Proceedings of the Fifth Conference on
Human Computation and Crowdsourcing
(HCOMP 2017)

Evaluating Visual Conversational Agents via Cooperative Human-AI Games

Prithvijit Chattopadhyay,[1,2,4] **Deshraj Yadav,**[1,2,4] **Viraj Prabhu,**[2,4] **Arjun Chandrasekaran,**[2]
Abhishek Das,[2] **Stefan Lee,**[2,4] **Dhruv Batra,**[3,2] **Devi Parikh**[,3,2]
[2]Georgia Institute of Technology
[3]Facebook AI Research
{prithvijit3, deshraj, parikh}@gatech.edu
visualdialog.org

Abstract

As AI continues to advance, human-AI teams are inevitable. However, progress in AI is routinely measured in isolation, without a human in the loop. It is crucial to benchmark progress in AI, not just in isolation, but also in terms of how it translates to helping humans perform certain tasks, *i.e.*, the performance of human-AI teams.

In this work, we design a cooperative game – GuessWhich – to measure human-AI team performance in the specific context of the AI being a visual conversational agent. GuessWhich involves live interaction between the human and the AI. The AI, which we call ALICE, is provided an image which is unseen by the human. Following a brief description of the image, the human questions ALICE about this secret image to identify it from a fixed pool of images.

We measure performance of the human-ALICE team by the number of guesses it takes the human to correctly identify the secret image after a fixed number of dialog rounds with ALICE. We compare performance of the human-ALICE teams for two versions of ALICE. Our human studies suggest a counter-intuitive trend – that while AI literature shows that one version outperforms the other when paired with an AI questioner bot, we find that this improvement in AI-AI performance does not translate to improved human-AI performance. This suggests a mismatch between benchmarking of AI in isolation and in the context of human-AI teams.

1 Introduction

As Artificial Intelligence (AI) systems become increasingly accurate and interactive (*e.g.* Alexa, Siri, Cortana, Google Assistant), human-AI teams are inevitably going to become more commonplace. To be an effective teammate, an AI must overcome the challenges involved with adapting to humans; however, progress in AI is routinely measured in isolation, without a human in the loop. In this work, we focus specifically on the evaluation of visual conversational agents and develop a human computation game to benchmark their performance as members of human-AI teams.

Visual conversational agents (Das et al. 2017a; 2017b; de Vries et al. 2017; Strub et al. 2017) are AI agents trained to understand and communicate about the contents of a scene in natural language. For example, in Fig. 1, the visual conversational agent (shown on the right) replies to answers questions about a scene while inferring context from the dialog history – Human: "What is he doing?" Agent: "Playing

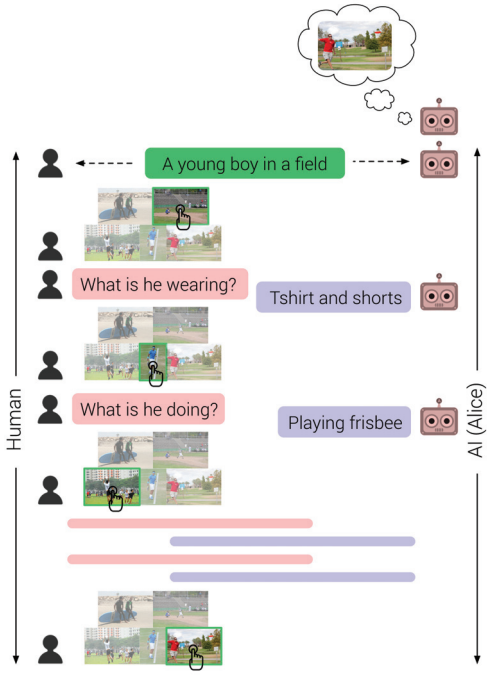


Figure 1: A human and an AI (a visual conversation agent called ALICE) play the proposed GuessWhich game. At the start of the game (top), ALICE is provided an image (shown above ALICE) which is unknown to the human. Both ALICE and the human are then provided a brief description of the image. The human then attempts to identify the secret image. In each subsequent round of dialog, the human asks a question about the unknown image, receives an answer from ALICE, and makes a best guess of the secret image from a fixed pool of images. After 9 rounds of dialog, the human makes consecutive guesses until the secret image is identified. The fewer guesses the human needs to identify the secret image, the better the human-AI team performance.

[1]The first two authors (PC, DY) contributed equally.

[4]Work done at Virginia Tech.

frisbee". These agents are typically trained to mimic large corpora of human-human dialogs and are evaluated automatically on how well they retrieve actual human responses (ground truth) in novel dialogs.

Recent work has evaluated these models more pragmatically by evaluating how well pairs of visual conversational agents perform on goal-based conversational tasks rather than response retrieval from fixed dialogs. Specifically, (Das et al. 2017b) train two visual conversational agents – a questioning bot QBOT, and an answering bot ABOT – for an image-guessing task. Starting from a description of the scene, QBOT and ABOT converse over multiple rounds of questions (QBOT) and answers (ABOT) in order to improve QBOT's understanding of a secret image known only to ABOT. After a fixed number of rounds, QBOT must guess the secret image from a large pool and both QBOT and ABOT are evaluated based on this guess.

(Das et al. 2017b) compare supervised baseline models with QBOT-ABOT teams trained through reinforcement learning based self-talk on this image-guessing task. They find that the AI-AI teams improve significantly at guessing the correct image after self-talk updates compared to the supervised pretraining. While these results indicate that the self-talk fine-tuned agents are better visual conversational agents, crucially, it remains unclear if these agents are indeed better at this task when *interacting with humans*.

