Modular Reinforcement Learning for Embodied Cognition

Dana H. Ballard
Computer Science Dept.
University of Texas at Austin
Outline: Modeling Human Cognitive Architectures

- Embodiment confers enormous advantages
- Information acquisition is fragmentary
- Control programs for ordinary behaviors
- Tiered cognitive architectures – influence of robotics
  - Middle tier is RL based; modules confer enormous advantage
  - Eye gaze model
  - Credit assignment model
  - Fast Inverse RL algorithm
Embodied Cognition

Body does computation that the brain does not have to repeat, making the brain’s computation simpler.

Maurice Merleau-Ponty 1906-1961
Computing with a physical device:
e.g. Find the shortest path in a graph between the two red nodes
Timescales
Study from Heider and Simmel (1944) replicated w eye-tracking added
Summary of Embodiment

Abstractions make use of body’s capabilities
e.g. motor control

Control structure is discrete and uses sparse
Inputs and outputs

A huge amounts of experiential indexing is tied to
elemental tokens
e.g. Heider and Simmel animation
Visual Routines

Tracing
Monkeys have to make a saccade To the end of line connected to the fixation point.

Search & Tracing
Monkeys have to make a saccade To the end of line that has the same colored stub as the fixation point.
Fixations use different computational mechanisms

Rao et al Vision Research 2002

Search using memory

Search using correlation
Every fixation has a specific purpose
Every fixation has a specific purpose
Summary of Evidence for ‘Visual Routines’ Theory of Vision (Ullman, Kosslyn 80s)

Visual tasks can be concatenated
   1) Monkey finds the stub of the correct color, then 2) traces

Different visual routines are used depending on context
   locate from memory vs locate with template

Information is not reacquired
   Once a color has been registered, there is no need to re-interrogate the optic array
Using the gaze system to orchestrate tasks

Human eye has a pronounced high-resolution fovea, loosely characterized as the "width of a thumb held at arm’s length."

Inside the fovea the resolution is 100 times that of the visual periphery.
Making a PBJ sandwich

Cognitive Architectures

Expert Systems
- ACT-R (Anderson)
- SOAR (Newell)
- Icarus (Langley)

Embodied
- Subsumption (Brooks)

3T (Bonasso)
- SMR (Ballard)

Not Embodied

Embodied

Embodied

Abstraction Layers
Disembodied Architectures have problems with level of description and variable binding e.g. ICARUS Driving simulation

Table 2. Primitive and nonprimitive ICARUS skills for the in-city driving domain.

<table>
<thead>
<tr>
<th>Skill Description</th>
<th>Code</th>
</tr>
</thead>
</table>

If u r going in the wrong direction, make a U-turn

If u r steering at an angle, straighten out

Make a right turn at the intersection
Cognitive Architectures: Influence from Robotics

<table>
<thead>
<tr>
<th>Bonasso’s 3T</th>
<th>RL modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planner</td>
<td>Scheduler</td>
</tr>
<tr>
<td>Sequencer</td>
<td>Modules</td>
</tr>
<tr>
<td>Reactive Skills</td>
<td>I/O routines</td>
</tr>
</tbody>
</table>
Modules
There is an enormous library of special purpose behaviors

- Intrinsic modules
- Allow indep. reward scaling

Scheduling
At any moment only a small subset are activated

Active Modules
Refine their behavior using RL
Negotiate for body’s resources via Routines
Decision-making venues

- **Env.**
- **Decision Theory**
  - Uncertainty
- **Game Theory**
  - Uncertainty
  - Active agent
- **Reinforcement Learning**
  - Uncertainty
  - Delayed Reward
A Monkey uses Secondary Reward
Do dopamine neurons report an error in the prediction of reward?

