Self-recognition as Optimisation

Timothy Atkinson and Nihat Engin Toklu

NNAISENSE, Lugano, Switzerland
{timothy, engin}@nnaisense.com

Abstract

We tackle the concept of 'self-recognition' in a simulated setting. We propose an experiment where two simultaneous reinforcement learning environments are controlled by two agents. Although each agent is given the control of its own environment, both agents receive the visual input of the *same* environment. The success threshold depends on self-recognition by definition as the agent must answer: am I seeing a mirror, or am I seeing a camera? We show that this experiment can be posed as an optimisation problem, solvable via evolutionary computation.

Introduction

Self-awareness is defined as "the capacity to become the object of one's own attention" (Morin, 2006). Self-awareness is studied also within the context of computational systems (Lewis et al., 2011). Here, we focus on *self-recognition*, which has been suggested as evidence of self-awareness (Gallup Jr, 1998), or seen as the implication of "some form of self-awareness" (Morin, 2011).

The classic 'mark-test' mirror test (Gallup Jr, 1970), used to detect self-recognition in animals equipped with visual stimuli, broadly proceeds as follows: (*a*) mark the animal with a dye (somewhere not directly visible to the animal) (*b*) observe the animal's frequency of touching the dyed area; (*c*) place a mirror in front of the animal (allowing the animal to see the dyed area) (*d*) observe again the animal's frequency of touching the dyed area. Observing a significant increase in the frequency of touching the dyed area is interpreted as the animal seeing itself in the mirror, observing the presence of the dye and therefore touching it. It is then argued that the animal must have an internalised notion of 'self' as to recognise the being in its visual stimuli as itself. A number of animals have been reported to pass various forms of the mirror test, including orangutans (Suárez and Gallup Jr, 1981), dolphins (Marten and Psarakos, 1994) and magpies (Prior et al., 2008).

Quoting Plotnik et al. (2006), the subject's behaviour is characterised as: (*i*) *social response*; (*ii*) *mirror inspection (looking behind the mirror)*; (*iii*) *repetitive mirror-testing behaviour (where the animal observes the effects of its own actions on the mirror image)*; and (*iv*) *self-directed behaviour (recognition of the mirror image as self)*. We argue that the most interesting parts of this process are (*iii*) and (*iv*). Therefore, we propose a new experimental setting based on reinforcement learning (RL) where expected reward can only be maximised through mirror-testing and ultimately self-recognition in a mirror. Our rewarding mechanism has a clear threshold which can only be surpassed through self-recognition. Experiments demonstrate that existing evolutionary algorithms can readily surpass this threshold. While similar efforts have been made for self-recognising artificial intelligence (Haikonen, 2007; Winfield, 2014; Pipitone and Chella, 2021), these were not framed as a pure optimisation problem.

The source code for our study is available online [1].

Self-recognition in Reinforcement Learning

We propose an experimental setting with two simultaneous RL environments: the 'mirror environment' E_M and the 'camera environment' E_C. Using the same policy (neural network *and* weights), two agents A_M and A_C operate in E_M and E_C, respectively. The distinctive feature is that both A_M and A_C observe E_M (i.e. receive their inputs from E_M). We argue that, by observing E_M, A_M can be thought as 'seeing' itself in the mirror (since its inputs are its own environment and 'body'). Meanwhile, A_C will be seeing the other agent's environment and body. A_M gets rewarded for completing the assigned task in E_M while A_C gets penalised for producing non-zero actions. To reach success, A_M must realise that it is observing itself ('passing the mirror test') and act to complete the task, while A_C must realise that it is observing another agent and stop. Reaching this realisation is non-trivial: initially, an agent has no way of knowing whether or not its observations are coming from the mirror. Therefore, the agent must 'understand' in time if its own actions cause updates on observations. We pose this framework as an optimisation problem of seeking policy parameters which maximise the total reward obtained by A_M and A_C.

[1] https://github.com/nnaisense/self-recog-as-optim.git

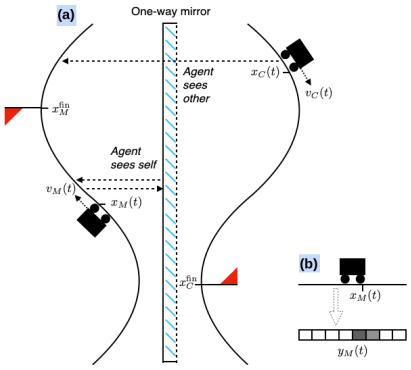

Figure 1: **(a)** The set up of the two environments E_M and E_C, with the agent seeing itself in E_M and the agent seeing the other environment in E_C. **(b)** Conversion of position $x_M(t)$ to visual input $y_M(t)$ passed to both agents.