GuessWhich. In this work, we propose to evaluate if and how this progress in AI-AI evaluation translates to the performance of human-AI teams. Inspired by the popular GuessWhat or 20-Questions game, we design a human computation game – GuessWhich – which requires collaboration between human and visual conversational AI agents. Mirroring the setting of (Das et al. 2017b), GuessWhich is an image-guessing game that consists of 2 participants – *questioner* and *answerer*. At the start of the game, the answerer is provided an image that is unknown to the questioner and both questioner and answerer are given a brief description of the image content. The questioner interacts with the answerer for a fixed number of rounds of question-answer (dialog) to identify the secret image from a fixed pool of images (see Fig. 1).

We evaluate human-AI team performance in GuessWhich, for the setting where the questioner is a human and the answerer is an AI (that we denote ALICE). Specifically, we evaluate two versions of ALICE for GuessWhich:

1. ALICE_{SL} which is trained in a supervised manner on the Visual Dialog dataset (Das et al. 2017a) to mimic the answers given by humans when engaged in a conversation with other humans about an image, and
2. ALICE_{RL} which is pre-trained with supervised learning and fine-tuned via reinforcement learning for an image-guessing task as in (Das et al. 2017b).

It is important to appreciate the difficulty and sensitivity of the GuessWhich game as an evaluation tool – agents have to understand human questions and respond with accurate, consistent, fluent and informative answers for the human-AI team to do well. Furthermore, they have to be robust to their own mistakes, *i.e.*, if an agent makes an error at a particular round, that error is now part of its conversation history, and it must be able to correct itself rather than be consistently inaccurate. Similarly, human players must also learn to adapt to ALICE's sometime noisy and inaccurate responses.

At its core, GuessWhich is a game-with-a-purpose (GWAP) that leverages human computation to evaluate visual conversational agents. Traditionally, GWAP (Von Ahn and Dabbish 2008) have focused on *human-human collaboration*, *i.e.* collecting data by making humans play games to label images (Von Ahn and Dabbish 2004), music (Law et al. 2007) and movies (Michelucci 2013). We extend this to human-AI teams and to the best of our knowledge, our work is the first to evaluate visual conversational agents in an interactive setting where humans are continuously engaging with agents to succeed at a cooperative game.

Contributions. More concretely, we make the following contributions in this work:

- We design an interactive image-guessing game (GuessWhich) for evaluating human-AI team performance in the specific context of the AIs being visual conversational agents. GuessWhich pairs humans with ALICE, an AI capable of answering a sequence of questions about images. ALICE is assigned a secret image and answers questions asked about that image from a human for 9 rounds to help them identify the secret image (Sec. 4).
- We evaluate human-AI team performance on this game for both supervised learning (SL) and reinforcement learning (RL) versions of ALICE. Our main experimental finding is that despite significant differences between SL and RL agents reported in previous work (Das et al. 2017b), we find *no significant difference* in performance between ALICE_{SL} or ALICE_{RL} when paired with human partners (Sec. 6.1). This suggests that while self-talk and RL are interesting directions to pursue for building better visual conversational agents, there appears to be a disconnect between AI-AI and human-AI evaluations – progress on former does not seem predictive of progress on latter. This is an important finding to guide future research.

2 Related Work

Given that our goal is to evaluate visual conversational agents through a human computation game, we draw connections to relevant work on visual conversational agents, human computation games, and dialog evaluation below.

Visual Conversational Agents. Our AI agents are visual conversational models, which have recently emerged as a popular research area in visually-grounded language modeling (Das et al. 2017a; 2017b; de Vries et al. 2017; Strub et al. 2017). (Das et al. 2017a) introduced the task of Visual Dialog and collected the VisDial dataset by pairing subjects on Amazon Mechanical Turk (AMT) to chat about an image (with assigned roles of questioner and answerer). (Das et al. 2017b) pre-trained questioner and answerer agents on this VisDial dataset via supervised learning and fine-tuned them via self-talk (reinforcement learning), observing that RL-fine-tuned QBOT-ABOT are better at image-guessing after interacting with each other. However,

as described in Section 1, they do not evaluate if this change in QBOT-ABOT performance translates to human-AI teams.

Human Computation Games. Human computation games have been shown to be time- and cost-efficient, reliable, intrinsically engaging for participants (Jain and Parkes 2013; Krause and Smeddinck 2011), and hence an effective method to collect data annotations. There is a long line of work on designing such Games with a Purpose (GWAP) (Von Ahn and Dabbish 2008) for data labeling purposes across various domains including images (Von Ahn and Dabbish 2004; Von Ahn, Liu, and Blum 2006; Law and Von Ahn 2009; Kazemzadeh et al. 2014), audio (Diakopoulos, Luther, and Essa 2008; Law et al. 2007), language (Aras et al. 2010; Chamberlain, Poesio, and Kruschwitz 2008), movies (Michelucci 2013) *etc*. While such games have traditionally focused on human-human collaboration, we extend these ideas to human-AI teams. Rather than collecting labeled data, our game is designed to measure the effectiveness of the AI in the context of human-AI teams.

