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*
2 *University of Chinese Academy of Sciences*
3 *CAS Center for Excellence in Brain Science and Intelligence Technology*

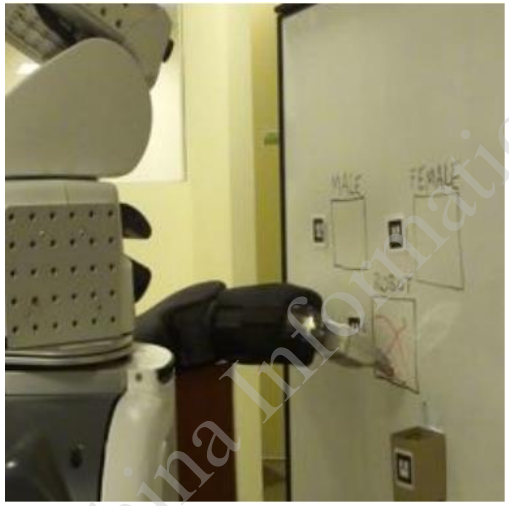
Outline

Outline

1 Introduction

Robot manipulation skill learning

Grasping



Pushing

Placing

. . .

Remarkable progress of deep reinforcement learning (DRL) in robot skill learning has recently been witnessed [1].

[1] Liu N, Lu T, Cai Y, et al. A Review of Robot Manipulation Skills Learning Methods. Acta Automatica Sinica, 2019, 45(3): 458-470.

1 Introduction

Robot manipulation skill learning

Manipulation skill learning for Multi-step complex task is still challenging

Reasons

- Higher exploration of state-action space
- Uninformative sparse rewards for policy learning

Reward Shaping [1]

Curriculum Learning [3]

Overcoming [2]

HER [4]

[1] Popov I, Heess N, Lillicrap T, et al. Data-efficient deep reinforcement learning for dexterous manipulation. arXiv preprint arXiv:1704.03073, 2017.
[2] Nair A, McGrew B, Andrychowicz M, et al. Overcoming exploration in reinforcement learning with demonstrations. IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 6292-6299.
[3] Fournier P, Sigaud O, Chetouani M, et al. Accuracybased Curriculum Learning in Deep Reinforcement Learning. arXiv preprint arXiv:1806.09614, 2018.
[4] Andrychowicz M, Wolski F, Ray A, et al. Hindsight experience replay. In Advances in Neural Information Processing Systems, 2017, 5048-5058.

1 Introduction

Robot manipulation skill learning

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)

