SCIENCE CHINA
Information Sciences

• MOOP •

January 2022, Vol. 65 114201:1–114201:3
https://doi.org/10.1007/s11432-019-2648-7

Manipulation skill learning on multi-step complex task based on explicit and implicit curriculum learning

Naijun LIU[1,2], Tao LU[1*], Yinghao CAI[1], Rui WANG[1] & Shuo WANG[1,2,3]

[1]*Institute of Automation, Chinese Academy of Sciences, Beijing* 100190, *China;*
[2]*University of Chinese Academy of Sciences, Beijing* 100049, *China;*
[3]*CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai* 200031, *China*

Received 14 July 2019/Accepted 10 September 2019/Published online 24 May 2021

Citation Liu N J, Lu T, Cai Y H, et al. Manipulation skill learning on multi-step complex task based on explicit and implicit curriculum learning. Sci China Inf Sci, 2022, 65(1): 114201, https://doi.org/10.1007/s11432-019-2648-7

Recently, remarkable progress of deep reinforcement learning (DRL) with regard to robot skill learning has been witnessed [1]. However, manipulation skill learning on multi-step complicated tasks with DRL still poses great challenges owing to its higher exploration of state-action space and uninformative sparse rewards. To alleviate these problems, some studies explored reward shaping [2] to guide the learning process toward good solutions, which inevitably requires significant expert knowledge and manual efforts. Other study overcomes exploration with demonstrations [3] which generally is costly and may result in suboptimal performance. Curriculum learning [4], which starts learning small domains with easier aspects of tasks and then gradually increases the difficulty level, is also applied to learn relatively easy manipulation skills [5]. Another method called hindsight experience replay (HER) [6] in hindsight imaged that every state that the agent finally reached was actually a goal, which showed appealing performance in some manipulation tasks with only sparse rewards available. However, it is still quite challenging for HER to directly train policies for multi-step manipulation tasks (e.g., picking and placing several objects).

To address the challenges in manipulation skill learning on multi-step complex tasks, we propose an efficient curriculum learning method called task auxiliary and task difficulty-HER (TATD-HER), which is endowed with a high-low level curriculum learning structure and combines the explicit and implicit curriculum learning mechanisms together for policy training. The experimental results demonstrate that our proposed learning paradigm is capable of achieving satisfactory performance in manipulation skill learning on multi-step complicated tasks.

Methodology. The idea behind TATD-HER is to generate auxiliary mechanisms to effectively guide policy search. The main novelty of TATD-HER is that it uses both explicit and implicit curriculum learning together for complex manipulation skill training. As TATD and HER jointly form an efficient high- and low-level curriculum learning framework, the TATD-HER method provides smoother gradients for policy learning on complicated tasks.

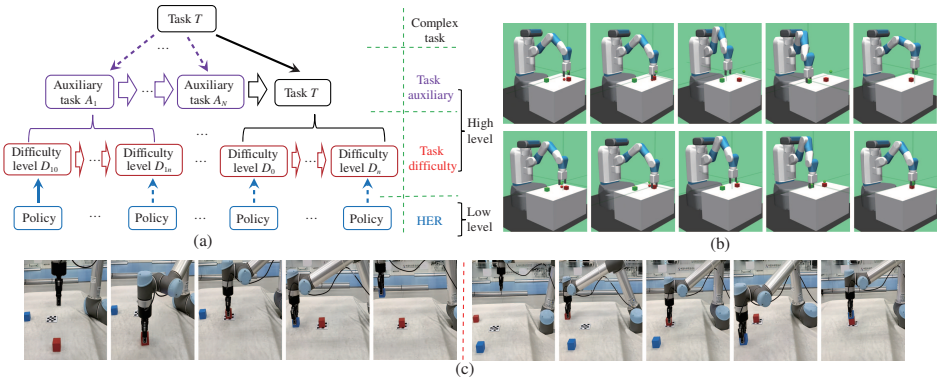
We first consider a standard reinforcement learning (RL) algorithm under the HER learning framework. At time step t, the robot agent takes an action a_t sampled from policy $\pi_\theta(s_t||g)$ and receives reward $r(s_t||g, a_t)$, where $||$ denotes concatenation, and then moves from state s_t to the next state s_{t+1} based on the transition dynamics $p(s_{t+1}|s_t, a_t)$. At the end of each episode, a trajectory sequence τ : $\langle s_0||g, a_0, s_1||g, a_1, \cdots, s_H||g\rangle$ is obtained. When interacting with the environment, in most cases, the robot agent fails to reach the goal. For the HER agent, a failed trajectory τ can be transformed to a successful one τ^{her} : $\langle s_0||g_0^h, a_0, s_1||g_1^h, a_1, \ldots, s_H||g_H^h\rangle$ with the hindsight technique, where g_i^h is the reached position after the ith state, $i = 0, 1, \ldots, H$.