Evaluating Conversational Agents. Goal-driven (non-visual) conversational models have typically been evaluated on task-completion rate or time-to-task-completion (Paek 2001), so shorter conversations are better. At the other end of the spectrum, free-form conversation models are often evaluated by metrics that rely on n-gram overlaps, such as BLEU, METEOR, ROUGE, but these have been shown to correlate poorly with human judgment (Liu et al. 2016). Human evaluation of conversations is typically in the format where humans rate the quality of machine utterances given context, without actually taking part in the conversation, as in (Das et al. 2017b) and (Li et al. 2016). To the best of our knowledge, we are the first to evaluate conversational models via team performance where humans are continuously interacting with agents to succeed at a downstream task.

Turing Test. Finally, our GuessWhich game is in line with ideas in (Grosz 2012), re-imagining the traditional Turing Test for state-of-the-art AI systems, taking the pragmatic view that an effective AI teammate need not appear human-like, act or be mistaken for one, provided its behavior does not feel jarring or baffle teammates, leaving them wondering not about what it is thinking but whether it is.

Next, we formally define the AI agent ALICE (Sec. 3), describe the GuessWhich game setup (Sec. 4 and 5), and present results and analysis from human studies (Sec. 6).

3 The AI: ALICE

Recall from Section 1 that our goal is to evaluate how progress in AI measured through automatic evaluation translates to performance of human-AI teams in the context of visual conversational agents. Specifically, we are considering the question-answering agent ABOT from (Das et al. 2017b) as ABOT is the agent more likely to be deployed with a human partner in real applications (*e.g*. to answer questions about visual content to aid a visually impaired user). For completeness, we will review this work in this section.

(Das et al. 2017b) formulate a self-supervised image-guessing task between a questioner bot (QBOT) and an answerer bot (ABOT) which plays out over multiple rounds of dialog. At the start of the task, QBOT and ABOT are shown a one sentence description (*i.e*. a caption) of an image (unknown to QBOT). The pair can then engage in question and answer based dialog for a fixed number of iterations after which QBOT must try to select the secret image from a pool. The goal of the QBOT-ABOT team is two-fold, QBOT should: 1) build a mental model of the unseen image purely from the dialog and 2) be able to retrieve that image from a line-up of images.

Both QBOT and ABOT are modeled as Hierarchical Recurrent Encoder-Decoder neural networks (Das et al. 2017a; Serban et al. 2016) which encode each round of dialog independently via a recurrent neural network (RNN) before accumulating this information through time with an additional RNN (resulting in hierarchical encoding). This representation (and a convolutional neural network based image encoding in ABOT's case) are used as input to a decoder RNN which produces an agent's utterance (question for QBOT and answer for ABOT) based on the dialog (and image for ABOT). In addition, QBOT includes an image feature regression network that predicts a representation of the secret image based on dialog history. We refer to (Das et al. 2017b) for complete model details.

These agents are pre-trained with supervised dialog data from the VisDial dataset (Das et al. 2017a) with a Maximum Likelihood Estimation objective. This pre-training ensures that agents can generally recognize objects/scenes and utter English. Following this, the models are fine-tuned by 'smoothly' transitioning to a deep reinforcement learning framework to directly improve image-guessing performance. This annealed transition avoids abrupt divergence of the dialog in face of an incorrect question-answer pair in the QBOT-ABOT exchange. During RL based self-talk, the agents' parameters are updated by gradients corresponding to rewards depending on individual good or bad exchanges. We refer to the baseline supervised learning based ABOT as ALICE_{SL} and the RL fine-tuned bot as ALICE_{RL}. (Das et al. 2017b) found that the AI-AI pair succeeds in retrieving the correct image more often after being fine-tuned with RL. In the following section, we outline our GuessWhich game designed to evaluate whether this improvement between ALICE_{SL} and ALICE_{RL} in automatic metrics translates to human-AI collaborations.

4 Our GuessWhich Game

We begin by describing our game setting; outlining the players and gameplay mechanics. A video of an example game being played can be found at `https://vimeo.com/229488160`.

Players. We replace QBOT in the AI-AI dialog with humans to perform a collaborative task of identifying a secret image from a pool. In the following, we will refer to ABOT as ALICE and the human player as H. We evaluate two versions of ALICE – ALICE_{SL} and ALICE_{RL}, where SL and RL correspond to agents *trained in a supervised setting* and *fine-tuned with reinforcement learning* respectively.

Gameplay. In our game setting, ALICE is assigned a secret image I^c (unknown to H) from a pool of images $\mathbb{I} = \{I_1, I_2, ..., I_n\}$ taken from the COCO dataset (Lin et al. 2014). Prior to beginning the dialog, both ALICE and H are

Figure 2: GuessWhich Interface: A user asks a question to ALICE in each round and ALICE responds with an answer. The user then selects an appropriate image which they think is the secret image after each round of conversation. At the end of the dialog, user successively clicks on their best guesses until they correctly identify the secret image.

provided a brief description (*i.e.* a caption) of I^c generated by Neuraltalk2 (Karpathy 2016), an open-source implementation of (Vinyals et al. 2015). H then makes a guess about the secret image by selecting one from the pool $\mathbb{I}$ based only on the caption, *i.e.* before the dialog begins.

In each of the following rounds, H asks ALICE a question q_t about the secret image I^c in order to better identify it from the pool and ALICE responds with an answer a_t. After each round, H must select an image I^t that they feel is most likely the secret image I^c from pool $\mathbb{I}$ based on the dialog so far. At the end of $k = 9$ rounds of dialog, H is asked to successively click on their best guess. At each click, the interface gives H feedback on whether their guess is correct or not and this continues until H guesses the true secret image. In this way, H induces a partial ranking of the pool up to the secret image based on their mental model of I^c from the dialog.

