Broadening and Building Beyond Classical Reinforcement Learning

Jacob Valdez

We humans have potential to self-actualize far beyond our immediate or even our own objectives, and our overall life experience is an open-book of this endeavour. However, such sacrifice is not as common in the world of deep reinforcement learning agents. Standard reinforcement learning agents practically embody the behavioralist fear-driven, classically-conditioned interpretation of behavior. Survival-oriented emotions like fear are certainly useful in momentarily narrowing focus to motivate brief behavioral responses. However, the spectrum of human affective experience is complemented by positive emotions such as curiosity, joy, and love, which though rarely directly serving any explicit objective, actually broaden and build on social, cognitive, and behavioral skills for many potential needs. I apply the Broaden and Build theory of positive emotions to the design and training of a novel deep reinforcement learning architecture Affective RL (AffectR) situated in a nonstationary lifelong, multimodal, multiagent, mixed-mode setting. The population is placed in progressively more human-like environments culminating in a photorealistic simulator where interventions are made when necessary to help each AffectR agent realize its full potential. Code: tinyurl.com/affectr

Motivation and Background

Emotion colors rich variation into the perceptual experience. It compliments deliberate rational yet bounded thought with a context-rich representation that weighs over indescribably many factors in guiding adaptive cognition and behavior, and the instinctive and learned context it carries stitches unifying threads into the social fabric of families, communities, and larger populations. What motivates farsighted and altruistic endeavor? The very term "motivate" almost seems to contradict the latter objective, yet in the dumpster of "intrinsic motivation", we find a range of emotions that promote unselfish activity. Consider: We feel the love of friends and family members. We've witnessed the compassion of healthcare workers during the pandemic. We read the joy of science in between the lines of academic journals, and many other touching emotions exert strong effects on our own lives.

Note that emotions motivating meaningful, self-actualizing behavior are not shortsighted survival-oriented ones. Escaping the hedonic treadmill, positive emotions take a broader, allocentric paradigm to behavior. Curiosity rarely has a known goal, yet it motivates open-ended discovery of useful skills and knowledge. Unselfish love deliberately places the individual's behavioral objective beyond self but inescapably reaps improved physical, mental, and social well-being. The broaden and build theory of positive emotion emphasizes that such positive emotions, though rarely directly serving any immediate survival oriented objective, actually facilitate broadening an individual's experience and building their social, cognitive, and behavioral skills for many potential needs.

Can this principle be applied to problem solving in general? I consider a subset of problem solving – reinforcement learning (RL) – where in the multiagent model-based deep RL setting, comparisons may reasonably be drawn between human affective experience and heuristics on RL agents. Formally, reinforcement learning models problem solving as a Markov decision process (S, A, P_a, R_a) where S is the state space, A is the action space, $P_a(s, s')$ is the probability that the environment will transition to state $s' \in S$ when action $a \in A$ is taken at state $s \in S$, and $R_a(s, s')$ is the reward received from a state-action trajectory s, a, s' and an accompanying learned action selection policy $\pi(a|s)$. Model-free approaches use some estimate of reward such as a Q-function of state and action $Q(s, a)$ while model-based RL learns a world-model such as $f(s'|s, a)$ or $f(s'|s)$ which it exploits to select actions optimally. These approaches are not mutually exclusive, and self-supervised, information-theoretic, and intrinsically motivated RL approaches exist which train using internally generated reward signals such as prediction, action entropy, and empowerment. In partially observable domains, the agent only receives a restricted observation o of the environment state s. Multiagent RL extends the classical agent-environment interaction process to many agents, often simultaneously learning and without a common observation or even objective as in the mixed-mode and competitive settings. Finally, deep RL uses neural networks to approximate an agent's Q-function, policy, world model, and other trainable functions.