The illustration of the proposed TATD-HER method is shown in Figure 1(a). The method is endowed with a high-low level curriculum learning structure. Task auxiliary and task difficulty (TATD) is at the high level, explicitly forming curriculum learning via generating auxiliary tasks in a meaningful order $\{A_1, A_2, \ldots, A_N\}$ which gradually illustrates more concepts and more complex ones according to the type of complex task T. For the auxiliary task A_k, we set the difficulty level D_{ki} such that it gradually increases from low level D_{k0} to the normal level D_{kn} in the training procedure,

$$D_{ki} = \begin{cases} \frac{D_{kn} - D_{k0}}{I_k} i + D_{k0}, \ i \leqslant I_k, \\ D_{kn}, \ i > I_k, \end{cases} \tag{1}$$

where i is the iteration step for training the auxiliary task A_k, and I_k is its maximum iteration steps. For the complex task T, the difficulty level D_i is similar to (1). HER is at the low level, which implicitly forms the curriculum learning technique by using the hindsight technique to train the same

* Corresponding author (email: tao.lu@ia.ac.cn)



Method	Reward shaping [2]	Overcoming [3]	HER [6]	DDPG [7]	PPO [8]	TATD-HER	TATD-HER (no TATD)	TATD-HER (no HER) [4]
Success rate (%)	2	92	5	3	1	96	5	7

(d)

Figure 1 (Color online) (a) Illustration of the high-low level TATD-HER method with explicit and implicit curriculum learning mechanisms. (b) Frames are captured from the final trained policy employed in the simulated environment. The robot arm picks and places two cubes to their target positions denoted by different color ball points. The first and second rows show two different manipulation circumstances. (c) Frames are captured from the final trained policy employed on the real-world UR5 robot corresponding to the simulated environment. (d) Performance comparison with other methods.

policy on different difficulty levels of the auxiliary tasks A_k and the complex task T with the deep deterministic policy gradient (DDPG) [7].

The neural network parameter ϕ of Q-function Q_ϕ in DDPG is updated with

$$\phi = \phi - \alpha \nabla_\phi \sum_B (Q_\phi(s^k||g^{kh}, a^k) - y(r^k, s'^k||g'^{kh}))^2, \quad (2)$$

where B is the stored transformed transition $(s^k,\ g^{kh},\ a^k, r^k, s'^k, g'^{kh})$ from the auxiliary task A_k. s^k, g^{kh}, a^k, r^k, s'^k, and g'^{kh} denote the state, reached position, action, received reward, next state, and next reached position, respectively.

$$y(r^k, s'^k||g'^{kh}) = r^k + \gamma Q_{\phi\text{targ}}(s'^k||g', \pi_{\theta\text{targ}}(s'^k||g')), \quad (3)$$

where $Q_{\phi\text{targ}}$ and $\pi_{\theta\text{targ}}$ are the target Q-function and target policy, respectively. The neural network parameter θ of policy π is updated with

$$\theta = \theta + \beta \nabla_\theta \sum_B Q_\phi(s^k||g^{kh}, \pi_\theta(s'^k||g'^{kh})). \quad (4)$$

The detailed algorithm is shown in Algorithm 1.

Algorithm 1 TATD-HER curriculum learning method

Input: Complicated task T.
Output: Policy π_θ.

```
Generate auxiliary tasks in a meaningful order according to T: {A_1, A_2, ..., A_N} ← T;
Initialize policy π_θ with random weights;
for k = 1 : N do
    for i = 0 : k_steps do
        Set difficulty level D_ki for A_k with (1);
        Update policy parameter θ with (4);
    end for
    Policy π_θ converges on auxiliary task A_k;
end for
for i = 0 : T_steps do
    Set difficulty level D_i for T with (1);
    Update policy parameter θ with (4);
end for
Policy π_θ converges on task T.
```

Experiments and results. We instantiate and evaluate our method on a simulated robot with parallel grippers to pick two cubes and to place them at two different target positions (one target position on the desk, and the other in the air), in MuJoCo environment interfaced with Open-AI Gym, as is shown in Figure 1(b). To achieve the designed multi-step manipulation task T, the robot agent has to learn a sequence of skills: picking the first cube from a random position; placing the first cube on its target position on the desk; picking the second cube from the random position; moving the gripper to the target position of the second cube.