4.1 Pool Selection

When creating a pool of images, our aim is to ensure that the game is challenging and engaging, and not too easy or too hard. Thus, we construct each pool of images $\mathbb{I}$ in two steps – first, we choose the secret image I^c, and then sample similar images as distractors for I^c. Fig. 2 shows a screenshot of our game interface including a sample image pool and chat.

Secret Image Selection. VisDial v0.5 is constructed on $68k$ COCO images which contain complex everyday scenes with 80 object categories. ABOT is trained and validated on VisDial v0.5 *train* and *val* splits respectively. As the images for both these splits come from COCO-train, we sample secret images and pools from COCO-validation to avoid overlap.

To select representative secret images and diverse image pools, we do the following. For each image in the COCO validation set, we extract the penultimate layer ('fc7') activations of a standard deep convolutational neural network (VGG-19 from (Simonyan and Zisserman 2015)). For each of the 80 categories, we average the embedding vector of all images containing that category. We then pick those images closest to the mean embeddings, yielding 80 candidates.

Generating Distractor Images. The distractor images are designed to be semantically similar to the secret image I^c. For each candidate secret image, we created 3 concentric hyper-spheres as euclidean balls (of radii increasing in arithmetic progression) centered on the candidate secret image in fc7 embedding space, and sampled images from each sphere in a fixed proportion to generate a pool corresponding to the secret image. The radius of the largest sphere was varied and manually validated to ensure pool difficulty. The sampling proportion can be varied to generate pools of varying difficulty. Of the 80 candidate pools, we picked 10 that were of medium difficulty based on manual inspection.

4.2 Data Collection and Player Reward Structure

We use AMT to solicit human players for our game. Each Human Intelligence Task (HIT) consists of 10 games (each game corresponds to one pool) and we find that overall 76.7% of users who started a HIT completed it *i.e.* played all 10 games. We note that incomplete game data was dis-

carded and does not contribute to the analysis presented in subsequent sections.

We published HITs until 28 games with both ALICE$_{SL}$ and ALICE$_{RL}$ were completed. This results in a total of 560 games split between the agents, with each game consisting of 9 rounds of dialog and 10 rounds of guessing. Workers are paid a base pay of \$5 per HIT (∼\$10/hour).

To incentivize workers to try their best at guessing the secret image, workers are paid a two-part bonus – (1) based on the number of times their best guess matched the true secret image after each round (up to \$1 per HIT), and (2) based on the rank of the true secret image in their final sorting at the end of dialog (up to \$2 per HIT).

This final ranking explicitly captures the workers' mental model of the secret image (unlike the per-round, best-guess estimates), and is closer to the overall purpose of the game (identifying the secret image at the end of the dialog). As such, this final sorting is given a higher potential bonus.

4.3 Evaluation

Since the game is structured as a retrieval task, we evaluate the human-AI collaborative performance using standard retrieval metrics. Note that the successive selection of images by H at the end of the dialog tells us the rank of the true secret image in a sorting of the image pool based on H's mental model. For example, if H makes 4 guesses before correctly selecting the secret image, then H's mental model ranked the secret image 5th within the pool.

To evaluate human-AI collaboration, we use the following metrics: (1) Mean Rank (MR), which is the mean rank of the secret image (*i.e.* number of guesses it takes to identify the secret image). Lower values indicate better performance. (2) Mean Reciprocal Rank (MRR), which is the mean of the reciprocal of the rank of the secret image. MRR penalizes differences in lower ranks (e.g., between 1 and 2) greater than those in higher ranks (e.g., between 19 and 20). Higher values indicate better performance.

At the end of each round, H makes their best guess of the secret image. To get a coarse estimate of the rank of the secret image in each round, we sort the image pool based on distance in fc7 embedding space from H's best guess. This can be used to assess accuracy of H's mental model of the secret image after each round of dialog (e.g., Fig. 4b).

5 Infrastructure

We briefly outline the backend architecture of GuessWhich in this section. Unlike most human-labeling tasks that are one-way and static in nature (i.e., only involving a human labeling static data), evaluating AI agents via our game requires live interaction between the AI agent and the human. We develop a robust workflow that can maintain a queue of workers and pair them up in real-time with an AI agent.

We deploy ALICE$_{SL}$ and ALICE$_{RL}$ on an AWS EC2 (AWS 2017) GPU instance. We use Django (a Model-View-Controller web framework written in Python) which helps in monitoring HITs in real-time. We use (RabbitMQ 2017), an open source message broker, to queue inference jobs that generate dialog responses from the model. Our backend is asynchronously connected to the client browser via websockets such that whenever an inference job is completed, a websocket polls the AI response and delivers it to the human in real-time. We store and fetch data efficiently to and from a PostgreSQL database. Fig. 3 shows a schematic diagram of the backend architecture. Our complete backend infrastructure and code is publicly available on github.com/VT-vision-lab/GuessWhich for others to easily make use of our human-AI game interface.

