Integrating Reinforcement Learning with Models of Representation Learning

Matt Jones (mcj@colorado.edu) & Fabián Cañas (canas@colorado.edu)
University of Colorado, Department of Psychology & Neuroscience
Boulder, CO 80309 USA

Abstract

Reinforcement learning (RL) shows great promise as a model of learning in complex, dynamic tasks, for both humans and artificial systems. However, the effectiveness of RL models depends strongly on the choice of state representation, because this determines how knowledge is generalized among states. We introduce a framework for integrating psychological mechanisms of representation learning that allows RL to autonomously adapt its representation to suit its needs and thereby speed learning. One such model is formalized, based on learned selective attention among stimulus dimensions. The model significantly outperforms standard RL models and provides a good fit to human data.

Keywords: Reinforcement learning; attention; generalization

Introduction

Most challenging tasks people face are inherently dynamic and interactive. Choices affect not just immediate outcomes but also future events, and hence subsequent decisions that must be made. Normative and descriptive theories of learning in dynamic environments have advanced dramatically in recent years with the development of Reinforcement Learning (RL), a mathematical and computational theory drawing on machine learning, psychology, and neuroscience (e.g., Schultz, Dayan, & Montague, 1997; Sutton & Barto, 1998).

However, RL currently faces a fundamental challenge relating to the issue of knowledge representation. Dynamic tasks tend to be highly complex, with an enormous number of possible states (situations) that can arise. Therefore, efficient learning must rely on generalization from past states that are similar to the current one. Similarity, in turn, depends on how states are represented, including the features by which they are encoded and the relative attention allocated to those features (Medin, Goldstone, & Gentner, 1993). Thus representation is critical to the effectiveness of RL algorithms, because representation determines the pattern of generalization by which past experience is used to make new decisions.

Although there has been little research on representation in the context of RL, representation and representation learning have long been topics of psychological study. Empirical research in a number of domains, including perceptual learning, attention, categorization, object recognition, and analogy, has uncovered principles and mechanisms by which people learn to modify how they encode objects and situations in the service of learning, inference, and decision making. Here we describe a framework for a natural synthesis of these ideas with RL algorithms, which leads to models that learn representations for dynamic tasks. A specific model is presented that is based on principles of attention learning from the categorization literature (Kruschke, 1992; Nosofsky, 1986). Two sets of simulation studies are reported, which demonstrate both the power and the psychological validity of this approach.

Reinforcement Learning

RL comprises a family of algorithms for learning optimal action in dynamic environments. RL models characterize a task as a Markov Decision Process, in which the environment at any moment exists in one of a set of *states*, each associated with a set of actions available to the learner. The chosen action determines both the immediate reward received, if any, and the state of the environment on the next time step. This general framework encompasses most tasks of psychological interest (Sutton & Barto, 1998).

The key insight behind most RL algorithms is to learn a *value* for each possible state or action. This value represents the total future reward that can be expected starting from that point. Formally, given any state s and action a, the state-action value is defined as

$$Q(s,a) = E\left[\sum_{k\geq 0}\gamma^k r_{t+k} \middle| s_t = s, a_t = a\right], \quad (1)$$

where t is the current timestep; s, a, and r are the state, action, and received reward on each step; and $\gamma \in [0, 1]$ is a discount factor representing the relative value of immediate versus delayed rewards. This approach allows action selection to be based directly on the Q-values. Here we use a Luce-choice or Gibbs-sampling rule, with inverse-temperature parameter θ.

$$P(a_t = a) \propto e^{\theta \cdot Q(s_t,a)} \quad (2)$$

Once an action is selected, its value is updated according to the immediate reward and the values associated with the state that follows. One of the best-studied algorithms for learning action values, Q-learning (Watkins & Dayan, 1992), uses the update rule

$$\Delta Q(s_t,a_t) = \varepsilon_{\text{val}} \cdot \delta, \quad (3)$$

where $\varepsilon_{\text{val}} \in (0, 1]$ is a learning rate, and δ is the *temporal-difference (TD) error*, defined as

$$\delta = r_t + \gamma \cdot \max_a \{Q(s_{t+1},a)\} - Q(s_t,a_t). \quad (4)$$

The TD error represents the difference between the original action-value estimate, $Q(s_t, a_t)$, and a new estimate based on the immediate reward and ensuing state. The expression $\max_a\{Q(s_{t+1}, a)\}$ represents an estimate of the value of the new state, s_{t+1}, based on the best action that could be performed in that state.