Outline

2 Method

TATD-HER

Illustration

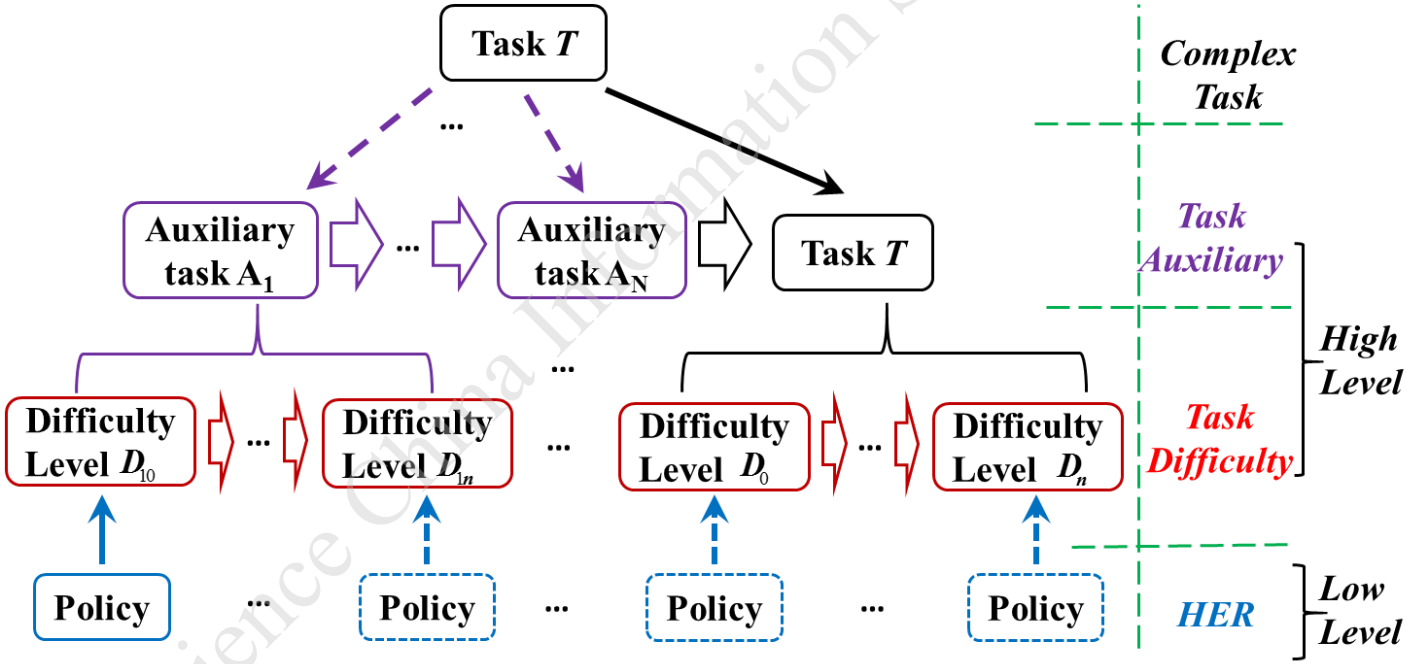


Illustration of the high-low level TATD-HER Method

2 Method

TATD-HER

Task Auxiliary

$$\{A_1, A_2, \cdots, A_N\} \leftarrow T$$

In a meaningful order
more concepts and more complex ones

Task Difficulty

$$A_1 \rightarrow \cdots \rightarrow A_N$$

More and more difficult

To auxiliary task A_k, difficulty level D_{ki}

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

D_{k0} low difficulty level D_{kn} normal difficult level

i iteration step D_{kn} maximum iteration steps for A_k

2 Method

TATD-HER

HER

$$\tau : < s_0||g, a_0, s_1||g, a_1, \cdots, s_H||g >$$

$$\tau^{her} : < s_0||g_0^h, a_0, s_1||g_1^h, a_1, \cdots, s_H||g_H^h >$$

Train policy with DDPG (Deep Deterministic Policy Gradient) [1]

Update Q-function

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

B: $(s^k, g^{kh}, a^k, r^k, s'^k, g'^{kh})$ stored transformed transiton

s^k state, g^{kh} reached position, a^k action

r^k received reward, s'^k next state, g'^{kh} next reached position

[1] Lillicrap T, Hunt J , Pritzel A, et al. Continuous control with deep reinforcement learning. International Conference on Learning Representations, 2016.

2 Method

TATD-HER

Update Q-function

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

$Q_{\phi\text{targ}}$ target Q-function $\pi_{\theta\text{targ}}$ target policy

Update Policy

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

β learning rate

2 Method

TATD-HER

Detailed algorithm

Algorithm 1 TATD-HER Curriculum Learning Method

Input: Complicated task T
Output: Policy π_θ

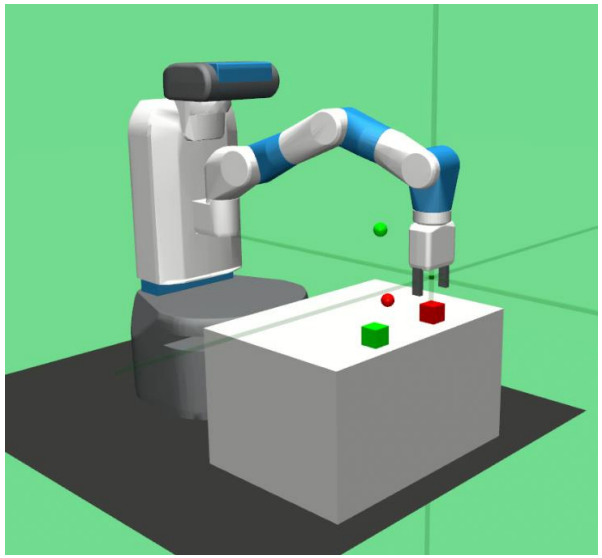
Generate auxiliary tasks in a meaningful order according to T: $\{A_1, A_2, \cdots, A_N\} \leftarrow T$.
Initialize policy π_θ with random weights.
for $k = 1 : N$ **do**
  for $i = 0 : k_{steps}$ **do**
    Set difficulty level D_{ki} for A_k with Eq. (1)
    update policy parameter θ with Eq. (4)
  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 Eq. (1)
  update policy parameter θ with Eq. (4)
end for
Policy π_θ converges on task T.

Outline

3 Experiment and Result

Experiments

Task



A sequence of skills

(1) Picking the first cube from the random position

(2) Placing the first cube on its target position on the desk

(3) Picking the second cube from the random position

4) Moving the gripper to the target position (in the air) of the second cube

Picking and placing two cubes to different target positions

Target positions denoted by the different color ball points

One target position **on the desk**, and the other one **in the air**

Interfaced with Open-AI Gym [1]

[1] Brockman G, Cheung V, Pettersson L, et al. Openai gym, 2016.

