Evaluating Visual Conversational Agents via Cooperative Human-AI Games

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, Devi Parikh

Introduction

- Al routinely measured in isolation, without a human in the loop
- Design a cooperative game GuessWhich to measure human-AI team performance
- Compare supervised baseline models with QBOT-ABOT teams trained through reinforcement learning based self-talk on this image-guessing task
- AI-AI teams improve significantly at guessing the correct image compared to the supervised pretraining
- Results indicate that self-talk fine-tuned agents are better visual conversational agents, remains unclear if these agents are better at this task when interacting with humans

Introduction

- ALICE_SL rained in a supervised manner on the Visual Dialog dataset
- ALICE_RL pre-trained with supervised learning and fine-tuned via reinforcement learning
- Evaluate human-AI team performance on this game for both supervised learning (SL) and reinforcement learning (RL) versions of ALICE
- Main finding: Despite significant differences between SL and RL agents reported in previous work, they find no significant difference in performance between ALICE_SL and ALICE_RL when paired with human partners
- Disconnect between AI-AI and human-AI evaluations

GuessWhich Game

- Players: Replace QBOT in AI-AI dialog with humans to perform a collaborative task of identifying a secret image from a pool
- Gameplay: ALICE is assigned a secret image from a pool of images from COCO dataset
- Prior to beginning dialog, both ALICE and H are given a brief description (caption) of image
- H asks ALICE a question qt about the secret image to identify it from the pool and ALICE responds with an answer a
- After each round, H selects an image based on the dialog so far
- At the end of 9 rounds of dialog, H asked to successively click on best guess
- Interface gives H feedback on whether guess is correct, continues until H guesses true image

Guess Which

Total bonus so far: \$ 0 Possible bonus for this game: \$ 0.2 Game 1/10



and answer your questions, out i am not perfect. I make quite a few mistakes. I hope we can work together to find the image! Let's do this! Note: My knowledge of English is limited. Sometimes if I don't know the right word, I say UNK.



Based on your understanding of the image description, pick the image that you think is the most relevant.



You can update your guess now. You will get a bonus if you correctly guess the target image. Bonus won will be shown at the end of the game. Once you're ready, ask me a question! You can only ask me questions (avoid using "Thanks", "Okay" for example), and please keep your questions relevant to image content, or your work will be rejected.



•• white

Click on the image that is your best guess

Evaluation

- Mean Rank (MR): mean rank of the secret image (i.e. number of guesses it takes to identify the secret image)
- Lower values indicate better performance
- Mean Reciprocal Rank (MRR): 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

Results



Mean Rank vs. Number of Games

—Alice (RL) —Alice (SL) —Random (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.

fc7 Mean Rank vs. Round of Dialog



—Alice (RL) fc7 —Alice (SL) fc7 —Random (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.

Results



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

Dialog 1 (Human-Alice_{SL})



A man sitting on a couch with a laptop



What color is the couch? Brown



Is there a television? Yes



Do you see any lamps? Yes



Are there any windows? Yes



What color is the floor? Brown



What color are the walls? Beige



Are there any doors? No



Do you see any pictures on the walls? No

Are there curtains at the window? Yes



Pool



Dialog 2 (Human-Alice_{RL})

N	A man sitting on a couch with a laptop	
	How many people are in the picture? 1 person	
	What color is the man's shirt? it is black	
	What color is the mans pants? He has black pants	
	What color is the laptop? It is black	Ren .
S.	What color are the walls? They are beige	and a
	How big is the room? It looks pretty large	
	What color is the floor? It is brown	
1	Is there a tv in the room? No I don't see TV	and a
	Is there a coffee table in the room? I	



Rank: 3