The simplest implementations of *Q*-learning and other RL algorithms use a *tabular* (i.e., lookup-table) representation, in which a different set of action values is learned separately for every possible state that can occur. Tabular representations are impractical for most realistic tasks, because the number of states grows exponentially with the number of state variables. Therefore, most implementations of RL utilize some form of *generalization*, whereby knowledge about one state is extended to other, similar states. This approach dramatically speeds learning by reducing the amount of information that must be retained and updated, and by allowing the learner to draw on a richer set of past experiences when making each new decision.

Central to the success of generalization in all learning tasks (not just RL) is the choice of representation. In order for generalization to be effective, states (or stimuli in general) must be encoded so that stimuli that are perceived or treated as similar tend to be associated with similar outcomes or appropriate actions (Shepard, 1987). Such a representation facilitates generalization, and hence learning, because it leads the learner to draw on precisely those past experiences that are most relevant to the current situation.

Unfortunately, the choice of representation is a notoriously difficult problem, and the field of machine learning is far from having automated algorithms that discover useful representations for learning novel tasks. Successful applications of RL have instead tended to rely on hand-coded human knowledge for encoding states. For example, the state representation in Tesauro's (1995) celebrated backgammon program, TD-gammon, was based on complex features (configurations of pieces) suggested by expert human players. Likewise, psychological research in RL generally avoids the problem of representation by using small sets of stimuli with clearly defined features, so that the subject's representation can be confidently assumed by the modeler and is unlikely to change during the course of learning (e.g., Fu & Anderson, 2006). Arguably, representation is where the real challenge often lies, and therefore starting a model with a hand-coded representation, or using experimental stimuli with unambiguous features, presupposes the most difficult and interesting aspects of learning (Schyns, Goldstone, & Thibaut, 1998).

Selective Attention in Category Learning

One behavioral domain in which generalization and representation have been extensively studied is category learning. Much of the research on category learning has aimed to understand the internal representations that humans develop to facilitate classification of objects and inference of unobserved features. All of these models serve, in one way or another, to allow generalization of category knowledge from previously encountered to novel stimuli.

The most direct mechanism for generalization in categorization is embodied by exemplar models (Medin & Schaffer, 1978). In these models, the psychological evidence (E) in favor of classifying a stimulus (s) into a given category (c) is given by summing its similarity to all previously encountered exemplars (s'), weighting each exemplar by its association to c.

$$E(s,c) = \sum_{s'} sim(s,s') \cdot w(s',c) \qquad (5)$$

The property of exemplar models most relevant to the current investigation is the similarity function. Rather than being fixed, a large body of evidence indicates that similarity changes during the course of learning as a consequence of shifts of attention among the stimulus dimensions (e.g, Nosofsky, 1986). This flexibility is modeled by expressing similarity as a decreasing function of distance in psychological space, with each stimulus dimension, i, scaled by an attention weight, α_i (Nosofsky, 1986). Here we assume an exponential similarity-distance function, in accord with empirical evidence and normative Bayesian analysis (Shepard, 1987).

$$sim(s,s') = e^{-\sum_i \alpha_i |s_i - s'_i|} \qquad (6)$$

The effect of attention on similarity is to alter the pattern of generalization between stimuli so as to fall off more rapidly with differences along dimensions with greater attention weights. When stimuli differ only on unattended dimensions, their differences are unnoticed and hence generalization between them is strong. This adaptation of generalization leads to improved performance when attention is shifted to task-relevant dimensions, because the learner generalizes between stimuli that have common outcomes while discriminating between stimuli that are meaningfully different.