3 Experiment and Result

Experiments

Auxiliary tasks

Auxiliary Task A_1: picking the first cube (red cube) and place it on its target position, ignoring the second cube (green cube).

Reward function

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

Auxiliary Task A_2: picking two cubes and placing them on their target positions on the desk

Reward function

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

x_{1t}, x_{2t} positions of the first cube and the second cube, δ threshold

g_1, g_2 target positions of the first cube and second cube

3 Experiment and Result

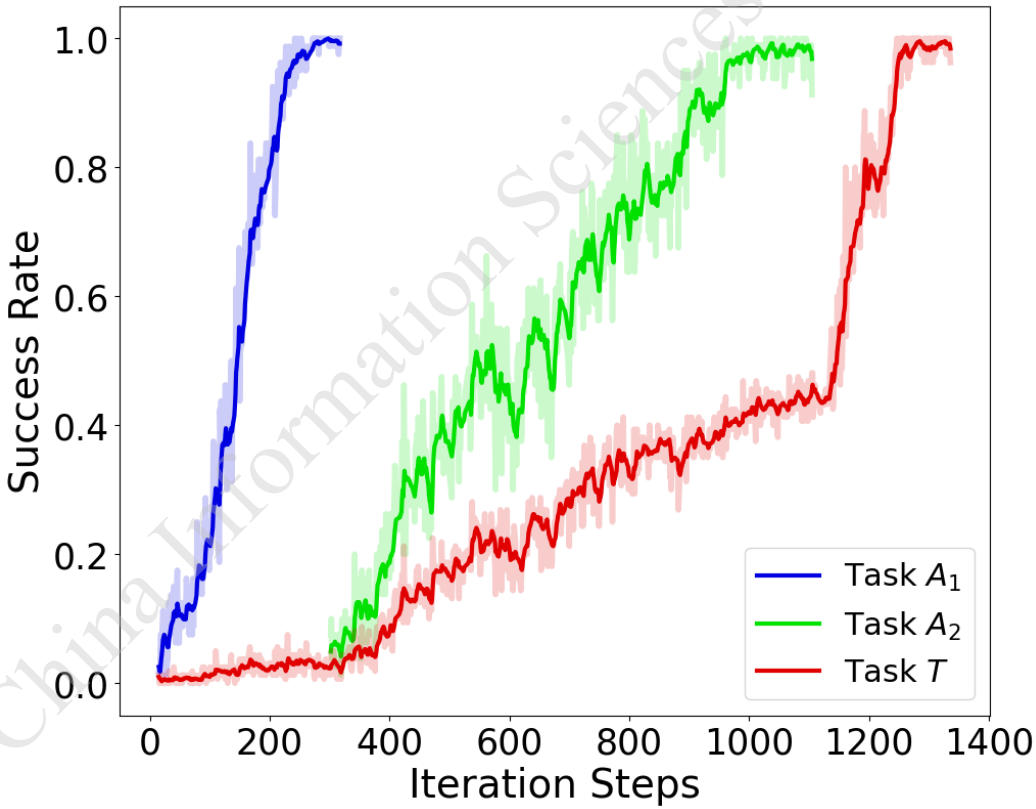
Results

Learning curves

Blue curve: for task A_1

green curve: for task A_2

red curve: for complex T



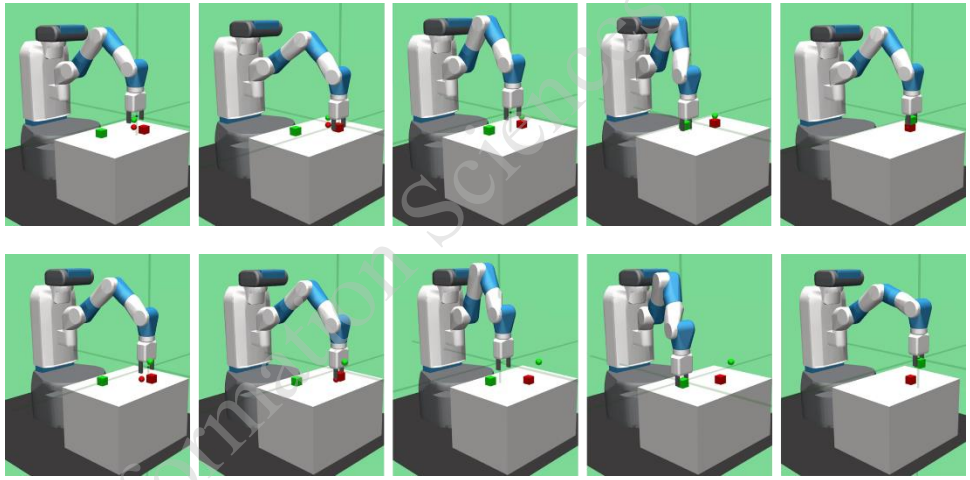
Learning curves of TATD-HER method

Our method succeed in learning policy on complex ***T***

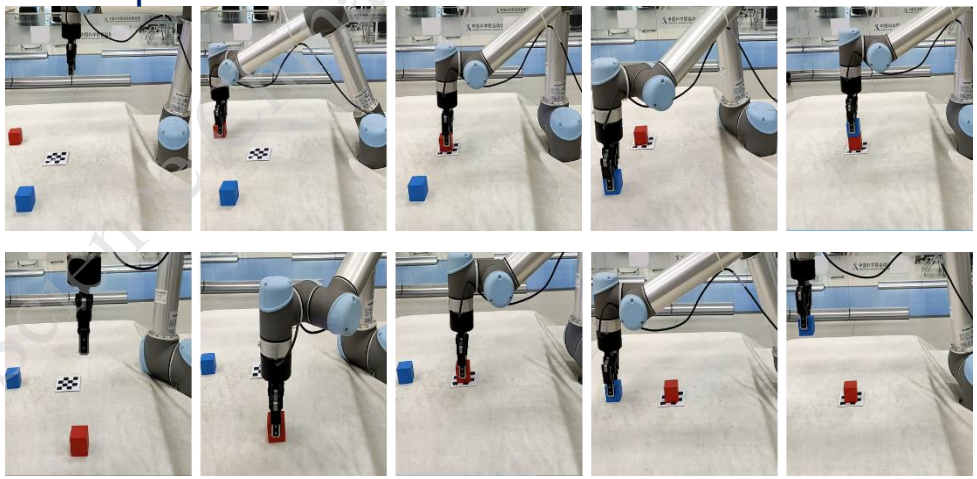
3 Experiment and Result

Results

Frames from the final trained policy
Employed in the simulated environment