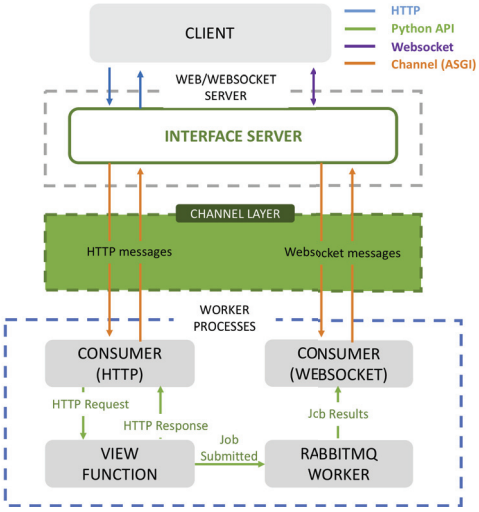


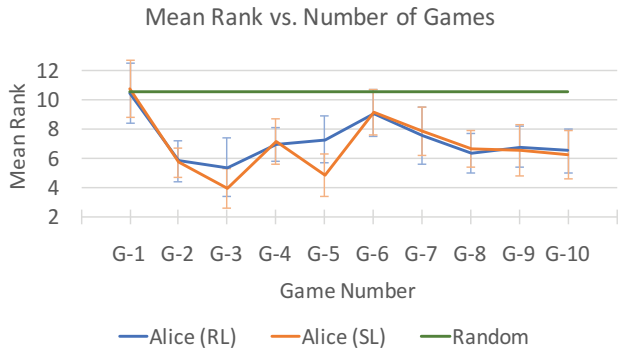
Figure 3: We outline the backend architecture of our implementation of GuessWhich. Since GuessWhich requires a live interaction between the human and the AI, we design a workflow that can handle multiple queues and can quickly pair a human with an AI agent.

6 Results

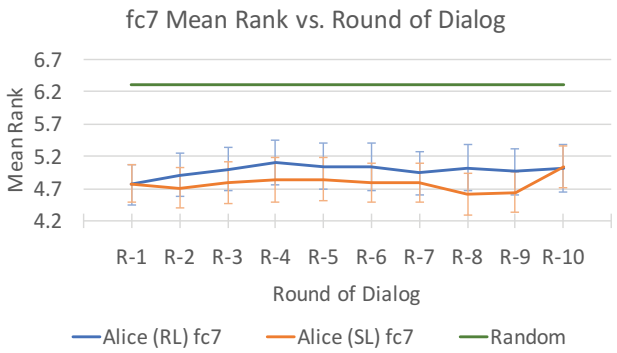
6.1 ALICE$_{SL}$ vs. ALICE$_{RL}$

We compare the performance of the two agents ALICE$_{SL}$ and ALICE$_{RL}$ in the GuessWhich game. These bots are state-of-the-art visual dialog agents with respect to emulating human responses and generating visually discriminative responses in AI-AI dialog. (Das et al. 2017b) evaluate these agents against strong baselines and report AI-AI team results that are significantly better than chance on a pool of ∼10k images (rank ∼1000 for SL, rank ∼500 for RL). In addition to evaluating them in the context of human-AI teams we also report QBOT-ALICE team performances for reference.

In Table 1, we compare the performances of human-ALICE$_{SL}$ and human-ALICE$_{RL}$ teams according to Mean Rank (MR) and Mean Reciprocal Rank (MRR) of the secret image based on the guesses H makes at the end of dialog. We observe that at the end of each game (9 rounds of dialog), human subjects correctly guessed the secret image on their 6.86th attempt (Mean Rank) when ALICE$_{SL}$ was their teammate. With ALICE$_{RL}$ as their teammate, the average number of guesses required was 7.19. We also observe that ALICE$_{RL}$ outperforms ALICE$_{SL}$ on the MRR metric. On



(a) ALICE$_{SL}$ and ALICE$_{RL}$ perform about the same for most games and outperform a baseline model that makes a string of random guesses at the end of each game.



(b) ALICE$_{SL}$ and ALICE$_{RL}$ perform about the same, and clearly outperform a baseline model that randomly chooses an image. As described in Sec. 4.3, this is only a coarse estimate of the rank of the secret image after each round of dialog.

Figure 4: Mean rank (MR) of secret image across (a) number of games and (b) rounds of dialog. Lower is better. Error bars are 95% confidence intervals from 1000 bootstrap samples.

Team	MR	MRR
Human-ALICE$_{SL}$	6.86 ± 0.53	0.27 ± 0.03
Human-ALICE$_{RL}$	7.19 ± 0.55	0.25 ± 0.03

Table 1: Performance of Human-ALICE teams with ALICE$_{SL}$ and ALICE$_{RL}$ measured by MR (lower is better) and MRR (higher is better). Error bars are 95% CIs from 1000 bootstrap samples. Unlike (Das et al., 2017b), we find no significant difference between ALICE$_{SL}$ and ALICE$_{RL}$.

Team	ALICE$_{SL}$	ALICE$_{RL}$
Human	6.9	7.2
QBOT (SL)	5.6	5.3
QBOT (RL)	4.7	4.7

Table 2: Performance of Human-ALICE and QBOT-ALICE teams measured by MR (lower is better). Error bars are 95% confidence intervals from 1000 bootstrap samples. We observe that AI-AI teams outperform human-AI teams.

both metrics, however, the differences are within the standard error margins (reported in the table) and not statistically significant. As we collected additional data, the error margins became smaller but the means also became closer. This interesting finding stands in stark contrast to the results reported by (Das et al. 2017b), where ALICE$_{RL}$ was found to be significantly more accurate than ALICE$_{SL}$ when evaluated in an AI-AI team. Our results suggest that the improvements of RL over SL (in AI-AI teams) do not seem to translate to when the agents are paired with a human in a similar setting.

MR with varying number of games. In Fig. 4a, we plot the mean rank (MR) of the secret image across different games. We see that the human-ALICE team performs about the same for both ALICE$_{SL}$ and ALICE$_{RL}$ except Game 5, where ALICE$_{SL}$ seems to marginally outperform ALICE$_{RL}$. We compare the performance of these teams against a baseline model that makes a string of random guesses at the end of the game. The human-ALICE teams outperforms this random baseline with a relative improvement of about 25%.

