TRL: Discriminative Hints for Scalable Reverse Curriculum Learning

Chen Wang[1], Xiangyu Chen[1], Zelin Ye[1], Jialu Wang[1], Ziruo Cai[1], Shixiang Gu[2,3], Cewu Lu[1]

[1]Shanghai Jiao Tong University [2]University of Cambridge [3]MPI Tübingen

Motivation

- Model-free deep reinforcement learning (RL) methods have been successful in a wide variety of domains (Atari, AlphaGo Zero, etc.).
- Tasks with sparse rewards and large state space remain challenging for traditional deep RL. Goal-oriented tasks are among the most typical problems, where a reward can only be received when the final goal is achieved.
- We derive **TRL**, an experience-based tendency reward mechanism, which generates additional hints of tendency to help existing deep RL methods overcome Goal-oriented problems with large state space efficiently.

(a) (b) (c) (d) (e)

We evaluate our model in five experiments with different levels of difficulties: (a) Maze, (b) Super Mario Bros, (c) Grasping task in MuJoCo, (d) Conveyance challenge in MuJoCo, (e) Recurrent Pick-and-place in MuJoCo.

Backgrounds

RL learns the **optimal policy** $\pi_\theta(a_t|s_t)$ for an agent according to a **reward function** $r(s_t, a_t)$ by maximizing γ-discounted cumulative returns $J(\theta) = \mathbb{E}_\pi[\sum_t \gamma^t r(s_t, a_t)]$. There have been several lines of approaches to tackling sparse reward tasks.

Curriculum learning, e.g.Reverse curriculum gengration [3]

- show promising results in solving a variety of sparse reward problems but is lack of efficiency in larger state space tasks with longer planning horizon

Hierarchical RL and intrinsic motivations, e.g.HDRL [2], Vime [4]

- drastically reduce the search dimension of the problem
- improve the exploration even under no rewards
- both approaches assume forward-based exploration, which becomes increasingly difficult as the dimensionality and the time horizon increase, unless strong structured prior is provided

RL with demonstrations, e.g.Shaping from Demonstration [1]

- prove successful in many manipulation tasks but make strong assumptions about the quality of demonstrations received

Tendency Reinforcement Learning

TRL is derived using **discriminative learning** on past experiences during an automated reverse curriculum, which not only provides dense learning signals on which states are likely to lead to success, but also allows the agent to retain only this tendency classifier instead of the whole histories of experience.

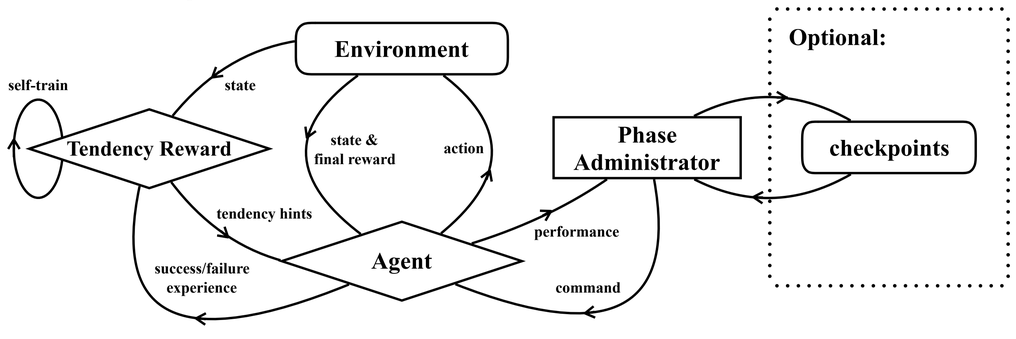


Fig 1: The schematic illustration of TRL.

TRL performs efficiently on solving Goal-oriented tasks and is also compatible to an **optional checkpoint scheme** with very small quantity of key states to tackle difficult robot manipulation challenges directly from perception. TRL also shows robustness to misleading checkpoints.

Experiments without checkpoint scheme

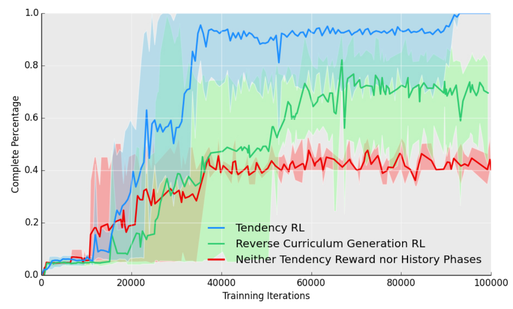


Fig 2: The training results of ablation experiments in Maze under three different conditions: using tendency reward without history phases (TRL), using history phases without tendency reward (Reverse curriculum generation) and using neither tendency reward nor history phases. TRL performs more efficiently.

