How to Learn and Represent Abstractions: An Investigation using Symbolic Alchemy

Badr AlKhamissi[1], Akshay Srinivasan[1], Zeb Kurth-Nelson[2], Sam Ritter[2]

[1]Sony AI | [2]DeepMind

Abstract

Alchemy is a new meta-learning environment rich enough to contain interesting abstractions, yet simple enough to make fine-grained analysis tractable. Further, Alchemy provides an optional symbolic interface that enables meta-RL research without a large compute budget. In this work, we take the first steps toward using *Symbolic Alchemy* to identify design choices that enable RL agents to learn various types of abstraction. Then, using a variety of behavioral and introspective analyses we investigate how our trained agents use and represent abstract task variables, and find intriguing connections to the neuroscience of abstraction. We conclude by discussing the next steps for using meta-RL and Alchemy to better understand the representation of abstract variables in the brain.

The Alchemy Benchmark

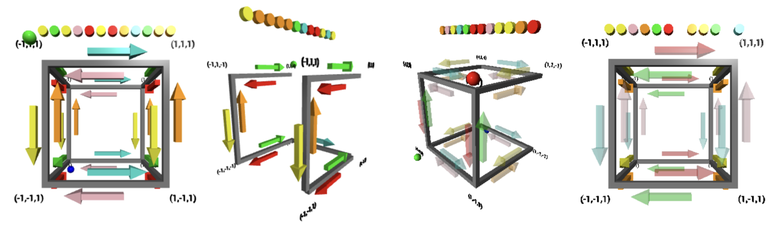

Figure: Visualizations of chemistries created using the Symbolic Alchemy Visualization Tool, which we are releasing for public use. The coordinates on the cube's vertices indicate the latent state and thus reward of the stone in that position. Translucent arrows indicate that the agent has not yet discovered the effect of that color. The stones are represented by a red, blue and green spheres. The available potions are shown floating above the cube with their corresponding color, and when a potion is consumed it will no longer be visible. In some chemistries, edges of the graph may be missing as can be seen in the 2nd and 3rd snapshots.

Stones and Potions The goal of the agent within each trial is to transform three stones into a more valuable form by applying a sequence of different potions on each of them. Each potion is characterized by one of 6 hues that dictates the transformative effect it has on a specific stone according to some `chemistry' that is sampled at the beginning of each episode.

Chemistry The chemistry dictates the causal structure that governs each episode. It can be visualized as a cube, where each vertex correspond to a specific stone value in the latent space.

Contributions

1. We show a way in which researchers can achieve high performance in Symbolic Alchemy -- albeit with some tricks -- without having access to vast computational resources as is usually required for deep-RL research.
2. We propose a hypothesis based on empirical results into why previous agents failed to solve Symbolic Alchemy despite being much more powerful than the one used in this work.
3. We release a tool for visualizing the chemistry and latent space of any given episode in Symbolic Alchemy to better help researchers debug the behavior of their agents.
4. We present a new kind of behavioral analysis that can be done on Alchemy to test whether the agent succeeded or failed in acquiring specific pieces of abstract knowledge.
5. Finally, we demonstrate that, just as in animals, single-units of the LSTM and transformer encode abstract task variables. Moreover, single-unit analyses revealed evidence for distinct functional roles for LSTM and transformers units. We draw a connection between this observation and the differential roles of PFC and hippocampus observed in recent neuroscience experiments.

A2C EPN Architecture

We design a biologically inspired architecture that maps to functions and neural structures in the brain. Specifically, we build on the work of Wang et al. (2018) in which the prefrontal cortex (PFC) is conceptualized as forming a gated recurrent neural network (characterized as an LSTM) and augment it with an episodic memory which is connected to the LSTM via a single modified transformer block from Ritter et al. (2020) that represents the hippocampus.

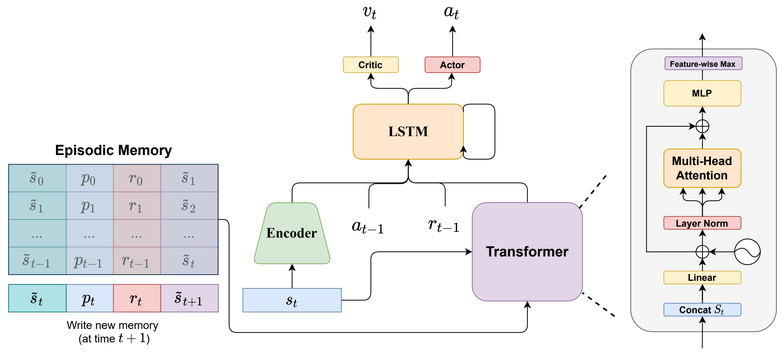

Figure: In the center of the architecture is an LSTM that takes as input the action and reward of the previous timestep, an encoded version of the current state, and a distilled representation of the relevant memory entries. The encoder is just a two-layered MLP. The episodic memory stores the features of the stone being transformed $\widetilde{s}_t$, a one-hot encoding of the applied potion color p_t, the reward r_t, and the resultant features of the stone after transformation $\widetilde{s}_{t+1}$. The transformer block architecture is fully described in Section **??**. The output of the LSTM is finally given to the policy and value networks, each of which is a linear layer, to generate an action and the value estimate respectively.