AI-ALICE teams versus human-ALICE teams. In Table 2, we compare team performances by pairing three kinds of questioners – human, QBOT (SL) and QBOT (RL) with ALICE$_{SL}$ and ALICE$_{RL}$ (6 teams in total) to gain insights about how the questioner and ALICE influence team performances. Interestingly, we observe that AI-ALICE teams outperform human-ALICE teams. On average, a QBOT (SL)-ALICE$_{SL}$ team takes about 5.6 guesses to arrive at the correct secret image (as opposed to 6.86 guesses for a human-ALICE$_{SL}$ team). Similarly, a QBOT (RL)-ALICE$_{RL}$ team takes 4.7 guesses as opposed to a human-ALICE$_{RL}$ team which takes 7.19 guesses. When we compare AI-AI teams (see Row 2 and 3) under different settings, we observe that teams having QBOT (RL) as the questioner outperform those with QBOT (SL). Qualitatively, we found that QBOT (SL) tends to ask repeating questions in a dialog and that questions from QBOT (RL) tend to be more visually grounded compared to QBOT (SL). Also, note that among the four teams ALICE does not seem to affect performance across SL and RL.

Since we observe that QBOT (RL) tends to be a better questioner on average compared to QBOT (SL), as future work, it will be interesting to explore a setting where we evaluate QBOT via a similar game with the human playing the role of answerer in a QBOT-human team.

MR with varying rounds of dialog. Fig. 4b shows a coarse estimate of the mean rank of the secret image across rounds of a dialog, averaged across games and workers. As explained in Sec. 4.3, image ranks are computed via distance in embedding space from the guessed image (and hence, are only an estimate). We see that the human-ALICE team performs about the same for both ALICE$_{SL}$ and ALICE$_{RL}$ across rounds of dialog in a game. When compared with a baseline

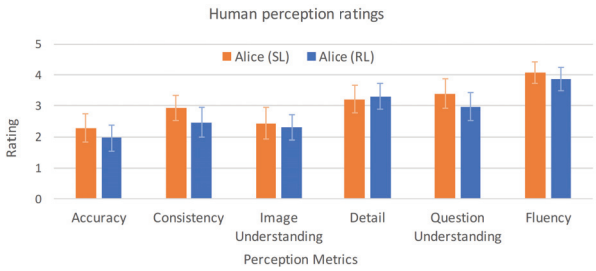


Figure 5: Worker ratings for ALICE$_{\text{SL}}$ and ALICE$_{\text{RL}}$ on 6 metrics. Higher is better. Error bars are 95% confidence intervals from 1000 bootstrap samples. Humans perceive no significant differences between ALICE$_{\text{SL}}$ and ALICE$_{\text{RL}}$ across the 6 feedback metrics.

agent that makes random guesses after every round of dialog, the human-ALICE team clearly performs better.

Statistical tests. Observe that on both the metrics (MR and MRR), the differences between performances of ALICE$_{\text{SL}}$ and ALICE$_{\text{RL}}$ are within error margins. Since both standard error and bootstrap based 95% confidence intervals overlap significantly, we ran further statistical tests. We find no significant difference between the mean ranks of ALICE$_{\text{SL}}$ and ALICE$_{\text{RL}}$ under a Mann-Whitney U test ($p = 0.44$).

6.2 Human perception of AI teammate

At the end of each HIT, we asked workers for feedback on ALICE. Specifically, we asked workers to rate ALICE on a 5-point scale (where 1=Strongly disagree, 5=Strongly agree), along 6 dimensions. As shown in Fig. 5, ALICE was rated on – how accurate they thought it was (accuracy), how consistent its answers were with its previous answers (consistency), how well it understood the secret image (image understanding), how detailed its answers were (detail), how well it seemed to understand their questions (question understanding) and how fluent its answers were (fluency).

We see in Fig. 5 that humans perceive both ALICE$_{\text{SL}}$ and ALICE$_{\text{RL}}$ as comparable in terms of all metrics. The small differences in perception are not statistically significant.

6.3 Questioning Strategies

Fig. 6 shows the distribution of questions that human subjects ask ALICE in GuessWhich. Akin to the format of the human-human GuessWhat game, we observe that binary (yes/no) questions are overwhelmingly the most common question type, for instance, "Is there/the/he ...?" (region shaded yellow in the figure), "Are there ...?" (region shaded red), etc. The next most frequent question is "What color ...?". These questions may be those that help the human discriminate the secret image the best. It could also be that humans are attempting to play to the perceived strengths of ALICE. As people play multiple games with ALICE, it is possible that they discover ALICE's strengths and learn to ask questions that play to its strengths. Another common question type is counting questions, such as "How many ...?". Interestingly, some workers adopt the strategy of querying ALICE with a single word (e.g., nouns such as

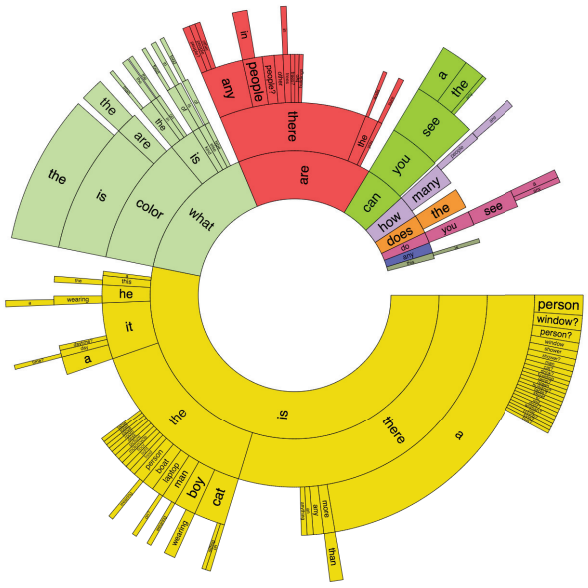


Figure 6: Distribution of first n-grams for questions asked to ALICE. Word ordering starts from the center and radiates outwards. Arc length is proportional to the number of questions containing the word. The most common question-types are binary – followed by 'What color..' questions.

