COGNITIVE ARCHITECTURES FOR INTROSPECTING DEEP REINFORCEMENT LEARNING AGENTS

Konstantinos Mitsopoulos, Sterling Somers & Christian Lebiere
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{cmitsopoulos,cl}@cmu.edu, sterling@sterlingsomers.com

Robert Thomson
Army Cyber Institute
Unites States Military Academy
West Point, NY, 10996, USA
robert.thomson@westpoint.edu

ABSTRACT

In this work we demonstrate the use of Cognitive Architectures in the Reinforcement Learning domain, specifically to serve as a common ground to understand and explain Reinforcement Learning agents in Human Ontology terms. Cognitive Architectures could potentially act as an adaptive bridge between Cognition and modern AI, sensitive to the cognitive dynamics of human user and the learning dynamics of AI agents.

1 WHY COGNITIVE ARCHITECTURES

Cognitive models leverage scalable and efficient learning mechanisms, capable of replicating human sensemaking processes to transform loosely-structured data from heterogeneous sources into a more structured form in a manner similar to human learning (Lebiere et al., 2013). Cognitive models can be instantiated within a cognitive architecture. A cognitive architecture is a computational framework representing cognitive processes including attention, memory, and pattern-matching, organized in a structure analogous to that of the human brain (Laird et al., 2017). It provides a principled platform for developing behavior models constrained by human processes and predictive of human performance, but also with the means to both reproduce and mitigate limitations in human reasoning such as cognitive biases. The cognitive models developed for this work use the ACT-R cognitive architecture (Anderson et al., 2004; Anderson, 2007). ACT-R is organized into modules for declarative memory, procedural knowledge, vision, audition, and motor control. In this project we use only the declarative memory module and, in particular, the instance-based learning (IBL) theory of decision making based on storage and retrieval processes from memory.

ACT-R is a hybrid architecture, composed of both symbolic and sub-symbolic processing. The hybrid nature of ACT-R makes for a particularly compelling candidate for bridging AI and Cognitive Science. The symbolic aspects of the architecture are inherently interpretable: the models not only fit data, but the introspectability of the models themselves provide an understanding of the representational content and cognitive processes underlying behavior. Further, the sub-symbolic components, which constrain the architecture, has similar characteristics of learning and adaptivity to machine learning models and thus has the potential to interface with other sub-symbolic systems like neural networks (Jilk et al., 2008).

2 METHODS

In order to showcase the usage of Cognitive Architectures along with an RL agent, in various settings, we developed a modular turn-based OpenAI Gym gridworld. For our purposes we designed

the following scenario: an RL agent (orange) and a predator (red) coexist in a 10×10 grid (figure 1). The agent's goal is to reach a target location (green) without getting caught. The agent can move up, down, left, right or do nothing ('noop'). The predator protects the goal by calculating an A* path from the player to the goal and moves itself to a square within that path. Second, if a player comes within 2 steps from the agent, the agent will attempt to chase the player. Finally, if the predator gets further than 3 steps from the goal, it will return to the goal to defend. There are two general solutions to reaching the goal: 1) attempt to evade the predator until it makes a mistake (it can chose an A* path when chasing that takes it out of the way of the goal) or 2) bait the predator from 2 steps away with a noop, causing the predator to attack the player. In that case, the predator chases the player but can never catch it unless it makes a mistake (e.g. run into a wall) - the A* only provides a path to the last player location, and as long as the player is not there on the next step, it will simply follow. An 'expert' strategy, therefore, is to do a 'noop' action to bait the agent, evade, and move towards the goal. By using this scenario, we focus on showing how Cognitive modeling of both agent and human behavior can lead to a better understanding of their strategies.

2.1 Reinforcement Learning

We consider the standard RL setting where an agent interacts with an environment in discrete timesteps. An agent receives an image 10x10 as input which is fed to a block of three convolutional layers (4×4,3×3,3×3 kernels and 3,1,1 strides respectfully). The resulted tensor is flattened and passed to a 2-layer MLP (512 units per layer) from which the policy and state-value function are estimated. The agent was trained with the A2C algorithm (Mnih et al., 2016).

2.2 Instance Based Model

Instance-based learning theory (Gonzalez et al.) is the claim that implicit expertise is gained through the accumulation and recognition of previously experienced events or instances. IBL was formulated within the principles and mechanisms of cognition in ACT-R, and makes use of the dynamics of knowledge storage retrieval from long-term memory. For instance, the dynamics of an instance's sub-symbolic activation (e.g., frequency and recency of presentation, similarity to other instances) provide a validated mechanism for determining which instances are likely to be retrieved for a given situation, and what factors came into play. Models of decision-making and problem-solving in ACT-R over the past 10 years have seen increasing use of IBL to account for implicit learning phenomena.

