Reinforcement Learning with Human Teachers

Evidence of Feedback and Guidance with Implications for Learning Performance

Andrea Thomaz Cynthia Breazeal

Presented by Jonas Schwertfeger
Brown University
CS296-3

March 14th, 2007

Introduction

As robots enter consumer markets it is desirable that they are able to

- perform various tasks (multi-purpose robots)
- perform tasks not known during manufacturing stage

Thus, need possibility to alter behavior of robots. How can we do this?

- Reprogram them using some form of programming language
  - Time-consuming
  - Extensive technical knowledge required
- Build robots that have ability to learn from humans
  - We need to understand how people want to teach
  - Incorporate these insights into machine learning frameworks

Characterization of Human-Trainable Systems

- Implicit vs. explicit training
  - Implicit: Agent (robot) learns through passive observation of user's behavior
  - Explicit: Human teaches agent through interaction
- Who leads interaction in case of explicit training
  - Agent queries human regardless of what human wants to or can teach
  - Human decides on when to teach what
- Balance between autonomous exploration and human guidance

Thomaz and Breazeal specifically look at explicit learning with human as interaction leader. Framework used: Reinforcement learning with human teachers.

Experiments in Sophie's Kitchen

Thomaz and Breazeal investigate human-agent interaction with a computer game platform called Sophie's Kitchen

- Sophie likes to bake a cake but needs to learn how
- Kitchen has five objects: *Flour*, *eggs*, a *spoon*, a *bowl* and a *tray*
- Bowl can be in state *empty*, *flour*, *eggs*, *both* or *mixed*
- Tray can be in state *empty*, *batter* or *baked*
- Any of the objects can be at one of three locations: *Shelf*, *table* or *oven*
- Sophie is able to carry out five actions: *Go left*, *go right*, *pick object up*, *put object down* and *use a picked up object on another object*

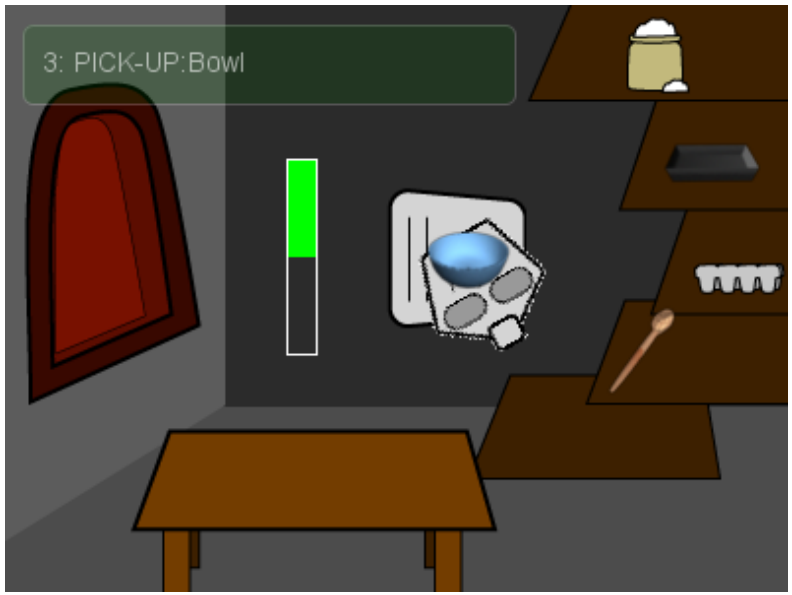
Sophie's Kitchen as a Markov Decision Process

- A world $W = (L, O, \Sigma, T)$ is a finite set of k locations $L = \{l_1, \ldots, l_k\}$ and n objects $O = \{o_1, \ldots, o_n\}$
- Each object o_i can be in one of an object-specific number of states. Ω_i is the set of states o_i can be in, and $O^* = (\Omega_1 \times \ldots \times \Omega_n)$ is the entire object configuration space
- W defines a set of legal (physically possible) states $\Sigma \subset (L \times L^O \times O^*)$, where $L^O = (L \times \ldots \times L)$ is the object location space
- Further, W defines a transition function $T : \Sigma \times A \mapsto \Sigma$, where A is the action space

Sophie's Kitchen as an Interactive Java Program

- In initial state S_0 all objects and the agent are at location *shelf*
- A successful completion of the task includes
  - putting flour and eggs into bowl,
  - stirring ingredients using spoon,
  - transferring batter into tray
  - and putting tray into oven
- Human rewards actions using mouse, $r \in [-1, 1]$
  - Object-specific and general rewards are possible
  - But no distinction is made for learning algorithm

- Standard Q-Learning is used as algorithm ($\alpha = .3, \gamma = .75$)



First Round Results and Analysis

18 participants played the game. Important findings from activity logs and interviews:

- Many people assumed object-specific rewards were future directed (guidance)
- Very often rewards were not related to last action but pertained to future action
  - 15 of 18 participants gave rewards to bowl/tray sitting empty on shelf

Interpretation: Humans do not only want to reward past actions but want to influence the agent's action selection process; learning algorithms should pay attention to this

Incorporating Guidance

Thomaz and Breazeal added "guidance communication channel" to their simulation

- Agent pauses 1.5s in each learning iteration before deciding on action
- Click with right mouse button on object communicates guidance message to agent ("pay attention to this object")
- Agent registers guidance to bias action selection
- Q-Learning algorithm is extended to incorporate bias

Extending Q-Learning for Guidance

Basic algorithm:

while learning **do**
  a = random select weighted by $Q[s, a]$ values
  execute a, and transition to s' (small delay to allow for human reward)
  sense rewared, r
  update values: $Q[s, a] \leftarrow Q[s, a] + \alpha(r + \gamma(\max_{a'} Q[s', a']) - Q[s, a])$
end while

Extending Q-Learning for Guidance

Basic algorithm:

while learning **do**
  a = random select weighted by $Q[s, a]$ values
  execute a, and transition to s' (small delay to allow for human reward)
  sense rewared, r
  update values: $Q[s, a] \leftarrow Q[s, a] + \alpha(r + \gamma(\max_{a'} Q[s', a']) - Q[s, a])$
end while

Extended algorithm:

while learning **do**
>> **while** waiting for guidance **do**
>> **if** receive human guidance message **then**
>> g = *guidance − object*
>> **end if**
>> **end while**
>> **if** received guidance **then**
>> a = random selection of actions containing g
>> **else**
  a = random select weighted by $Q[s, a]$ value
>> **end**
  execute a, and transition to s' (small delay to allow for human reward)
  sense rewared, r
  update values: $Q[s, a] \leftarrow Q[s, a] + \alpha(r + \gamma(\max_{a'} Q[s', a']) - Q[s, a])$
end while

Second Round Results and Analysis

Extended game was played by additional 12 participants (with easier task)

1 expert played 10 times wo/guidance, 10 times with

Measure	Mean no guide	Mean guide	Chg
# trials	6.4	4.5	30%
# actions	151.5	92.6	39%
# failures	4.4	2.3	48%
# f before s	4.2	2.3	45%
# states	43.5	25.9	40%

11 non-experts played with guidance

Measure	Mean no guide	Mean guide	Chg
# trials	28.5	14.6	49%
# actions	816.4	368	55%
# failures	18.9	11.8	38%
# f before s	18.7	11	41%
# states	124.4	62.7	50%

(Non-experts played wo/guidance on full task?)

Discussion

Findings:

- People assume they can guide an agent, not only give feedback
- Q-Learing can be extended with guidance component
- Guidance provides more successful teaching experience
- Guidance helps keep exploration of agent in smaller, positive portion of state space