No prediction
Reward occurs

Reward predicted
Reward occurs

Reward predicted
No reward occurs

(No CS) R

CS R

-1 0 1 2 s

(No R)
Biological implementation of reward

- Dopamine
- TD-like signals observed
Reinforcement Learning Primer: Before Learning

You are here

State → Action → Reward

Reward for taking action
By trying different actions from different starting points, gradually learn the expected reward value from any starting point.
RL Definitions

An MDP consists of a 4-tuple \((S, A, T, R)\)

- \(S\) being the set of possible states,
- \(A\) the set of possible actions,
- \(T\) the transition model describing the probabilities \(P(s_{t+1}|s_t, a_t)\) of reaching a state \(s_{t+1}\) from state \(s_t\) at time \(t\) and executing action \(a_t\),
- \(R\) is the expected value of the reward \(r_t\), distributed according to \(P(r_t|s_t, a_t)\) and is associated with the transition from state \(s_t\) to some state \(s_{t+1}\) when executing action \(a_t\).
Q-Learning variant of Temporal Difference Learning

The goal of RL is to find a policy $\pi$ that maps from the set of states $S$ to actions $A$ so as to maximize the expected total discounted future reward

$$V^\pi(s) = E^\pi \left( \sum_{t=0}^{\infty} \gamma^t r_t \right)$$  \hspace{1cm} (1)

Alternatively, the values can be parametrized by state and action pairs, denoted by $Q^\pi(s, a)$.

$$Q^*(s, a) = \sum_r r P(r|s, a) + \gamma \sum_{s'\in S} P(s'|s, a) \max_{a'} Q^*(s', a')$$  \hspace{1cm} (2)

Temporal difference learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \delta_Q$$  \hspace{1cm} (3)

$$\delta_Q = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t).$$  \hspace{1cm} (4)
Modules have to agree on an action

\[ M_i = \{ S_i, A, T_i, R_i \} \]

One way: Average

\[
Q(s_t, a_t) = \sum_{i=1}^{M} Q(s_t^{(i)}, a_t^{(i)})
\]

Better way: Softmax using

\[
P(a_t^{(j)} | Q(s_t^{(1)}, a_t), ..., Q(s_t^{(M)}, a_t)) = \frac{e^{Q(s_t^{(j)}, a_t^{(j)}) / \tau}}{\sum_{i=1}^{M} e^{Q(s_t^{(i)}, a_t^{(i)}) / \tau}}
\]
MDPs or POMDPS?

POMDP model captures
The fundamentals …

But …

MDP can take advantage of
1) In many embodied situations
   independent modules are possible

2) In a dynamic situation, an aliased
   state situation may be resolved in time
A human exhibits multitasking

- Pickup purple boxes
- Stay on sidewalk
- Avoid blue obstacles
Modules learning in a multi-tasking context

Walter
Overhead view of trajectory

Initial performance

After 100 episodes

After 150 episodes
State Spaces definitions
1. Visual Routine

2a. Policy

2b. $V$ is value of Policy

$V(s) = \max_a Q(s,a)$

Module for Litter Cleanup

Heading from Walter’s perspective
Learned Module Behaviors

- **Litter**
- **Sidewalk**
- **Obstacles**
Eye Gaze controversy- A job for modular RL!

Eye gaze models invoke bottom up saliency; image features are
The source of gaze locations

Modular RL is top-down agenda driven; active modules compete
For the use of gaze to reduce uncertainty
What directs gaze? One answer: Saliency

Salient locations are a good candidate for fixations but by and large, the task dominates.
Humans perform the sidewalk navigation task with visual distractions
Human gaze data

Substitution
Human gaze data

Not executing instructed task

Substitution
Human gaze data: Task determines fixation point
Which module should get the gaze vector?

Before Observation

Without gaze, the location in state space becomes increasingly uncertain.
Which module should get the gaze vector?