Performance

Table: Evaluation episode scores comparing the effect of the canonical representation (indicated by an `o') proposed by the Alchemy benchmark to the modified one (indicated by an `x'). The VMPO result taken from the original Alchemy paper. The no bottleneck column indicate the average score of the evaluation episodes with no missing edges.

Agent	Input	Output	Mem	Score ± SEM	Score (w/o Mem)	No Bottleneck
A2C EPN	x	x	x	272.85 ± 1.78	168.88	309.90
	o	x	x	243.83 ± 2.21	171.59	300.79
	o	o	x	156.34 ± 1.57	156.70	181.41
	o	o	o	158.91 ± 1.60	153.40	182.10
VMPO	o	o	-	155.40 ± 1.60	-	-
Random Heuristic	-	-	-	146.07 ± 1.55	-	172.11
Ideal Observer	-	-	-	284.42 ± 1.59	-	313.31

We observe that our A2C EPN agents adapts its strategy throughout the course of the episode. Concretely, in the first trial it performs a lot of exploratory actions by trying out potions that do not have an effect on the stone (the orange area) while in the last trial it shifts towards a more exploitative strategy by performing more actions that improved the value of the stone as it acquired knowledge about the episode's chemistry. This ability to adapt is indicative of good meta-learning performance and is similar in behavior to what we see from the Ideal Observer.

Behavioral Results

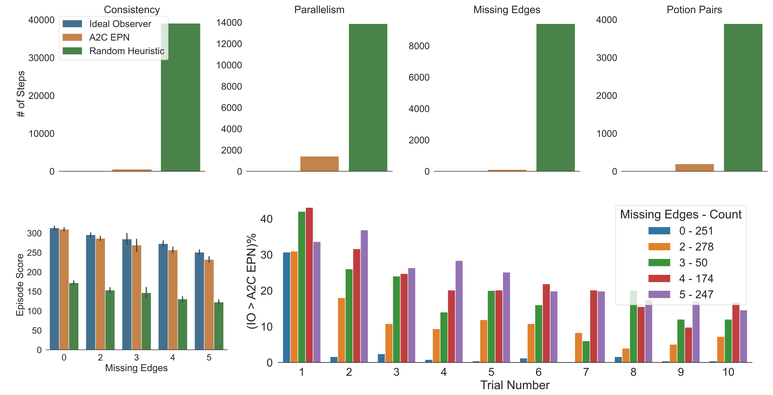

Figure: **Top**: Here, we report the number of times the agent took a null transition that it should have avoided if it understood the corresponding concept. The Ideal Observer is zero in all cases demonstrating perfect understanding of each principle. Our agent performs significantly better than the Random Heuristic but still takes a lot of actions where it could know better in the Parallelism and Potion Pairs analyses. **Bottom-left**: The average episode score as a function of the number of missing edges for each agent. **Bottom-right**: The percentage of episodes where the Idea Observer scores more than the A2C EPN agent at a given trial.

Single-Unit Activations

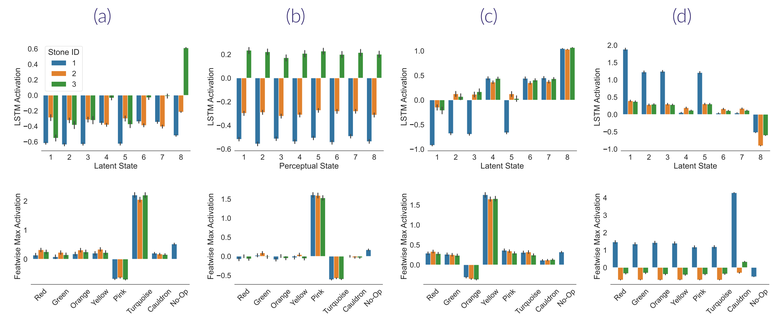

Figure: **Top**: LSTM units. **(a)** The activation of unit 192 as a function of the latent state. This unit is only responsive when stone 3 is in latent state 8 (the state with the highest reward). **(b)** The same unit as in (a) but as a function of the stone's perceptual state. It implies that the unit is responsive regardless of the stone's perceptual features. **(c)** The activations of unit 84 has a magnitude proportional to the reward of its corresponding state regardless of which stone is used. Note that states $\{4, 6, 7\}$ have a reward of $+1$ while the rewards of states $\{2, 3, 5\}$ is -1. **(d)** The activations of unit 26 are more positive in the states with negative reward. **Bottom**: Transformer units. **(a)** Unit 34 has a high activation when the agent choose the turquoise potion while a negative activation when it chooses its opposite color. **(b)** Unit 56 is the opposite of (a). **(c)** Unit 60 is similarly responsive when the agent chooses potion with a specific color (here yellow) and has a negative response for its opposite color. **(d)** Unit 9 is selective when the agent chooses an action that uses stone 1 regardless of the potion color.

Inspired by single-cell recordings in neuroscience, where single neurons are usually shown to be selective for a specific abstract concept, we probed our model to see if it will give rise to similar selectivity by analysing the activations of single units in the LSTM and that of the transformer.