The influence of attention on generalization has extensive support, both theoretically (Medin et al., 1993) and empirically (Jones, Maddox, & Love, 2005; Nosofsky, 1986). An important question suggested by this research is how attention can be learned. One proposal is that attention weights are updated in response to prediction error (Kruschke, 1992). In a classification task, prediction error (δ) is simply the difference between the category evidence, $E(s, c)$, and the actual category membership given as feedback to the learner (e.g., +1 if $s \in c$ and -1 otherwise). The updating rule for attention is then based on gradient descent on this error, squared and summed over categories.

$$\Delta\alpha_i = -\varepsilon_{\text{att}} \cdot \frac{\partial}{\partial \alpha_i} \left(\frac{1}{2} \sum_c \delta_c^2 \right). \qquad (7)$$

This mechanism for attention learning has been implemented in ALCOVE, a highly successful model of human category learning (Kruschke, 1992). ALCOVE learns to shift attention to stimulus dimensions that are most relevant to predicting category membership and away from dimensions that are non-diagnostic. This leads to adaptation of generalization, which in turn speeds learning.

Attention Learning in RL

Because of the strong empirical support for attention learning in the categorization literature, we believe it is a potentially fruitful topic for study in the context of RL.

Selective attention may be especially relevant in this domain because most interesting RL tasks have complex state spaces of high dimensionality, and learning to distinguish relevant from irrelevant dimensions should be expected to greatly speed learning in such tasks.

The present investigation addresses two questions regarding the relationship between attention learning and RL. The first question is a psychological one, of whether attention learning as observed in supervised tasks such as categorization also operates in the dynamic tasks modeled by RL. This extrapolation is not trivial, because RL relies on TD error, which is an internally generated signal based in part on the learner's own value estimate of the ensuing state (see Eq. 4). It is an empirical question whether this internal signal can drive attention shifts and other forms of representation learning in the same way that external feedback does. A companion paper (Cañas & Jones, 2010) reports a behavioral experiment that supports an affirmative answer to this question, and the data from that experiment are modeled below.

The second question is a computational one, of whether the formal equations that describe RL and attention learning can be coherently integrated, and whether the resulting model will exhibit efficient learning. This normative question is important psychologically because computational power constitutes a significant motivation for expecting attention learning to play a role in human RL. If the two are computationally compatible, then the potential significance of RL is greatly increased, in that RL is capable of autonomously adapting the representations on which it operates.

Comparison of the equations describing Q-learning and attention learning reveals there is indeed a natural, highly complementary integration. The strength of Q-learning, and RL algorithms in general, is in the sophisticated updating signals they compute, which take into account both external reward and internal consistency of value estimates (Eq. 4). The updating itself is fairly trivial, consisting of adjusting the existing estimate by a proportion of the error (Eq. 3). Attention learning, and models of category learning more generally, have the opposite character. Their updating signals are fairly trivial (prediction error relative to external feedback), but the updates themselves are complex, driving adaptation of sophisticated internal representations. This complementary relationship suggests the solution of using the TD error signal from RL to drive representation learning, and in particular to update attention weights.

We refer to the model resulting from this integration as Q-ALCOVE. Q-ALCOVE estimates action values via similarity-based generalization among states, directly analogous to ALCOVE (Eq. 5).

$$Q(s,a) = \sum_{s'} sim(s,s') \cdot w(s',a) \tag{8}$$

The Q-values are used to generate action probabilities according to the response-selection rule used by both Q-learning and ALCOVE (Eq. 2). The w parameters, which act as pre-generalization action values, are updated analogously to both Q-learning (Eq. 3) and ALCOVE, using the same TD error signal as in Q-learning (Eq. 4).

$$\Delta w(s_t, a_t) = \varepsilon_{\text{val}} \cdot \delta \tag{9}$$

Similarity between states in Q-ALCOVE is defined identically to stimulus similarity in ALCOVE (Eq. 6), except that a normalization term is included that fixes the total similarity (i.e., the integral of the generalization gradient) to 1. We have found that attention learning in tasks requiring continuous prediction only functions well when normalization is included.

Learning of attention weights follows the same rule as in ALCOVE, except for the critical substitution of classification error with TD error. In addition, we only differentiate δ with respect to $Q(s_t, a_t)$ and not $Q(s_{t+1}, a)$, because the motivation behind Q-learning is to use $Q(s_{t+1}, \cdot)$ to adjust $Q(s_t, \cdot)$. Nevertheless, changing α also affects $Q(s_{t+1}, \cdot)$, and further analytical work is needed to understand the impacts of this fact on model behavior and predictions. The resulting rule for attention learning is

$$\Delta\alpha_i = \varepsilon_{\text{att}} \cdot \delta \cdot \frac{\partial}{\partial\alpha_i} Q(s_t, a_t) \tag{10}$$

The intuition behind attention learning in Q-ALCOVE is that, after feedback, the model adjusts attention weights to reduce generalization from states that contributed to error and to increase generalization from states that suggested more correct predictions. Over the course of experience, attention should shift to those dimensions that are most diagnostic of correct actions and their values.