In IBL, assuming that we seek a decision d given a new state i, with context $\mathbf{f}$, the inferential process consists of three steps: 1) The current state is compared to all N instances in declarative memory and a matching score $M(i,j) \propto Sim(\mathbf{f}_i, \mathbf{f}_j)$ is assigned to each instance j in the memory; the matching score describes how similar the features $\mathbf{f}$ between the two instances are. 2) A recall probability $P_j = softmax(M(i,j))$ is used to estimate how probable it is for a memory instance to be recalled given the current state context. 3) Finally, a blending retrieval process returns a value chosen by the model, for that particular instance, as a weighted average of past decisions $u = \sum_{j=1}^{N} P_j d_j$. The weighted average form of the blending mechanism is derived as a special case of the general blending equation which can be viewed as an optimization process that returns the value that best satisfies conflicting pieces of knowledge stored in the memory. In the case where the decisions take continuous values the value u is assigned as the model's decision d. In the discrete case, and assuming that the decisions are one-hot encoded the resulted value is equal to the probability of selecting a specific decision. The final decision is the one that maximizes the value u, or in other words the most probable one.

3 Cognitive Modeling

The cognitive model is very simplistic, intended to mimic behavior, as opposed to making human-like rational decisions given some expected reward. It simply uses past behavior to guide future behavior, without considering the consequence of behavior. The goal here is to develop models that reproduce human and network behavior rather than generate independent behavior.

Instances in the cognitive model are comprised of symbolic state and action descriptions. An instance for this environment consists of a chunk structure containing the following slots: *angle to goal*, *distance to goal*, *angle to predator*, *distance to predator*, *distance to left wall*, *distance to right*

(a) Human preference

(b) Network preference

Figure 1: Trajectory preferences human vs. network.

wall, *distance to upper wall*, distance to lower wall; as well as one-hot encoded action descriptions: *left*, *right*, *up*, *down*, and *NOOP*.

Data is gathered for the model by recording either human[1] or artificial agent during free-play of the game. At each step of the game, the observation and the corresponding action is recorded. The raw data is converted into symbolic data and later loaded into the model's memory. For illustrative purposes we gathered 500 episodes from the network ('network model') and 500 episodes of an expert human player ('human model').

We did a coarse parameter search for both models, assessing the move-by-move predictions of the model with 80% of their data, testing on the remaining 20%. We fit the human model to human data with approximately 75% accuracy, and the network model to network data with approximately 85% accuracy. Move-for-move accuracy is not particularly telling, as different moves could result in the same success. For example, moving left then down, as opposed to down then left, may result in a different trajectory but may not change the overall outcome nor strategy (as you end up at the same point). Given those models, we then tested each during free play for 500 episodes and measured behavioral performance.

3.1 Model Results

We measured the success rate of the network, the human, the network model, and the human model. The network reached the goal 100% of the time; the human reached the goal 98% of the time, the network model 94% of the time, and the human model 94% of the time.

The behavioral results of the model are presented in figure 2. Figure 2a illustrates the last 7 steps of successful plays of the game for the network (solid, aqua), network-model (dashed aqua), human (solid, green), and human model (dashed, grey). The y-axes indicates distance (measured in moves), and the x-axis indicates steps. The vertical lines indicate the average step where the 'noop' was pressed. The graph illustrates a close fit with respect to trajectories for the network-model to the network and the human-model to the human. Although both the human and the network (and their respective models) both use a 'noop' strategy (see figure 2a,b), the results suggest a slight difference in preference for when to press the 'noop'. Specifically, the human trajectory suggests that the human prefers to move directly towards the predator (see figure 1a), whereas the network appears to prefer to move to a position diagonally to the predator (see figure 1b).

3.1.1 Cognitive Salience

In an analogue to using gradients to determine which pixels most influence a decision (Selvaraju et al., 2017), we take the derivative of the blending equation with respect to the features, to determine their influence on the decision. While the features in this illustrative work correspond to features of the environment we have shown in other work ((Somers et al., 2019)) the use of abstract features, such as abstract relationships between entities in the environment.

As illustrated in figure 3, the saliences can be helpful to explain, in a human-interpretable ontology, what features influence a decision. The plot (right, figure 3a), shows the salience of the: angle-to-goal, distance-to-goal, angle-to-predator, and distance-to-predator (respectively), and indicates exactly what we might expect: the angle to the goal has the most influence in determining the direction selected. In figure 3b, however, the salience for distance-to-predator has far more influence on choosing the 'noop' action. It is important to note that the salience calculation is model-agnostic.