RL systems leverage powerful conditioning techniques to optimize decision processes, yet their single objective paradigm contrasts to the motivational riches of human affective experience. Affect (colloquially, feeling or emotion) is often considered along three principal dimensions: arousal, valency, and motivational intensity. Arousal may be objectively measured by the degree of an individual's sympathetic axis activation, and in healthy individuals, increased arousal means greater potential to integrate information and respond appropriately. Valency measures individuals' subjective positive negative evaluation of a situation and is associated with the physiological dimension of pleasure. Motivational intensity describes an individual's subjective propensity to meaningfully act. I instill related characteristics in a novel RL system Affective RL (AffectR).

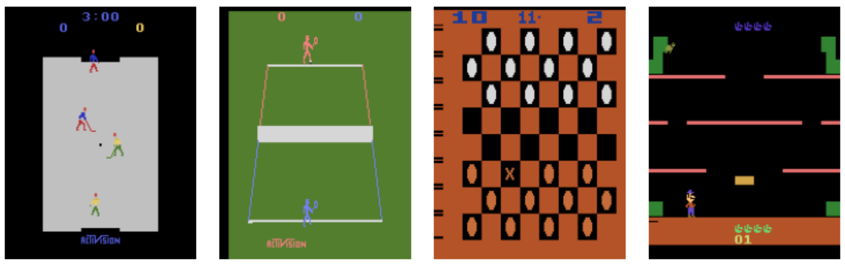


Figure 1. Simulator snapshots of select Atari games employed for multiagent reinforcement learning.

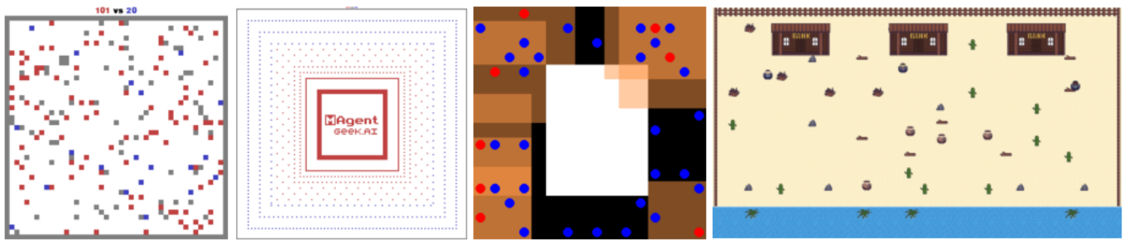


Figure 2.Select multiagent `PettingZoo` environments. Some of these environments are neither purely competitive nor cooperative but involve mixed-mode social interaction.

Methods

Each AffectR agent is composed from an encoder $f_{enc} : h, o \to h$, processor $f_{proc} : h \to h$, and decoder $f_{dec} : h \to a$ module. Internally, each module utilizes a residual dot-product attention mechanism to manipulate information. The encoder assigns partially learned embedding keys to each observation channel, and recurrently queries the observation space to update its representation. The processor performs self-attention similarly. The decoder extracts the agent's outputs with the same embedding keys as the encoder. These modules update asynchronously according to learned computation activity hyperparameters $\{\delta_{enc}, \delta_{proc}, \delta_{dec}\} \in \theta$. After running at time $t^{prev}_{\cdot}$, each module waits until its respective timeout has elapsed $t = t^{prev}_{\cdot} + \delta_{\cdot}$ before proceeding to apply its output. This is differentiably implemented as:

$$h \leftarrow \Theta(t - (t^{prev}_{enc} + \delta_{enc}))f_{enc}(h, o) + (1 - \Theta(t - (t^{prev}_{enc} + \delta_{enc})))h$$
$$h \leftarrow \Theta(t - (t^{prev}_{proc} + \delta_{proc}))f_{proc}(h) + (1 - \Theta(t - (t^{prev}_{proc} + \delta_{proc})))h$$
$$a \leftarrow \Theta(t - (t^{prev}_{dec} + \delta_{dec}))f_{dec}(h) + (1 - \Theta(t - (t^{prev}_{dec} + \delta_{dec})))a$$

where $\Theta(\tau) = \sigma(\beta\tau)$ and each respective timeout updates by $t^{prev}_{\cdot} \leftarrow \Theta(t - (t^{prev}_{\cdot} + \delta_{\cdot}))t + (1 - \Theta(t - (t^{prev}_{\cdot} + \delta_{\cdot})))t^{prev}_{\cdot}$.