After Observation

Making a measurement with a fixation reduces this uncertainty
Sprague’s hypothesis

\[
\text{loss}_b = \mathbb{E} \left[ \max_a \left( Q_b(s_b, a) + \sum_{i \in B, i \neq b} Q^E_i(s_i, a) \right) \right] - \sum_i Q^E_i(s_i, a_E)
\]

Reward where \( b \) updates using gaze  
Average reward no updates
<table>
<thead>
<tr>
<th>obs</th>
<th>side</th>
<th>can</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Obs.png" /></td>
<td><img src="image" alt="Side.png" /></td>
<td><img src="image" alt="Can.png" /></td>
</tr>
</tbody>
</table>

- **A priori uncertainty**
- **A posteriori uncertainty**
Performance Comparison

Gaze services random task
Gaze services tasks in order
Gaze services task w highest reduction in reward uncertainty

Sprague et al ACM TAP 2007
John Senders '60s
Summary of Part I:

In building human cognitive models, embodiment constraints are essential.

RL modules allow one to incrementally extend the capabilities of the model.

RL modules provides a good account of gaze fixations.
Outline: Modeling Human Cognitive Architectures

- Embodiment confers enormous advantages
- Information acquisition is fragmentary
- Control programs for ordinary behaviors
- Tiered cognitive architectures – influence of robotics
- Middle tier is RL based; modules confer enormous advantage
  - Eye gaze model
  - Credit assignment model
  - Fast Inverse RL algorithm
Credit Assignment: how do modules learn their share of reward?

Situations where you need to calibrate reward:

- Solving a problem with a new set of modules
- Using a set of modules where only the global reward is known and the problem is to assign credit appropriately
Modules architecture: Episodes

A

inactive modules

scheduler

active modules

episode 1

time

episode 2

active modules

B

active modules

\( r^{(l)}_t \)

\( r^{(k)}_t \)

episode 2

time
Is this really a problem? In a multi-agent RL setting, a Kalman filtering model allows agents to learn the correct policy even when they can’t solve the credit assignment problem?

It is when different sets of modules are used, since small errors in reward estimates can lead to large offsets in state values and consequently inconsistent policies.

\[ V'^\pi(s) = \frac{c}{1-\gamma} + V^\pi(s) \]
Ten modules working in subsets fail to compute learn correctly on a simple grid world task.
Key assumption: a module knows the estimated rewards of its co-activated module set

\[ \hat{r}(i) \leftarrow (1 - \beta) \hat{r}(i) + \beta (G - \sum_{j \in M, j \neq i} \hat{r}(j)) \]
Prey & Predators problem

15 food sources
5 Predators

Adapted from Singh and Cohn
Learning reward given only global reward for prey and predator problem
Variance weighted estimates are superior

Problem: A module that has a high variance estimate can corrupt others’ estimates

Solution: weight estimates by the variances in reward

\[
\beta_i = \frac{(\sigma^{(i)})^2}{\sum_{j=1}^{N} (\sigma^{(j)})^2} = \frac{(\sigma^{(i)})^2}{\sum_{j \neq i}^{N} (\sigma^{(j)})^2 + (\sigma^{(i)})^2}
\]
Learning Q tables
And policies given only
Global reward in sidewalk venue
Learning rewards given only
Global reward for sidewalk venue
Summary of Global Reward Calibration

- Modules work in different subsets each episode.
- Modules know each other’s reward estimates.
- Modules weight estimates by variance estimates.
Human performance data shows unexpected regularities.
Inverse Reinforcement Learning

Wouldn’t it be great to estimate the rewards that humans are using based on their behavior?
Bayesian IRL

Observations from a subject:

\[ O_{\chi} = \{(s_1, a_1), (s_2, a_2) \ldots (s_k, a_k)\} \]

Probability of seeing the individual states/actions, given the reward, is a product of individual probs.

\[ Pr_{\chi}(O_{\chi}|R) = Pr_{\chi}((s_1, a_1)|R)Pr_{\chi}((s_2, a_2)|R) \ldots Pr_{\chi}((s_k, a_k)|R) \]
Bayesian IRL