Simulation Studies

Two sets of simulations were carried out to evaluate the behavior of Q-ALCOVE. The first set was based on Gridworld, a common benchmark task for RL models. These simulations aimed to test whether the attention-learning mechanism in Q-ALCOVE operates as predicted, to shift attention toward relevant dimensions and away from irrelevant dimensions. If so, a second question was whether selective attention leads to significant improvements in learning speed, and how such a potential advantage depends on the dimensionality of the task. The second set of simulations was based on a human behavioral experiment (Cañas & Jones, 2010) designed to test whether humans can learn selective attention using internal value estimates (i.e., TD error) as feedback, as proposed here. These simulations aimed to evaluate Q-ALCOVE's viability as a psychological model.

Directional Gridworld

Gridworld is a class of tasks with a long tradition as a benchmark for RL algorithms (e.g., Sutton & Barto, 1998). The states of a Gridworld task form a rectangular lattice of dimensionality D. We call the present task Directional Gridworld, because it was set up in such a way that one dimension was relevant and the others were irrelevant.

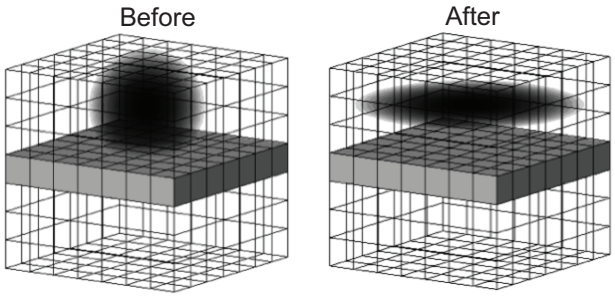

Figure 1. State space for 3-dimensional Directional Gridworld task. Grey states are goal states. Black cloud depicts Q-ALCOVE's generalization gradient, at the start of learning (left) and after 300 time steps (right).

Figure 1 illustrates the Directional Gridworld task for the case of $D = 3$ (the generalization gradients in the figure are discussed below). Each dimension has 7 levels, for a total of 7^D states. In each state, the learner has $2D$ available actions, corresponding to motion in either direction along any dimension. For simplicity, we assume that actions are deterministic and move the learner by 1 step in the chosen direction. Actions on the boundaries that would take the learner outside the space have no effect.

States are encoded as vectors corresponding to their values on the D dimensions. Other than this, the model has no prior knowledge of the topology of the environment or of the meanings or effects of actions. The spatial interpretation is only a convenient metaphor, and the task is not meant as a model of spatial navigation that might involve specialized psychological mechanisms. The stricter interpretation is as an abstract problem space (e.g., Newell & Simon, 1972).

The highlighted states (Fig. 1) spanning the center of the space are *goal states*. Whenever the learner reaches a goal state, a reward of 10 is provided. On the next step, the learner is taken to a random state maximally distant from the goal region. All actions that do not lead to a goal produce a reward of -1. The learner's task is to choose actions so as to maximize total temporally discounted reward (Eq. 1, with γ set to .5). Thus, optimal behavior consists of repeatedly moving in a straight line from the boundary to the nearest goal state.

For all values of the dimensionality D, the goal states form a hyperplane through the center of the space. The dimension perpendicular to the goal region is relevant to optimal action selection, as the learner needs to move in opposite directions depending on which side of the goal region it is on. All other dimensions are irrelevant. Indeed, it can easily be shown that the optimal Q-values for any state depend only on the state's position on the relevant dimension. Therefore, the most efficient generalization strategy for learning Q-values is to average over all states at each level of the relevant dimension but to learn separate values for each level. This strategy can be achieved by strong attention to the relevant dimension and zero attention to all other dimensions. A primary question was whether Q-ALCOVE would learn such an attention distribution.

