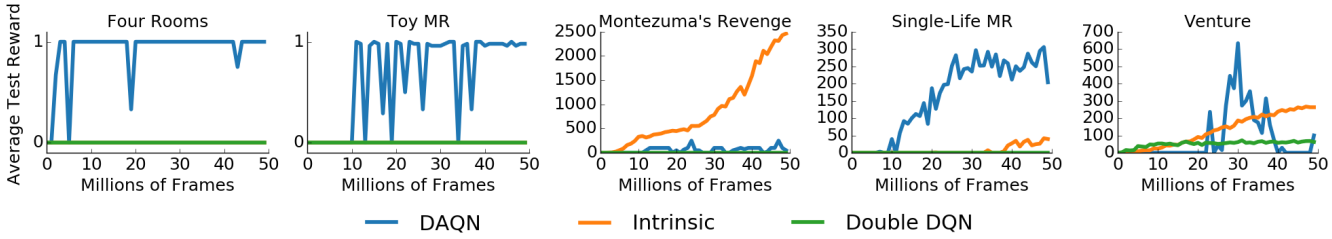

Figure 5: Average reward in the Four Rooms, Toy MR, Atari MR, Single-Life Atari MR, and Atari Venture domains using the following models: DAQN (blue), Double DQN (green) and IM (orange). In Four Rooms and Toy MR, both IM and Double DQN fail to score an average reward above zero, and are thus overlapping. We use the raw IM and Double DQN data from Bellemare et al. [3] on Montezuma's Revenge and Venture. All other plots show our implementations' results.

exploration, which is largely absent from existing state-of-the-art methods.

The main drawback to our approach is the requirement for a hand-annotated state-projection function that nicely divides the state-space. However, for our method allows this function need only specify abstract states, rather than abstract transitions or policies, and thus requiring minimal engineering on the part of the experimenter. In future work, we hope to learn this state-projection function as well. We are exploring methods to learn from human demonstration, as well as methods that learn only from a high-level reward function. Ultimately, we seek to create compositional agents that can learn layers of knowledge from experience to create new, more complex skills. We also plan to incorporate a motivated exploration algorithm, such as IM [3], with our L_0 learner to address our difficulty with time-based traps in MR.

Our approach also has the ability to expand the hierarchy to multiple levels of abstraction, allowing for additional agents to learn even more abstract high-level plans. In the problems we investigated in this work, a single level of abstraction was sufficient, allowing our agent to reason at the level of rooms and sectors. However, in longer horizon domains, such as inter-building navigation and many real-world robotics tasks, additional levels of abstraction would greatly decrease the horizon of the L_1 learner and thus facilitate more efficient learning.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant numbers IIS-1426452, IIS-1652561, and IIS-1637614, DARPA under grant numbers W911NF-10-2-0016 and D15AP00102, and National Aeronautics and Space Administration under grant number NNX16AR61G.

REFERENCES

[1] Pierre-Luc Bacon, Jean Harb, and Doina Precup. 2017. The Option-Critic Architecture.. In *AAAI*. 1726–1734.
[2] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. 2013. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research* 47 (jun 2013), 253–279.
[3] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. 2016. Unifying Count-Based Exploration and Intrinsic Motivation. In *NIPS*.
[4] Ronen I Brafman and Moshe Tennenholtz. 2002. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3, Oct (2002), 213–231.
[5] Thomas G Dietterich. 2000. Hierarchical reinforcement learning with the MAXQ value function decomposition. *J. Artif. Intell. Res.(JAIR)* 13 (2000), 227–303.
[6] Carlos Diuk, Andre Cohen, and Michael L Littman. 2008. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 240–247.
[7] Nakul Gopalan, Marie desJardins, Michael L. Littman, James MacGlashan, Shawn Squire, Stefanie Tellex, John Winder, and Lawson L.S. Wong. 2017. Planning with Abstract Markov Decision Processes. In *International Conference on Automated Planning and Scheduling*.
[8] Ken Kansky, Tom Silver, David A Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, and Dileep George. 2017. Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics. *arXiv preprint arXiv:1706.04317* (2017).
[9] Tejas D. Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Joshua B. Tenenbaum. 2016. Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation. In *NIPS*.
[10] Amy McGovern, Richard S Sutton, and Andrew H Fagg. 1997. Roles of macro-actions in accelerating reinforcement learning. In *Grace Hopper celebration of women in computing*, Vol. 1317.
[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.
[12] Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Rémi Munos. 2017. Count-based exploration with neural density models. *arXiv preprint arXiv:1703.01310* (2017).
[13] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112, 1-2 (1999), 181–211.
[14] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-Learning.. In *AAAI*. 2094–2100.
[15] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. 2017. FeUdal Networks for Hierarchical Reinforcement Learning. In *ICML*.