"people", "pictures", etc.) or a phrase (e.g., "no people", "any cars", etc.). This strategy, while minimizing human effort, does not appear to change ALICE's performance. Fig. 7 shows a game played by two different subjects.

7 Challenges

There exist several challenges that are unique to human computation in the context of evaluating human-AI teams, for instance, making our games engaging while still ensuring fair and accurate evaluation. In this section, we briefly discuss some of the challenges we faced and our solutions to them.

Knowledge Leak. It has been shown that work division in crowdsourcing tasks follows a Pareto principle (Little 2009), as a small fraction of workers usually complete a majority of the work. In the context of evaluating an AI based on performance of a human-AI team, this poses a challenge.

Recently, (Chandrasekaran et al. 2017) showed that human subjects can predict the responses of an AI more accurately with higher familiarity with the AI. That is, a human's knowledge gained from familiarity with their AI teammate, can bias the performance of the human-AI team – knowledge from previous tasks might leak to later tasks. To prevent a biased evaluation of team performance due to human subjects who have differing familiarity with ALICE, every person only plays a fixed number of games (10) with ALICE. Thus, a human subject can only accept one task on AMT, which involves playing 10 games. The downside to this is that our ability to conduct a fair evaluation of an AI in an interactive, game-like setting is constrained by the number of unique workers who accept our tasks.

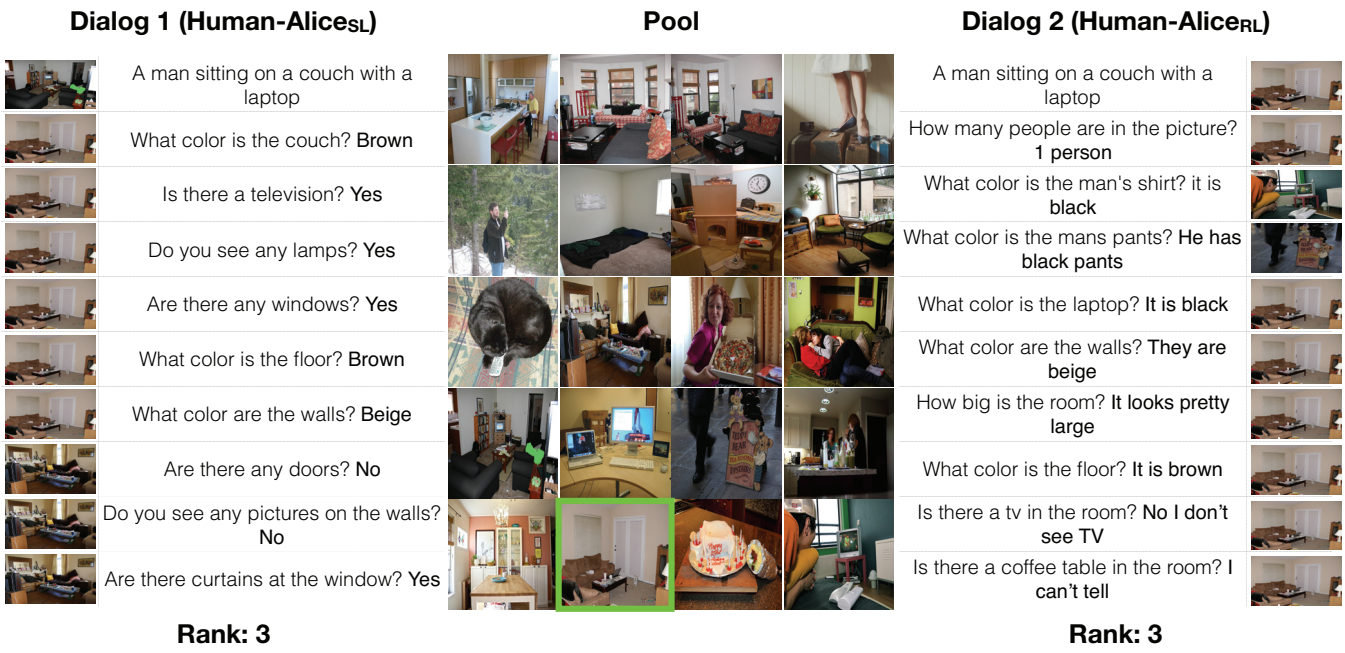


Figure 7: We contrast two games played by different workers with ALICE$_{SL}$ and ALICE$_{RL}$ on the same pool (secret image outlined in green). In both cases, the workers are able to find the secret image within three guesses. It is also interesting to note how the answers provided by ALICE are different in the two cases.