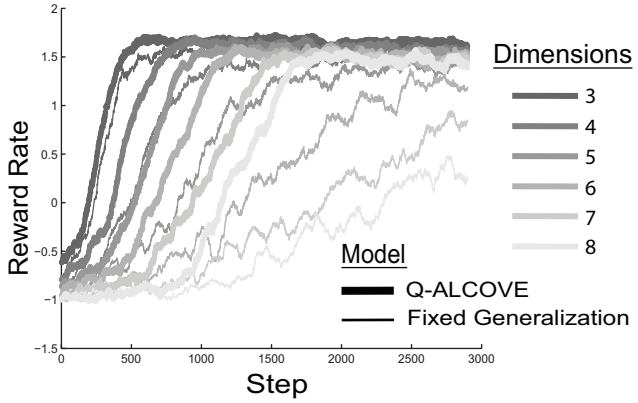

Figure 2. Learning curves in Directional Gridworld for Q-ALCOVE and version with fixed generalization.

Two models were simulated in addition to Q-ALCOVE. The first was tabular Q-learning, which learns actions values independently for all states. The second was a fixed-generalization model obtained from Q-ALCOVE by setting the attention-learning rate, ε_{att}, to 0. Q-ALCOVE was run using $\varepsilon_{att} = .01$. All models were run with value-learning rate $\varepsilon_{val} = 1$ and choice parameter $\theta = .5$. The models' value estimates (w or Q) were initialized at 0 at the start of each run. The initial values for attention weights were set to .4 for both Q-ALCOVE and the fixed-generalization model. This value was chosen so as to maximize performance of the fixed-generalization model on 3 dimensions.

Figure 2 shows average learning curves for Q-ALCOVE and the fixed-generalization model for Directional Gridworlds of 3 to 8 dimensions. Performance for tabular Q-learning was poor enough, especially at higher dimensionalities, that it is omitted. Each curve indicates reward rate, smoothed with a rectangular window of 100 time steps, and averaged over 4 separate runs of the model. As can be seen, Q-ALCOVE learns more quickly with attention learning than without, and the magnitude of this advantage grows rapidly with the number of dimensions. This result suggests that attention plays an indispensable role in natural tasks of much higher dimensionality.

Figures 1 and 3 illustrate how Q-ALCOVE's attention-learning mechanism facilitates learning, in the case of three dimensions. Figure 3 shows the attention weights for a single run, which increase for the relevant dimension and decrease toward 0 for the irrelevant dimensions. This shift of attention leads to the change in the generalization gradient depicted in Figure 1. The initial gradient (left) is spherical, reflecting the model's lack of knowledge of the dimensions' predictive validities. After 300 time steps (right), the gradient has been reshaped so that there is strong generalization between states as long as they match on the relevant dimension and very little generalization otherwise. Thus the model has learned the anisotropy of the task, which allows it essentially to estimate a common set of Q-values for all the states at each stratum (as an average over all the ws), while keeping the values for different strata separate.

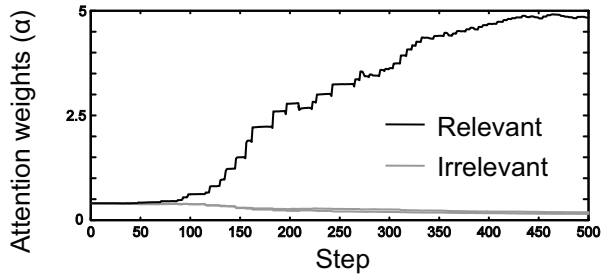

Figure 3. Dynamics of attention weights for one run of Q-ALCOVE in 3-dimensional Directional Gridworld.

The consequences of all three models' patterns of generalization are illustrated in Figure 4, which shows a two-dimensional slice through the center of the three-dimensional state space. Within each state, the four arrows indicate the model's estimated *Q*-values for the four actions within the plane. These values are from a single run of each model, after 300 time steps. Darker arrows indicate greater *Q*-values. The values for tabular Q-learning are irregular, reflecting the fact that they were learned separately for each state. In most states there has not been enough experience to obtain reliable estimates. The *Q*-values estimated by the fixed-generalization model are more accurate, because each draws on knowledge from neighboring states, so experience is used more efficiently. However, there is still irregularity among states at a given stratum (insufficient generalization across the irrelevant dimension) and too much smoothing (excess generalization) along the relevant dimension. Q-ALCOVE's estimated *Q*-values are much more accurate, allowing the model to select correct actions more reliably.

