Imitating Interactive Intelligence
Stanford CS379c Lecture
April 20, 2021

Greg Wayne, DeepMind
Interactive Agents Group
Interactive Agents
Creating agents that cooperatively interact with humans.

Why interaction?

1. Interaction with humans is the best test of intelligence (Turing, 1951).

2. Agents that interact with humans (answering questions, helping, learning socially) could profoundly enable people.

Our long-term goal: produce agents that can learn socially from humans as does a child or peer.
Social Learning is the Source of Human Intelligence
(Non-verbal intelligence test)

Problem solving involving spatial understanding

Involving e.g. non-verbal imitation

Herrmann et al., Science, 2007: Cited in Tomasello, 2019
What’s an Agent?
Typical Training Paradigm

Agent interacts with environment and receives programmed reward for successes.

Example: play Go and receive reward = 1 upon winning, reward = -1 upon losing, reward = 0 at other moments.

\[
\sum_{a_t} \pi_{\theta}(a_t \mid o_{\leq t}) \mathbb{E}_{\pi_{\theta}}[R_t \mid o_{\leq t}, a_t]
\]

\[
R_t = \sum_{t' \geq t} r_{t'} = r_t + r_{t+1} + r_{t+2} \ldots
\]
From Untrained Agents to Agents that Interact

Human children: organic process of nurtured and self-directed learning

Untrained AIs: nurtured and self-directed learning hard to implement

- [Lack human objectives]: We do not (yet) understand human drives and motivations at an algorithmic level: complicated, species-specific, hard-to-guess.

- [Feedback from scratch]: Agents begin at *tabula rasa* (blank slate / monkeys typing on typewriters). Intractable for humans to watch untrained agents and give reinforcing feedback until agents reach competence in practical amounts of time.

- [Ambiguity in communication]: Even simple instructions can be ambiguous. “Go near the door.” What is “near”? 

Therefore, it is difficult to (a) write down an objective for agent development; (b) provide feedback for untrained agents; and (c) formalize reward for even very simple communicative interactions.
Imitation Learning for Creating Behavioral *Priors*

Increasingly commonly: use supervised learning as an initial basis for behavior. Then improve from there.

**GPT-3**

**AlphaGo**

- In AlphaGo, dataset of human play was later replaced
  - Self-play in a win-lose game is a good curriculum
  - For general cooperative interactions, don’t have win-lose or a dataset

$$
\mathbb{E}_{s,a \sim \mathcal{D}} \left[ \log \pi_\theta(a \mid s) \right]
$$
Strategy: Create Dataset of Interactions
Playroom Virtual Environment
Eliciting Diverse Interactions

Prompt: Ask a question about the color of an object.

Setter: What color are the shelves in the corner?

Solver: They are red.
From Prompts to Instructions

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Prompt</th>
<th>Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q&amp;A</td>
<td>Ask the other player to count something</td>
<td>Count the lamps on the floor</td>
</tr>
<tr>
<td></td>
<td>Ask a question about the color of something</td>
<td>How many toys are on the bed?</td>
</tr>
<tr>
<td></td>
<td>Ask the other player where something is</td>
<td>What color is the chair?</td>
</tr>
<tr>
<td></td>
<td>Ask the other player to lift something</td>
<td>Where is the blue duck?</td>
</tr>
<tr>
<td></td>
<td>Ask the other player to go somewhere</td>
<td>Lift any object which is close to the magenta table</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Can you lift the rocket on the bookshelf?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Grab the red toy on the table and lift it</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Go as far away from the table as possible</td>
</tr>
</tbody>
</table>

