Informational Trade-offs of Learning from Expert Demonstration

Rebekah Gelpi*
(rebekah.gelpi@mail.utoronto.ca)
Department of Psychology, University of Toronto
Toronto, ON, Canada

Yikai Tang*
(yikai.tang@mail.utoronto.ca)
Department of Psychology, University of Toronto
Toronto, ON, Canada

William A. Cunningham (wil.cunningham@utoronto.ca)
Department of Psychology, University of Toronto
Toronto, ON, Canada

Keywords: deep learning; reinforcement learning; cultural transmission; pedagogy; social learning

Introduction

Human beings develop in a highly complex social and physical environment. Behaving appropriately in this environment requires learning detailed action sequences, where intermediate actions do not provide obvious instrumental rewards. Alongside a high degree of general-purpose intelligence, humans have adapted to this computational challenge through a deep reliance on learning through the cultural transmission of information from teachers or other social sources (Boyd et al., 2011; Mesoudi et al., 2006). This deep cognitive adaptation is expensive, requiring a large investment of each generation of humans in providing for and teaching the subsequent generation, and an extended period of childhood longer than that observed in other animals (Gopnik, 2020). During this time, children are both dependent on caregivers for resources, and spending a large amount of energy on brain development.

Nevertheless, learning from expert demonstrators obviates the need to engage in time-consuming and even possibly dangerous exploration to discover solutions already known by other members of society, and allows for cultures to develop new tools and technologies by allowing its members to build upon previous knowledge cumulatively (Tennie et al., 2009).

Teaching provides many opportunities for learning above and beyond serving as another source of information for a learner. Because teachers are intentional agents, it is possible to make strong assumptions behind the rationale for their behavior, leading to stronger inferences about the data than if it had been independently discovered (Shafto et al., 2014). However, here we focus on a simpler phenomenon: teachers tend to be more skilled, and observing an expert demonstrator can improve learning by providing learners with access to examples of success before they are able to succeed themselves. Indeed, prior work has found that using expert demonstrations to pretrain or guide exploration can substantially improve learning speed and performance in RL agents (e.g. Gulcehre et al., 2019; Zhang & Ma, 2018).

To investigate the benefits of expert demonstration, we develop and test a simple grid world game in which an agent either learns through self-directed exploration, observation of a pre-trained expert demonstrator, or a combination of both of varying proportions.

Method

We implement a 10 × 10 grid world in which one agent, two bushes, and one wolf are located at coordinates in space. All the objects are randomly distributed throughout the world. The agent and the bushes have a certain energy level when they are instantiated. The agent's action space involves basic movements (up, down, left, and right) and eating, each consuming energy to perform. When the agent eats while adjacent to a bush, its energy level increases and the bush's energy level decreases. When an agent's energy level decreases to zero, the agent will 'die'; bushes with an energy level of zero no longer provide energy. Unlike the agent and the bushes, the wolf has unlimited energy. It intermittently hunts the agent with a predetermined action policy. The agent is rewarded when it eats bushes and when it survives for 50 turns, but it is punished when eaten by the wolf or when it starves.

Model Architecture

The agent contains a deep Q-learning neural network (DQN) that takes in the location and identity of nearby objects as well as its own hunger level as its observation of the world. Observations are first input into an LSTM followed by a linear policy that outputs the estimated Q-value of the five possible state-action pairs (four cardinal directions plus eating). The agent also contains a replay buffer that stores past experiences, either from self-directed exploration or from a pre-trained expert demonstrator. After each epoch, the neural network samples a batch of multi-state game sequences, and updates its policy estimates based on the rewards obtained in these states.

Experimental Conditions

We trained the agent for 200,000 games in one of five conditions. Each game is initialized with varying agent energy levels (between 15 and 100) and ends after 50 steps or when the agent dies. Individual games sometimes include a wolf, and sometimes do not. As a result, agents learn

about games that have differing optimal policies for survival (e.g. seek out food first, or avoid the wolf first).