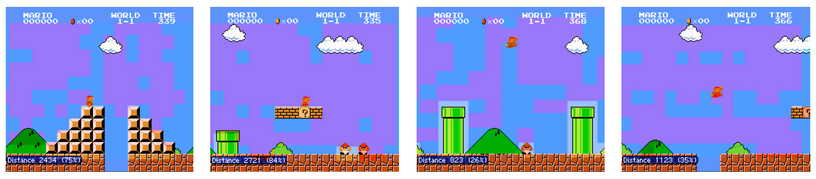
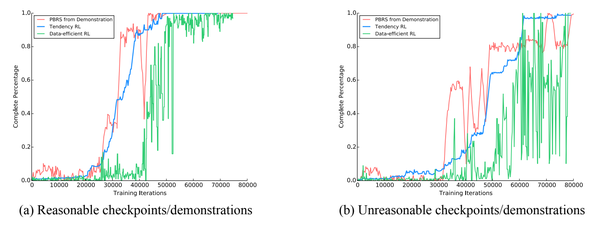


Fig 3: The distribution of tendency hints in Mario game. The purple color represents positive hints while the blue ones indicate negative hints. These figures show the guiding effect of the tendency reward, which helps the agent when it is far away from the final goal.

Fig 4: The generated phases in training process. The points indicate the starting points of Mario. Points with same color belong to a same phase.

Experiments with checkpoint scheme



(a) Reasonable checkpoints/demonstrations (b) Unreasonable checkpoints/demonstrations

Fig 5: Learning curves of TRL and other demonstrations RL methods with checkpoints / demonstrations of different qualities. PBRS from demonstration may outperform TRL with well hand-engineered reward shaping, while TRL is able to achieve a satisfactory performance only using crude checkpoints.

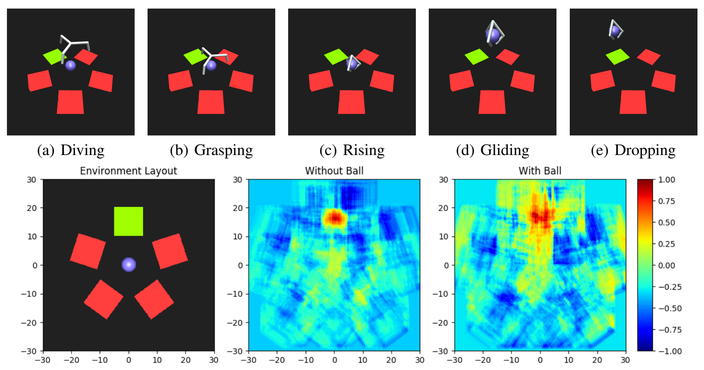


Fig 6: Five key states for the conveyance challenge and the final learnt tendency heat map. Since catching and dropping require more accuracy, the positive area of "without ball" is smaller.

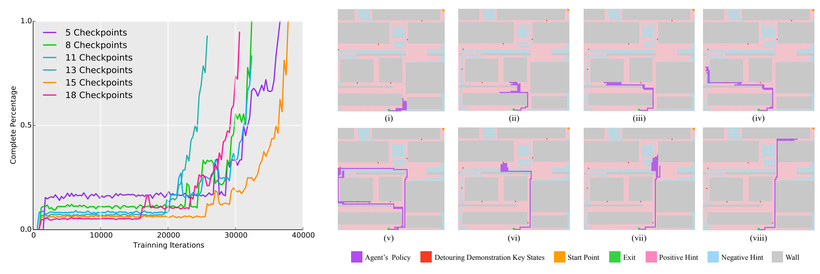


Fig 7: Learning curves of the grasping task with different numbers of checkpoints and the robustness test with several misleading checkpoints in large Maze task. The result shows that TRL is stable and robust to different quantity and quality of checkpoints.

References

[1] Brys, T., Harutyunyan, A., Suay, H. B., Chernova, S., Taylor, M. E., and Nowé, A. (2015). Reinforcement learning from demonstration through shaping. In *IJCAI*, pages 3352–3358.

[2] D.Kulkarni, T. and Karthik Narasimhan, Ardavan Saeedi, J. T. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *nips*.

[3] Florensa, C., Held, D., Wulfmeier, M., Zhang, M., and Abbeel, P. (2017). Reverse curriculum generation for reinforcement learning. In Levine, S., Vanhoucke, V., and Goldberg, K., editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 482–495. PMLR.

[4] Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. (2016). Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117.