[1]Human player data come from authors of the paper

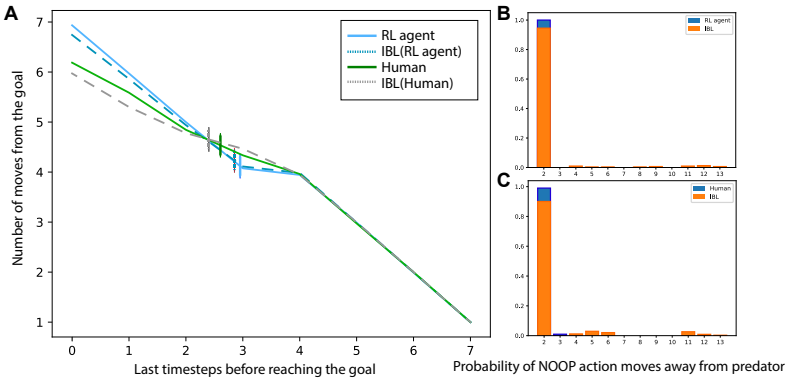
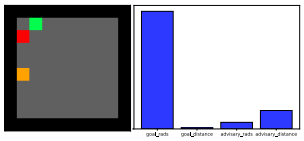
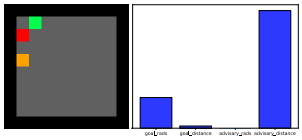


Figure 2: Behavioral fit of network model and human model to data

The image in 3 is from data collected from the network model but we expect similar saliences from a trace of the human model.



(a) Step before noop is chosen



(b) Step when noop is chosen

Figure 3: Contrasting salience examples, highlighting salience in noop decision.

4 DISCUSSION

This paper has focused mainly on specific use cases of modeling one-shot decisions in a cognitive architecture. Largely our work has been focused on explainable artificial intelligence (XAI) as the model provides a bridge between the complexities of the network and human-interpretable ontology terms. Cognitive architectures can play a large role, to that end, by elucidating differences between human understanding of a task and the AI understanding of the task. For example, a model that has knowledge of both human and AI play might be able to detect a difference in play preference, or strategies; and be able to anticipate the need for explanation from divergence in expectation between models of the AI and the human user, or help choose training examples to understand AI performance more efficiently, by reducing discrepancy between human model expectations and AI behavior.

Another area that our work extends is in Human-Machine collaboration domain. For example, we can have cognitive models that exhibit human-like behavior collaborating in a task with a RL agent. Another direction is to use cognitive model to mediate interaction between Human and RL agent. A cognitive model could provide an introspectable model of the agent to provide explanations of that agent’s behavior in terms understandable by the human agent. Conversely, a cognitive model of the human teammate could be used by the AI agent to generate predictions of human behavior to optimize its own actions.

The models presented here are meant to illustrate the use of cognitive models across domains. Although the task was quite simple, there is already a point of interest between cognition and AI. We have observed, in informal studies, the emergence of the ‘noop’ strategy in humans. Unlike how we imagine the learning trajectory of the network, human players likely exhibit one- or close-to-one -shot learning, after encountering the predator. It is very likely that a human player creates a mechanistic-style mental model that allows them to predict how the predator will act. Modeling the development of such a mental model would provide additional insights into the behavior of the human user that can in turn be reflected in the AI learner.

4.1 ACKNOWLEDGEMENTS

This research has been funded by DARPA contract FA8650-17-C-7710 and AFRL/AFOSR award FA9550-18-1-0251.

REFERENCES

John R. Anderson. *How Can The Human Mind Occur In The Physical Universe?* Oxford University Press, New York, NY, 2007. ISBN 9780195324259.

John R. Anderson, Dan J. Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 111(4):1036–60, oct 2004. ISSN 0033-295X. doi: 10.1037/0033-295X.111.4.1036.

Cleotilde Gonzalez, J.F. Lerch, and Christian Lebiere. Instance-based learning in dynamic decision making. *Cognitive Science*, 27:591–635.

David J Jilk, Christian Lebiere, Randall C O'Reilly, and John R Anderson. Sal: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 20(3):197–218, 2008.

John E Laird, Christian Lebiere, and Paul S Rosenbloom. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *Ai Magazine*, 38(4):13–26, 2017.

Christian Lebiere, Peter Pirolli, Robert Thomson, Jaehyon Paik, Matthew Rutledge-Taylor, James Staszewski, and John R Anderson. A functional model of sensemaking in a neurocognitive architecture. *Computational intelligence and neuroscience*, 2013.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Sterling Somers, Constantinos Mitsopoulos, Christian Lebiere, and Robert Thomson. Cognitive-level salience for explainable artificial intelligence. In *Proceedings of the 17th International Conference of Cognitive Modeling*, pp. 235–240, 2019.