The Spores Task

Psychologically, the core assumption of Q-ALCOVE is that attention learning can be driven by internally generated TD-error signals, not just overt feedback. A behavioral experiment, reported by Cañas and Jones (2010), tested this hypothesis using a two-step task, in which Action 1 determined Stimulus 2, but reward was not received until after Action 2. The basic question was whether selective attention to the dimensions of Stimulus 1 could be learned, when the only immediate feedback was the identity of Stimulus 2.

Figure 5 illustrates the design of the task. Stimulus 1 (a cartoon mushroom spore) varied along two dimensions and was sampled from a circular set. This set was probabilistically divided into two regions, which had different consequences for the outcome of Action 1 (two options for how to grow the spore). The border between regions was oriented so that one dimension was more relevant than the other. The second step was designed so that the two possibilities for Stimulus 2 (two colors of mushrooms) each had a different optimal choice for Action 2 (selling the mushrooms to a troll or a goblin). Under these optimal actions, Stimulus 2a led to more reward than Stimulus 2b.

RL models in general predict subjects will learn internal values for Stimuli 2a and 2b (or their pairings with choices of Action 2), and these values will be used to generate internal feedback (TD error) for Action 1. This will in turn

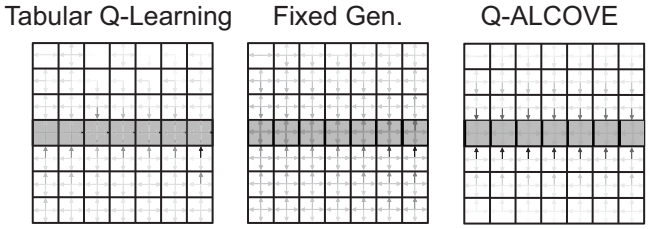

Figure 4. *Q*-values obtained from all three models after 300 steps in 3-dimensional Directional Gridworld. Shown is a 2-dimensional slice through the center of the state space. Arrows in each state correspond to the four actions within the plane. Darker arrows indicate greater *Q*-values.

allow subjects to learn to choose Action 1 so as to maximize the probability of obtaining Stimulus 2a (the more valuable mushroom). The key additional prediction of Q-ALCOVE is that TD error will also drive learning of attention to the more relevant dimension of Stimulus 1, to improve the effectiveness of generalization among stimuli.

Results revealed that subjects who learned the first step of the task also learned to selectively attend to the more relevant dimension (see Cañas & Jones, 2010, for details). Simulations of Q-ALCOVE corroborated this conclusion. Q-ALCOVE and the fixed-generalization version of the model were fit to the data of each subject using maximum likelihood. Aggregating over all 150 subjects, Q-ALCOVE fit reliably better, $\chi^2(150) = 1913.8$, $p \approx 0$. The difference between fits of the models was significant at the .05 level for 55 individual subjects. These results support the central hypothesis that attention learning, as embodied by Q-ALCOVE, was involved in learning the task.

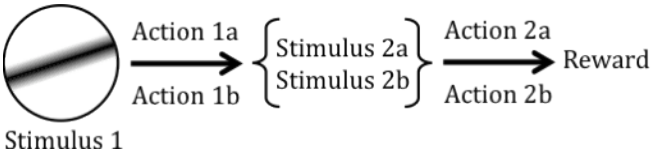

Figure 5. Structure of the Spores task.

Conclusions

Despite its computational power and neurological support, the basic principles behind RL are inherently limited by the representations it operates on. We argue here for a tight linkage between RL and mechanisms of representation learning established in other domains of psychology. Specifically, TD error, the engine behind nearly all RL models, can also drive updating of state representations. Representations thereby adapt so the pattern of generalization among states is tuned to the structure of the task, which in turn facilitates learning of optimal actions. RL's capacity to autonomously drive construction of representations that serve its needs greatly increases its power and flexibility, and hence its potential as a model of complex human learning.