Our experiments modify the classic mountain car task (Singh and Sutton, 1996). Contained in each environment E_i, $i \in \{M, C\}$, is a car at time t with position $x_i(t)$ and velocity $v_i(t)$. The state of agent A_i in E_i is denoted $s_i(t)$. The system is visualised in Figure 1. In each iteration $t+1$, the agent outputs actions $a_i(t+1) \in [-1, 1]$ which is integrated into the environment state according to,

$$v_i(t+1) = v_i(t) + \alpha a_i(t+1) - \beta \cos\left(3 * x_i(t+1)\right),$$
$$x_i(t+1) = \texttt{clamp}\left(x_i(t) + v_i(t+1), -1.6, 0.6\right),$$

with $\alpha = 0.001$ and $\beta = 0.0025$. Initial conditions are set as $x_i(0) = -\frac{1}{2}$ and $v_i(0) = 0$ such that their initial dynamics are identical. The critical difference between them lies in the observations. *Both* agents receive a 32-variable visual representation $y_M(t)$ of $x_M(t)$ (see Figure 1b), such that $a_i(t+1) = \texttt{Policy}(y_M(t), s_i)$. The reward in each time step is, $r_i(t) = -\texttt{min}\left(\texttt{abs}(x_i(t) + x_i^{\text{fin}})/D^{\text{fin}}, 1\right)$ where $x_M^{\text{fin}} = \pi/6$ and $x_C^{\text{fin}} = -\frac{1}{2}$ and $D^{\text{fin}} = \frac{x_M^{\text{fin}} - x_C^{\text{fin}}}{2}$. A_M is rewarded for maximising being at the top of the hill in E_M, and A_C for staying at the origin in E_C. The episode terminates at $t = 200$.

There exists a threshold expected episodic reward above which we can definitively claim the presence of self-recognition. Specifically, for an agent which *cannot* discriminate between E_M and E_C such that $x_M(t)$ and $x_C(t)$ follow the same distribution, $r_M(t) + r_C(t) \leq -1$, the total episodic reward across both settings is bounded: $R = \sum_{t \in \{0,1,2,\ldots,200\}} r_M(t) + r_C(t) \leq -200$. Therefore, when we see $R > -200$, the agent has learned some degree of self-recognition.

Experiments and Discussion

To solve the proposed environment, we use Separable Natural Evolution Strategies (SNES) (Schaul et al., 2011) with the ClipUp optimiser (Toklu et al., 2020) with initial radius $r = 4.5$. As in Salimans et al. (2017), we do not adapt

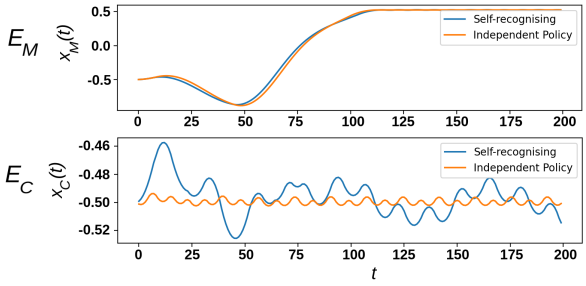

Figure 2: Position trace of learned agents, where 'self recognising' refers to a single policy trained to recognise itself across both environments, and 'independent policy' refers to individually trained policies for both environments. On E_M, both agents climb the hill at a similar rate. On E_C, the self-recognising agent takes much larger movements, suggesting that the agent prioritises E_C for mirror testing. It is worth noting that the self-recognising agent terminates within 0.1 of the goal position in 87% of scenarios, suggesting that performance can be further improved.

σ. The policy is a 64-neuron recurrent network with `ELU` activation for the hidden layer (Clevert et al., 2015) and `tanh` activation for the output layer. Each hidden activation is passed through layer normalisation (Ba et al., 2016). The agent's initial hidden state $s_i(0)$ is randomised, drawn from $\mathcal{N}(0, I_{64})$ and then passed through `ELU` activation and then layer normalisation, allowing the agent to take psuedo-randomised actions for mirror testing.