Provided by environment

Generated by the human
<table>
<thead>
<tr>
<th>Prompt</th>
<th>Full text</th>
</tr>
</thead>
<tbody>
<tr>
<td>go</td>
<td>Ask the other player to go somewhere</td>
</tr>
<tr>
<td>lift</td>
<td>Ask the other player to lift something</td>
</tr>
<tr>
<td>position object</td>
<td>Ask the other player to position something relative to something else</td>
</tr>
<tr>
<td>position yourself</td>
<td>Ask the other player to stand in some position relative to you</td>
</tr>
<tr>
<td>bring me</td>
<td>Ask the other player to bring you one or more objects</td>
</tr>
<tr>
<td>touch</td>
<td>Ask the other player to touch an object using another object</td>
</tr>
<tr>
<td>push object</td>
<td>Ask the other player to push an object around using another object</td>
</tr>
<tr>
<td>make a row</td>
<td>Ask the other player to put three or more specific objects in a row</td>
</tr>
<tr>
<td>arrange</td>
<td>Ask the other player to move a group of objects into a simple arrangement</td>
</tr>
<tr>
<td>put on top</td>
<td>Ask the other player to put something on top of something else</td>
</tr>
<tr>
<td>put underneath</td>
<td>Ask the other player to put something underneath something else</td>
</tr>
<tr>
<td>freestyle activity</td>
<td>Ask the other player to perform an activity of your choice</td>
</tr>
<tr>
<td>say what you see</td>
<td>Ask the other player to say what they are looking at or noticing right now</td>
</tr>
<tr>
<td>question about colour</td>
<td>Ask a question about the colour of something</td>
</tr>
<tr>
<td>question about existence</td>
<td>Ask the other player whether a particular thing exists in the room</td>
</tr>
<tr>
<td>describe location</td>
<td>Ask the other player to describe where something is</td>
</tr>
<tr>
<td>count</td>
<td>Ask the other player to count something</td>
</tr>
</tbody>
</table>

**Table 4: Prompts used in language games.**

<table>
<thead>
<tr>
<th>Modifier</th>
<th>Full text</th>
</tr>
</thead>
<tbody>
<tr>
<td>refer to objects by colour</td>
<td>Try to refer to objects by colour</td>
</tr>
<tr>
<td>refer to location by colour</td>
<td>Try to refer to the location by colour</td>
</tr>
<tr>
<td>use shape words</td>
<td>Try to use shape words like: circular, rectangular, round, pointy, long</td>
</tr>
<tr>
<td>refer to objects by location</td>
<td>Try to refer to objects by location</td>
</tr>
<tr>
<td>use proximity words</td>
<td>Try to use words like: near, far, close to, next to</td>
</tr>
<tr>
<td>use horizontal position words</td>
<td>Try to use words like: in front, behind, left of, right of, between</td>
</tr>
<tr>
<td>use vertical position words</td>
<td>Try to use words like: on top, beneath, above, below</td>
</tr>
<tr>
<td>use negation words</td>
<td>Try to use words like: not, isn't</td>
</tr>
<tr>
<td>use quantifier words</td>
<td>Try to use words like: some, all, most, many, none</td>
</tr>
<tr>
<td>not bed, door, or window</td>
<td>Do not use the words: bed, door, window</td>
</tr>
</tbody>
</table>

**Table 5: Modifiers used in language games.**

24 base prompts, ~10 modifiers

More than one year of video data

610,608 episodes

320,144 unique setter instructions of length 7.5 +/- 2 words
Human-Human Interaction
(dataset example)

Prompt: Ask the other player to describe where something is

Setter: Where is the orange color pillow

Solver: on the bed
Recorded Data

1. Setter Gives Instruction
   "please lift the object next to the magenta table"

2. Solver Completes Task

Action Sparsity
- Move
- Look
- Grab

Move Actions
- Forward
- Back
- Left
- Right
- Forward Left
- Forward Right

Look Actions
- Vertical
- Horizontal
Agent Model

\[ \mathbb{E}_{s, a \sim \mathcal{D}} \left[ \log \pi_\theta(a \mid s) \right] \]
Learning from Data

- Human demonstrations
- Behavioural cloning
- Forward RL
- Auxiliary learning
- Policy
- Inverse RL
- Reward model
- No programmed reward
- Auxiliary learning
Weaknesses of Behavioral Cloning

- Does not utilize environment interaction to learn how to respond to unusual contingencies
- Provides a relatively weak signal to train perceptual similarity

\[ \pi_\theta (a_t \mid o_{\leq t}) \]

Seen this

\[ \pi_\theta (a_t \mid o_{\text{novel, } \leq t}) \]

What to do now?