We generated data for 5 agents, corresponding to differing levels of experience received from a pre-trained expert demonstrator. In Condition 1, the agent learns solely through its own experiences of interacting with the environment, and does not receive any expert demonstration. In Conditions 2–5, a gradually increasing proportion of the agent's learning trials correspond to a game played by an expert demonstrator (12.5%, 25%, 50%, and 100%, respectively). Every 1000 epochs, the agent is presented with 900 test games with an initial energy level of 15 in the grid world. We test agents' performance by recording the number of steps survived on the test trials.

Results and Discussion

To assess the final performance of the model, we conducted a series of *t*-tests with Bonferroni correction for multiple comparisons to evaluate the performance of each fully trained model on 10000 new test games. We found that a proportion of 25% expert trials had a better performance than all other models (all $p < .001$), but also that models mixing both learning strategies outperformed the two that used only one or the other (all $p < .001$). Notably, the size of the performance increase from 25% expert trials compared to 100% expert trials (Cohen's $d = 1.08$) and self-directed learning (Cohen's $d = 1.40$) were both very large.

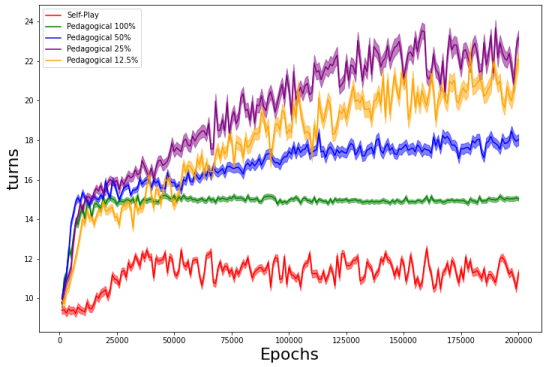

Figure 1. Average turns survived by agents for self-directed exploration (red), as well as 12.5% (yellow), 25% (purple), 50% (blue), and 100% expert demonstration (green) conditions. Results are averaged over 5 model runs. Shaded region indicates standard error value.

Overall, all pedagogical models substantially outperformed learning from self-directed exploration alone. Exposure to expert demonstrations led all agents to quickly improve well beyond the maximum average survival of the self-directed learning model. Nevertheless, not all forms of demonstrations were equally valuable. For example, being presented with only expert trials led agents to quickly stop improving their performance, with a ceiling achieved after 15 turns. This outcome reflected highly robust learning of how to avoid being eaten by a wolf, but an inability to reliably generalize a policy that included eating from the bushes to avoid starvation. In contrast, while other agents displayed a higher proportion of being eaten by a wolf, this was traded off against an ability to use self-directed learning to learn how to eat and thus survive longer on average.

Conclusions and Ongoing Research

These simulations suggest that learning from an expert can provide an immediate advantage over learning from one's own error-prone first attempts, and that even small amounts of expert guidance can provide a lasting boost to one's total learning (e.g. Gulcehre et al., 2019). Nevertheless, it also shows that relying too heavily on an expert can limit one's learning—serving as a "double-edged sword" (e.g., Bonawitz et al., 2011) that limits one's capacity for future exploration. Instead, success requires balancing expert knowledge with exploration, echoing the iterative innovation process that is characteristic of human cumulative culture (Tennie et al., 2009).

We are currently investigating how dynamically shifting reliance on an expert can optimize its benefits. For example, when one has little idea of the best action policy, heavily drawing from an expert is highly beneficial; as one gains more personal experience, however, relying on one's own innovations becomes progressively more advantageous.

Acknowledgments

We acknowledge the support of the Social Sciences and Humanities Research Council of Canada [SSHRC-506547] and the Natural Sciences and Engineering Research Council of Canada [RGPIN-2018-05946].

References

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*, 10918–10925.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1803), 20190502.

Mesoudi, A., Whiten, A., & Dunbar, R. (2006). A bias for social information in human cultural transmission. *British Journal of Psychology*, *97*(3), 405–423.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Tennie, C., Call, J., & Tomasello, M. (2009). Ratcheting up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2405–2415.

Gulcehre, C.*, Paine, T. L.*, et al. (2019). Making Efficient Use of Demonstrations to Solve Hard Exploration Problems. arXiv:1909.01387

Zhang, X., & Ma, H. (2018). Pretraining Deep Actor-Critic Reinforcement Learning Algorithms With Expert Demonstrations. arXiv:1801.10459