Use Bayes thm

\[ Pr_x(R | O_x) = \frac{Pr_x(O_x | R)Pr(R)}{Pr(O_x)} \]
Bayesian IRL

Let's make the assumption of the following PDF:

$$Pr_x(O_x|R) = \frac{1}{Z} e^{\alpha x E(O_x, R)}$$

And furthermore:

$$E(O_x, R) = \sum_i Q^*(s_i, a_i, R)$$

So that:

$$Pr_x((s_i, a_i)|R) = \frac{1}{Z_i} e^{\alpha x Q^*(s_i, a_i, R)}$$
Bayesian IRL

Priors that emphasize sparse rewards:

\[ P_{\text{Laplace}}(R(s) = r) = \frac{1}{2\sigma} e^{-\frac{|r|}{2\sigma}}, \forall s \in S \]

\[ P_{\text{Beta}}(R(s) = r) = \frac{1}{\left(\frac{r}{R_{\text{max}}}\right)^{\frac{1}{2}} \left(1 - \frac{r}{R_{\text{max}}}\right)^{\frac{1}{2}}}, \forall s \in S \]
Bayesian IRL Algorithm

Ramachandran & Amir

**Algorithm PolicyWalk**(Distribution $P$, MDP $M$, Step Size $\delta$)

1. Pick a random reward vector $\mathbf{R} \in \mathbb{R}^{\left|S\right|}/\delta$.
2. $\pi := \text{PolicyIteration}(M, \mathbf{R})$
3. Repeat
   (a) Pick a reward vector $\tilde{\mathbf{R}}$ uniformly at random from the
   neighbours of $\mathbf{R}$ in $\mathbb{R}^{\left|S\right|}/\delta$.
   (b) Compute $Q^\pi(s, a, \tilde{\mathbf{R}})$ for all $(s, a) \in S, A$.
   (c) If $\exists(s, a) \in (S, A), Q^\pi(s, \pi(s), \tilde{\mathbf{R}}) < Q^\pi(s, a, \tilde{\mathbf{R}})$
      i. $\tilde{\pi} := \text{PolicyIteration}(M, \tilde{\mathbf{R}}, \pi)$
      ii. Set $\mathbf{R} := \tilde{\mathbf{R}}$ and $\pi := \tilde{\pi}$ with probability
          $\min\{1, \frac{P(\tilde{\mathbf{R}}, \tilde{\pi})}{P(\mathbf{R}, \pi)}\}$
      Else
         i. Set $\mathbf{R} := \tilde{\mathbf{R}}$ with probability $\min\{1, \frac{P(\tilde{\mathbf{R}}, \tilde{\pi})}{P(\mathbf{R}, \pi)}\}$
4. Return $\mathbf{R}$
Fast Bayesian IRL

Modular rewards are scaled

Search the space of scale factors to make observed data most probable

\[ P(s_j, a_j | Q^*) = \frac{1}{Z} e^{\sum_i c_i Q(s_j^{*(i)}, a_j)} \]
Variability in trajectories due to variability of initial heading angle insteps of 10 degrees
Variability in trajectories due to variability in reward weight of obstacles reward in steps of 0.1 reward (renormalizing other reward weights)
Variability in trajectories due to variability in reward weight of litter reward in steps of 0.1 reward 
(renormalizing other reward weights)
Driving in a Virtual World
Multi-tasking revealed by gaze sharing in human data

Shinoda & Hayhoe, Vision Research 2001
Human Driving in VR: “follow” condition

Driver must avoid pedestrians, other cars while following the red car.
Robot driving with RL modules

RoboCar is constrained to stay in a selected lane but may decide to change lanes.
RoboCar has to follow the red car.
RoboCar has to avoid other cars.
Steering and Acceleration are separate modules
Q-tables

Steering and Acceleration are separate modules
Summary of Part II:
RL modules lead to computational efficiencies

Credit assignment – modules keep running estimates

Fast IRL – modules reduces the calculations to simple parameter estimation
Thanks to...

Nathan Sprague
Weilie Yi
Constantin Rothkopf

Dmitry Kit
Joseph Cooper
Rahul Iyer
Leif Johnson
Lewis Fishgold
Brian Sullivan