Alexandre de Brébisson, Jose Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2017. A Deep Reinforcement Learning Chatbot. *CoRR* abs/1709.02349 (2017). http://arxiv.org/abs/1709.02349

[26] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[27] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and Tell: A Neural Image Caption Generator. *CoRR* abs/1411.4555 (2014). arXiv:1411.4555 http://arxiv.org/abs/1411.4555

[28] Gellért Weisz, Pawel Budzianowski, Pei-Hao Su, and Milica Gasic. 2018. Sample Efficient Deep Reinforcement Learning for Dialogue Systems with Large Action Spaces. *CoRR* abs/1802.03753 (2018). http://arxiv.org/abs/1802.03753

[29] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*. Springer, 5–32.

[30] Qi Wu, Peng Wang, Chunhua Shen, Ian D. Reid, and Anton van den Hengel. 2017. Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. *CoRR* abs/1711.07613 (2017). arXiv:1711.07613 http://arxiv.org/abs/1711.07613

[31] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated Residual Transformations for Deep Neural Networks. *CoRR* abs/1611.05431 (2016). http://arxiv.org/abs/1611.05431

[32] Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic Coattention Networks For Question Answering. *CoRR* abs/1611.01604 (2016). http://arxiv.org/abs/1611.01604

[33] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *CoRR* abs/1502.03044 (2015). arXiv:1502.03044 http://arxiv.org/abs/1502.03044

[34] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2016. Boosting Image Captioning with Attributes. *CoRR* abs/1611.01646 (2016). arXiv:1611.01646 http://arxiv.org/abs/1611.01646

[35] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. *International Conference on Learning Representations, ICLR-2018* (2018).