The population size is $15,000$ and evolution is run for 4000 generations. Each candidate solution is evaluated for total episodic reward R when controlling each environment E_M or E_C in turn. The experiment is repeated 10 times. For comparison, we run the experiment in a setting with an independent policy for each E_i. In the self-recognition scenario, the median generations for the mean fitness of the population to surpass the threshold of -200 is **967**, demonstrating task solvability with existing techniques. This is substantially higher than the **141** generations needed with independent policies for each E_i meaning that the introduction of 'self-recognition' substantially complicates the overall task. This difference is statistically significant from the non-parametetric two-tailed Mann-Whitney U-test (Mann and Whitney, 1947) ($p = 10^{-4} \leq 0.05$) and the effect size is large according to the non-parametric Vargha-Delaney A Test (Vargha and Delaney, 2000) ($A = 0.96 \geq 0.71$).

While we are not claiming that the learned simple $64-$neuron RNNs are truly 'self-aware', it is interesting to note that the self-recognition test can readily be solved by existing evolutionary algorithms. From a more practical perspective, the substantial performance degradation obtained through the adaption of an existing environment to a self-recognition setting suggests a general template for describing harder, intrinsically hierarchical, RL benchmarks.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint*. arXiv:1607.06450.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint*. arXiv:1511.07289.

Gallup Jr, G. G. (1970). Chimpanzees: self-recognition. *Science*, 167(3914):86–87. doi: 10.1126/science.167.3914.86.

Gallup Jr, G. G. (1998). Can animals emphatize? Yes. *Scientific American*. Feature article: Animal self-awareness: A debate.

Haikonen, P. O. (2007). Reflections of consciousness: The mirror test. In *AAAI Fall Symposium: AI and Consciousness*, pages 67–71.

Lewis, P. R., Chandra, A., Parsons, S., Robinson, E., Glette, K., Bahsoon, R., Torresen, J., and Yao, X. (2011). A survey of self-awareness and its application in computing systems. In *2011 Fifth IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshops*, pages 102–107. doi: 10.1109/SASOW.2011.25.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60.

Marten, K. and Psarakos, S. (1994). Evidence of self-awareness in the bottlenose dolphin (tursiops truncatus). In *Self-awareness in animals and humans: Developmental perspectives*, pages 361–379. Cambridge University Press. doi: 10.1017/CBO9780511565526.026.

Morin, A. (2006). Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and cognition*, 15(2):358–371. doi: 10.1016/j.concog.2005.09.006.

Morin, A. (2011). Self-recognition, theory-of-mind, and self-awareness: What side are you on? *Laterality*, 16(3):367–383. doi: 10.1080/13576501003702648.

Pipitone, A. and Chella, A. (2021). Robot passes the mirror test by inner speech. *Robotics and Autonomous Systems*, 144:103838. doi: 10.1016/j.robot.2021.103838.

Plotnik, J. M., De Waal, F. B., and Reiss, D. (2006). Self-recognition in an asian elephant. *Proceedings of the National Academy of Sciences*, 103(45):17053–17057. doi: 10.1073/pnas.0608062103.

Prior, H., Schwarz, A., and Güntürkün, O. (2008). Mirror-induced behavior in the magpie (pica pica): evidence of self-recognition. *PLoS biology*, 6(8):e202. doi: 10.1371/journal.pbio.0060202.

Salimans, T., Ho, J., Chen, X., Sidor, S., and Sutskever, I. (2017). Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint*. arXiv:1703.03864.

Schaul, T., Glasmachers, T., and Schmidhuber, J. (2011). High dimensions and heavy tails for natural evolution strategies. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 845–852. doi: 10.1145/2001576.2001692.

Singh, S. P. and Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1):123–158. doi: 10.1023/A:1018012322525.

Suárez, S. D. and Gallup Jr, G. G. (1981). Self-recognition in chimpanzees and orangutans, but not gorillas. *Journal of human evolution*, 10(2):175–188. doi: 10.1016/S0047-2484(81)80016-4.

Toklu, N. E., Liskowski, P., and Srivastava, R. K. (2020). Clipup: A simple and powerful optimizer for distribution-based policy evolution. In *International Conference on Parallel Problem Solving from Nature*, pages 515–527. Springer. doi: 10.1007/978-3-030-58115-$2_3$6.

Vargha, A. and Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132. doi: 10.3102/10769986025002101.

Winfield, A. F. T. (2014). Robots with internal models: A route to self-aware and hence safer robots. In *The Computer After Me: Awareness and Self-Awareness in Autonomic Systems*, page 237–252. Imperial College Press. doi: 10.1142/9781783264186$_0$016.