Are novel observations similar to previously seen ones?

How to act?
Use Language to Instruct Similarity

Consider two visual movies similar if it is not possible to distinguish their instructions.

\[ D(\text{movie, instruction}) \]

Classify plausibility of movie-instruction pair from real versus shuffled dataset

- Movie 1
- Movie 2
- Movie 3
- Movie 4
- Movie 5

- Instruction 1
- Instruction 2
- Instruction 3
- Instruction 4
- Instruction 5
Language Matching Objective

Diagram:
- Image -> Augment (Crop, Rotate, Shear, Translate)
- ResNet
- Flatten
- Multi-Modal Transformer
- Prompt
- Prev. Lang.
- Tokenize & Embed
- Inter-Agent Comms.
- Text String
- Language Matching Loss
Weaknesses of Behavioral Cloning (2)
Goals can be more compact than policies.

Consider: robot designed to climb Mount Everest.
Policy is arguably very complicated. Must prescribe what to do in each scenario.
But the goal is simple: maximize altitude.
If the goal is known and success is measurable, then it is possible to practice with goal to acquire the policy.
Learning a Reward Model
version of GAIL (Ho and Ermon, 2016)

Discriminate between agent and human behavior using features from language matching.
Comparing Contributions on a Simple Task
“Put X on Bed”
Interactive Training

[Env]: ask a question about the color of something
[Setter]: what is the color of the wall
[Solver]: blue
Evaluation
From code to human interaction

A
- Total Loss
- GAIL Discriminator Loss
- Language Match Loss

B
- Same Object Lifted
- Object Mention Acc.
- Avg. Eval Reward

C
- Heatmap showing performance across different conditions:
  - Human
  - BGR·A
  - BG·A
  - B·A
  - B (no vis.)
  - B (no lang.)

Legend:
- B = Behavioural Cloning
- G = GAIL
- A = Auxiliary Losses
- R = Setter Replay
Inspecting Reward Model

1. Setter Perspective (t=3.8s)  
2. Solver Perspective (t=8.3s)  
3. Solver Perspective (t=36.7s)

- take the *white* robot and place it on the bed
- take the *red* robot and place it on the bed
Scaling and Transfer Performance

See Scaling Laws for Neural Language Models (Kaplan et al., 2019)
Human Evaluation Techniques
Human Evaluation (Observational)

A

Solver

Human
BGR\cdot A
BG\cdot A
B\cdot A
B
B (no vis.)
B (no lang.)

Setter

Human
BGR\cdot A
BG\cdot A
B\cdot A
B
B (no vis.)
B (no lang.)

Instruction Following

Question Answering

Success rate

Success rate

B

Instruction-following

Question-answering

Joint success rate
Human Evaluation (Interactive)
clear both tables
A Small Callback to the Programmer’s Apprentice
SHRDLU (Winograd, 1968)

Original screen display

Later color rendering (Univ. of Utah)
Human-Computer Interaction

Computers are being used today to take over many of our jobs. They can perform millions of calculations in a second, handle mountains of data, and perform routine office work much more efficiently and accurately than humans. But when it comes to telling them what to do, they are tyrants. They insist on being spoken to in special computer languages, and act as though they can't even understand a simple English sentence.

Let us envision a new way of using computers so they can take instructions in a way suited to their jobs. We will talk to them just as we talk to a research assistant, librarian, or secretary, and they will carry out our commands and provide us with the information we ask for. If our instructions aren't clear enough, they will ask for more information before they do what we want, and this dialog will all be in English.

Winograd, 1971
Discussion