The specific model proposed here draws on principles of selective attention from category learning and related domains (Nosofsky, 1986; Sutherland & Mackintosh, 1971). Shifting attention away from irrelevant dimensions allows

the learner to aggregate knowledge over states with similar outcomes, while attention toward relevant dimensions maintains discrimination of meaningful differences. Generalization in any learning task raises a bias-variance dilemma, in that more generalization reduces variance in parameter estimates but increases their bias. Selective generalization as modeled by attention learning is an elegant way of sidestepping this dilemma.

In a companion paper, we report empirical evidence supporting attention learning via TD error as a psychological mechanism (Cañas & Jones, 2010). Here we show how such a mechanism can be formalized in a mathematical model. Attention learning and RL in this model bootstrap off of each other, in that the internal value estimates generated by RL drive shifts of attention, and selective attention in turn improves RL's value estimates. This synergistic relationship, together with the elegance of the integration between the equations of Q-learning (Watkins & Dayan, 1992) and ALCOVE (Kruschke, 1992), suggests that RL and attention learning are similarly tightly coupled in the brain. The simulation studies reported here show that the unified model, Q-ALCOVE, is both computationally powerful and psychologically plausible.

Investigating attention is a useful first step because it acts to modify similarity directly, so that its effects on generalization are transparent. In further work, we plan to explore more complex psychologically supported mechanisms, such as stimulus-dependent attention (Aha & Goldstone, 1992), construction of new conjunctive features (Love, Medin, & Gureckis, 2004), and analogical mapping between structured stimulus representations (Markman & Gentner, 1993).

Psychological models that generalize knowledge based on pairwise similarity are closely related to kernel methods developed in statistics (Jäkel, Schölkopf, & Wichmann, 2007). Kernel methods add considerable flexibility to many learning algorithms, by allowing them to be recast from the raw stimulus space to a mathematical (Hilbert) space of functions (e.g., Cristianini & Shawe-Taylor, 2000). Q-ALCOVE can be viewed as a kernel method applied to RL. Viewed from the perspective of kernel methods, an important contribution of the present research is the proposal for adaptively modifying the kernel (i.e., generalization gradient) to improve learning. Learning the kernel has been a focus of recent research in machine learning (e.g., Micchelli & Pontil, 2007), but results thus far have been largely limited to existence theorems and global-search algorithms that seem psychologically implausible. Here we propose a simpler mechanism based on psychological principles. The mathematical results on kernel learning have been influential in guiding our design of well-behaved models and in inspiring more sophisticated mechanisms. Continuing to exploit this link to statistical and machine-learning techniques, while maintaining grounding in established psychological phenomena, seems promising for advancing the power and flexibility of psychological models.

References

Aha DW & Goldstone RL (1992). Concept learning and flexible weighting. *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, 534-539.

Cañas F & Jones M (2010). Attention and reinforcement learning: Constructing representations from indirect feedback. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Cristianini N & Shawe-Taylor J (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Fu W-T & Anderson JR (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, *135*, 184-206.

Jäkel F, Schölkopf B & Wichmann FA (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology, 51*, 343-358.

Jones M, Maddox WT & Love BC (2005). Stimulus generalization in category learning. *27th Annual Meeting of the Cognitive Science Society*, 1066-1071.

Kruschke JK (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Love B, Medin D & Gureckis T (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.

Markman AB & Gentner D (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.

Medin DL, Goldstone RL & Gentner D (1993). Respects for similarity. *Psychological Review, 100*, 254-278.

Medin DL & Schaffer MM (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238.

Micchelli CA & Pontil M (2007). Feature space perspectives for learning the kernel. *Machine Learning, 66*, 297-319.

Newell A & Simon HA (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Nosofsky RM (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Schultz W, Dayan P & Montague P (1997, March). Neural substrate of prediction and reward. *Science, 275*, 1593-1599.

Schyns PG, Goldstone RL & Thibaut J-P (1998). Development of features in object concepts. *Behavioral and Brain Sciences, 21*, 1-54.

Shepard RN (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Sutherland NS & Mackintosh NJ (1971). *Mechanisms of Animal Discrimination Learning*. NY: Academic Press.

Sutton R & Barto A (1998). *Reinforcement Learning: An Introduction*. The MIT Press.

Tesauro G (1995). Temporal difference learning and TD-gammon. *Communications of the ACM, 38*(3), 58-68.

Watkins CJCH & Dayan P (1992). Q-Learning. *Machine Learning*, *8*, 279-292.