As the policy action controls the robot gripper rather than the joints, we succeed in employing the final trained policies in real-world UR5 robot without additional training



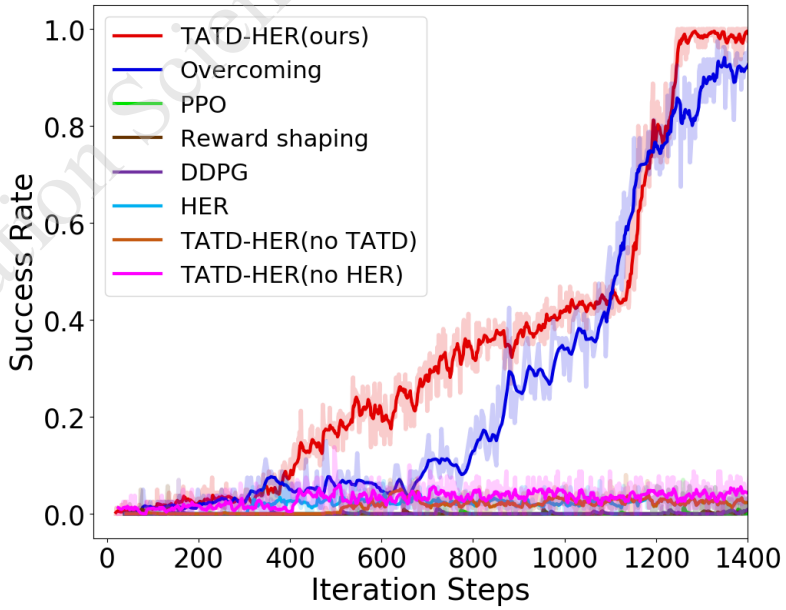
Frames from the final trained policy employed on the real-world UR5 robot

3 Experiment and Result

Experiments

- Baselines
  - Reward shaping
  - Overcoming
  - HER
  - DDPG
  - PPO
- Ablation study
  - TATD-HER(no TATD)
  - TATD-HER(no HER)

Performance comparison



Success rate of the final trained policies learned with different methods

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%

Our proposed method shows the best performance and needs no demonstration data.

Outline

3 Conclusion

Proposing a TATD-HER curriculum learning method to learn manipulation skills on multi-step complex task

Addressing the complicated task with both explicit and implicit curriculum learning mechanisms

Future works including applying our method to more various and complex tasks

Thanks