Engagement vs. Fairness. In order to improve user-engagement while playing our games, we offer subjects performance-based incentives that are tied to the success of the human-AI team. There is one potential issue with this however. Owing to the inherent complexity of the visual dialog task, ALICE tends to be inaccurate at times. This increases both the difficulty and unpredictability of the game, as it tends to be more accurate for certain types of questions compared to others. We observe that this often leads to unsuccessful game-plays, sometimes due to errors accumulating from successive incorrect responses from ALICE to questions from the human. In a few other cases, the human is misled by ALICE by a single wrong answer or by the seed caption that tends to be inaccurate at times. While we would like to keep subjects engaged in the game to the best extent possible by providing performance-based incentives, issuing a performance bonus that depends on both the human and ALICE (who is imperfect), can be dissatisfying. To be fair to the subjects performing the task while still rewarding good performance, we split our overall budget for each HIT into a suitable fraction between the base pay (majority), and the performance bonus.

8 Conclusion

In contrast to the common practice of measuring AI progress in isolation, our work proposes benchmarking AI agents via interactive downstream tasks (cooperative games) performed by human-AI teams. In particular, we evaluate visual conversational agents in the context of human-AI teams. We design a cooperative game – GuessWhich – that involves a human engaging in a dialog with an answerer-bot (ALICE) to identify a secret image known to ALICE but unknown to the human from a pool of images. At the end of the dialog, the human is asked to pick out the secret image from the image pool by making successive guesses. We find that ALICE$_{RL}$ (fine-tuned with reinforcement learning) that has been found to be more accurate in AI literature than it's supervised learning counterpart when evaluated via a questioner bot (QBOT)-ALICE team, is not more accurate when evaluated via a human-ALICE team. This suggests that there is a disconnect between between benchmarking of AI in isolation versus in the context of human-AI interaction. An interesting direction of future work could be to evaluate QBOT via QBOT-human teams.

We describe the game structure and the backend architecture and discuss the unique computation and infrastructure challenges that arise when designing such live interactive settings on AMT relative to static human-labeling tasks. Our code and infrastructure is publicly available on github.com/VT-vision-lab/GuessWhich.

Acknowledgements

We would like to acknowledge the effort provided by workers on Amazon Mechanical Turk. We are grateful to the developers of Torch (Collobert, Kavukcuoglu, and Farabet 2011) for building an excellent framework. This work was funded in part by NSF CAREER awards to DB and DP, ONR YIP awards to DP and DB, ONR Grant N00014-14-1-0679 to DB, ONR Grant N00014-16-1-2713 to DP, a Sloan Fellowship to DP, an Allen Distinguished Investigator award to DP from the Paul G. Allen Family Foundation, Google Faculty Research Awards to DP and DB, Amazon Academic

Research Awards to DP and DB, AWS in Education Research grant to DB, and NVIDIA GPU donations to DB. SL was partially supported by the Bradley Postdoctoral Fellowship. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

Aras, H.; Krause, M.; Haller, A.; and Malaka, R. 2010. Webpardy: harvesting qa by hc. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 49–52. ACM.

AWS. 2017. Amazon. https://aws.amazon.com/ec2/. [Online; accessed 04-May-2017].

Chamberlain, J.; Poesio, M.; and Kruschwitz, U. 2008. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, 42–49.

Chandrasekaran, A.; Yadav, D.; Chattopadhyay, P.; Prabhu, V.; and Parikh, D. 2017. It Takes Two to Tango: Towards Theory of AI's Mind. *arXiv preprint arXiv:1704.00717*.

Collobert, R.; Kavukcuoglu, K.; and Farabet, C. 2011. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017a. Visual Dialog. In *CVPR*.

Das, A.; Kottur, S.; Moura, J. M.; Lee, S.; and Batra, D. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*.

de Vries, H.; Strub, F.; Chandar, S.; Pietquin, O.; Larochelle, H.; and Courville, A. 2017. GuessWhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

Diakopoulos, N.; Luther, K.; and Essa, I. 2008. Audio puzzler: piecing together time-stamped speech transcripts with a puzzle game. In *Proceedings of the 16th ACM international conference on Multimedia*, 865–868. ACM.

Grosz, B. 2012. What question would turing pose today? *AI Magazine* 33(4):73.

Jain, S., and Parkes, D. C. 2013. A game-theoretic analysis of the esp game. *ACM Trans. Econ. Comput.* 1(1):3:1–3:35.

Karpathy, A. 2016. Neuraltalk2. https://github.com/karpathy/neuraltalk2. [Online; accessed 04-May-2017].

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. L. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*.

Krause, M., and Smeddinck, J. 2011. Human computation games: A survey. In *Signal Processing Conference, 2011 19th European*, 754–758. IEEE.

Law, E., and Von Ahn, L. 2009. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1197–1206. ACM.

Law, E. L.; Von Ahn, L.; Dannenberg, R. B.; and Crawford, M. 2007. Tagatune: A game for music and sound annotation. In *ISMIR*, volume 3, 2.

Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; and Jurafsky, D. 2016. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollr, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Little, G. 2009. How many turkers are there.(dec 2009).

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*.

Michelucci, P. 2013. Handbook of human computation. In *Springer*.

Paek, T. 2001. Empirical methods for evaluating dialog systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9*.

RabbitMQ. 2017. RabbitMQ. https://www.rabbitmq.com/. [Online; accessed 04-May-2017].

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Strub, F.; de Vries, H.; Mary, J.; Piot, B.; Courville, A. C.; and Pietquin, O. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *CHI*.

Von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM* 51(8):58–67.

Von Ahn, L.; Liu, R.; and Blum, M. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 55–64. ACM.