Agents receive egocentric vision V_t, local area text T_t, loss information $\mathcal{L}_t$, hyperparameters θ_t, and the previous action a_t as input. Agents output predicted vision V^{pred}_{t+1}, predicted text T^{pred}_{t+1}, predicted loss information $\mathcal{L}^{pred}_{t+1}$, new hyperparameters θ_{t+1}, and the environment action a_{t+1} to be taken. Predicted vision, text, and loss information are subtracted against that actually given by the environment.

$$V_t = V^{ext}_t - V^{pred}_t$$
$$T_t = T^{ext}_t - T^{pred}_t$$

Loss is increased by the overall previous predictive error and activity levels and is supplied as a multidimensional tensor including individual components $\mathcal{L}^{ext}_{t,0:7}$ and the weighted sum $\mathcal{L}^{ext}_{t,7}$ of those components.

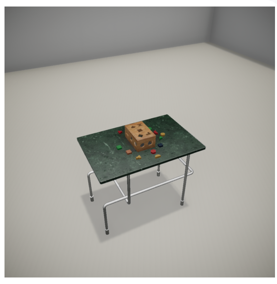
$$\mathcal{L}^{ext}_t = [l^{env}_t; reduce\_sum(V_t); reduce\_sum(T_t); reduce\_sum(\mathcal{L}_{t-1}); \Theta(t - (t^{prev}_{enc} + \delta_{enc})); \Theta(t - (t^{prev}_{proc} + \delta_{proc})); \Theta(t - (t^{prev}_{dec} + \delta_{dec})); \mathbf{c}_l \cdot \mathcal{L}^{ext}_{t,0:7}]$$
$$\mathcal{L}_t = \mathcal{L}^{ext}_t - \mathcal{L}^{pred}_t$$

Agents run batch gradient descent to minimize weighted loss sum on-policy after every 64 timesteps. Gradients are only allowed to flow through 16 computation steps. This means a staggered buffer of 80 timesteps are actually required to differentiate through 16 for 64 full frames. Unless specified, weights are not shared between agents. The environment action output layer changes with changing environments for individual agents.

This work is an active project and the AffectR architecture is still being implemented. From pretraining to multiagent simulation, agents and population will train in progressively more complex environments. Several simulator environments have already identified for training including Atari games (Figure 1), `PettingZoo` simulations (Figure 2), and `ThreeDWorld` scenes (Figure 3). Importantly, these environments are amenable to both quantitative metrics and qualitative human observation introducing the possibility of directly interacting with agents. I plan to conduct qualitative perturbation analysis on the agent and population level behaviors by both directly controlling a simulator agent and communicating to agents with the text modality.

The text modality represents a general communication channel that physically neighboring agents can interact through. I plan to pretraining agents individually on English corpera before introducing them to the AffectR population. Since agents are pretrained using natural language, I also plan to personally interact with AffectR individuals over their development to qualitatively asses their behavioral, cognitive, and personality characteristics.

Figure 3. Test scene from Three3World. Reused within the BSD 2-clause license from https://github.com/threedworld-mit/tdw/blob/master/Documentation/getting_started.md.



Discussion

This work draws many parallels between affective psychology and multi-agent deep reinforcement learning. Valency is expressed by the multidimensional loss signal $\mathcal{L}$. Encoder and processor activity rates may compare to arousal while processor and decoder activity rates liken to motivational intensity. Future developments will continue to entertain similarities between the two fields of study and apply the intrinsic motivations so instrumental to meaningful human endeavour to powerful yet socially-aware affective computing systems.

This work is personally funded and intrinsically motivated.