We set trajectory length $H = 150$, and learning rates $\alpha = 0.0001$ and $\beta = 0.0003$. The settings of policy network, Q-function network, and other parameters are referred to in HER [6]. The policy takes a concatenated vector including the gripper position, two cube's positions, and two target positions as input, and outputs 4-dimensional action vector to move the robot gripper in 3D space and close or open the gripper fingers.

To train policy on the designated manipulation task T, two auxiliary tasks, A_1 and A_2, are generated. For the auxiliary task A_1, the robot agent is trained to pick the first cube (red cube) and place it on its target position, ignoring the second cube (green cube). The sparse reward function for auxiliary task A_1 is set to

$$r_t = \begin{cases} 0, & \text{if } ||x_{1t} - g_1|| < \delta, \\ -1, & \text{otherwise}, \end{cases} \quad (5)$$

where x_{1t} is first cube position at time step t, g_1 is the target position of the first cube, and δ is the threshold. For auxiliary task A_2, the robot agent is trained to place the first cube and the second cube on their target positions, which are both set on the desk. The sparse reward function for A_2 is

$$r_t = \begin{cases} 0, & \text{if } ||x_{1t} - g_1|| < \delta \text{ and } ||x_{2t} - g_2|| < \delta, \\ -1, & \text{otherwise}. \end{cases} \quad (6)$$

The sparse reward for task T is identical to (6). For skills trained on each auxiliary task A_k, the difficulty levels D_{ki} are set to δ. D_{k0} and D_{kn} are set to 5 cm and 0.5 cm, respectively.

We compare our proposed TATD-HER method against the following methods: Reward shaping [2], Overcoming [3] (requiring demonstrations), HER [6], DDPG [7], PPO [8] (the state-of-the-art policy gradient DRL method), TATD-HER without TATD mechanism, and TATD-HER without HER mechanism. All policies share the same neural network. Figure 1(d) summarizes the success rates of the final trained policies. TATD-HER shows the best performance among the above-mentioned methods and needs no demonstration data. The reason is that TATD-HER combines both explicit and implicit curriculum learning for policy training, which efficiently guides the policy optimization toward good solutions. The ablation experiments suggest that TATD and HER are both crucial components of our method, as only policies trained with both explicit and implicit curriculum learning techniques can succeed in learning multi-step complex manipulation skills. The frames of the final trained policies with TATD-HER in the simulated environment are shown in Figure 1(b). As the policy action controls the robot gripper rather than the joints, we succeed in employing the final trained policies in a real-world UR5 robot without additional training, which is shown in Figure 1(c).

Conclusion. We proposed a TATD-HER curriculum learning method to learn manipulation skills on multi-step complex tasks. The method addresses the complicated task with both explicit and implicit curriculum learning mechanisms. The experimental results demonstrate the effectiveness of our proposed TATD-HER method and show that the combination of explicit and implicit curriculum learning techniques is crucial for learning complex manipulation skills. Future work involves applying our method to more complex tasks.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61773378, U1713222, U1806204) and Equipment Pre-research Field Fund (Grant No. 61403120407).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Liu N-J, Lu T, Cai Y-H, et al. A review of robot manipulation skills learning methods. Acta Autom Sin, 2019, 45: 458–470
2 Popov I, Heess N, Lillicrap T, et al. Data-efficient deep reinforcement learning for dexterous manipulation. 2017. ArXiv: 1704.03073
3 Nair A, McGrew B, Andrychowicz M, et al. Overcoming exploration in reinforcement learning with demonstrations. In: Proceedings of IEEE International Conference on Robotics and Automation, Brisbane, 2018. 6292–6299
4 Bengio Y, Louradour J, Collobert R, et al. Curriculum learning. In: Proceedings of International Conference on Machine Learning, Montreal, 2009. 41–48
5 Fournier P, Sigaud O, Chetouani M, et al. Accuracy-based curriculum learning in deep reinforcement learning. 2018. ArXiv: 1806.09614
6 Andrychowicz M, Wolski F, Ray A, et al. Hindsight experience replay. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 5048–5058
7 Lillicrap T, Hunt J, Pritzel A, et al. Continuous control with deep reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2016
8 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. 2017. ArXiv: 1707.06347