Hao-nan WANG, Ning LIU, Yi-yun ZHANG, Da-wei FENG, Feng HUANG, Dong-sheng LI, Yi-ming ZHANG, 2020. Deep reinforcement learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 21(12):1726-1744. https://doi.org/10.1631/FITEE.1900533

Deep reinforcement learning: a survey

Key words: Reinforcement learning; Deep reinforcement learning; Reinforcement learning applications

Corresponding author: Hao-nan WANG
E-mail: wanghaonan14@nudt.edu.cn
ORCID: https://orcid.org/0000-0002-0792-3858

Motivation

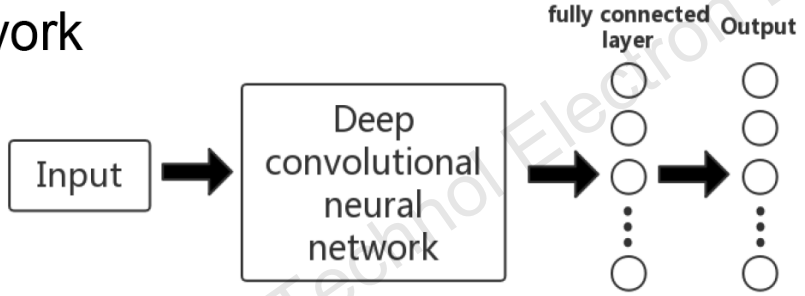
- Deep reinforcement learning (RL) is thought as one of the closest things that look anything like artificial general intelligence.
- RL is a typical representation of the closed-loop learning paradigm, which uses dynamic data and tags to bring feedback signals into the learning process. The processing and analysis differences between deep RL and traditional machine learning are huge.
- There is no complete analysis about different deep RL algorithms to show the relationship and development trend rather than focusing on one specific branch.
- There is no systematical categorization of graph workloads and applications.

Contents

1. Model-free reinforcement learning
   - RL based on the value function
   - RL based on policy gradient
2. Model-based reinforcement learning
   - Global and local models
   - Uncertainty-aware model
   - Model for complex observation
3. Advanced reinforcement learning
   - Exploration
   - Inverse reinforcement learning
   - Transfer reinforcement learning
4. Recent applications, challenges, and future

Model-free reinforcement learning

1. RL based on the value function
   - Deep Q-network



   - Developments of DQN

   We focus mainly on representative methods related to the overall structure of the system, the construction of training samples, and the structure of neural networks, including DDQN, noisy DQN, and Rainbow.

2. RL based on policy gradient
   - Improved methods based on the actor-critic
   - Improved methods based on the trust region

Model-based reinforcement learning

1. Global and local models

   For model-based algorithms, the first question is which one should be fitted if the dynamics is unknown: global dynamics models or local dynamics models? We introduce the related algorithms and the comparison between these two types of models.

2. Uncertainty-aware model

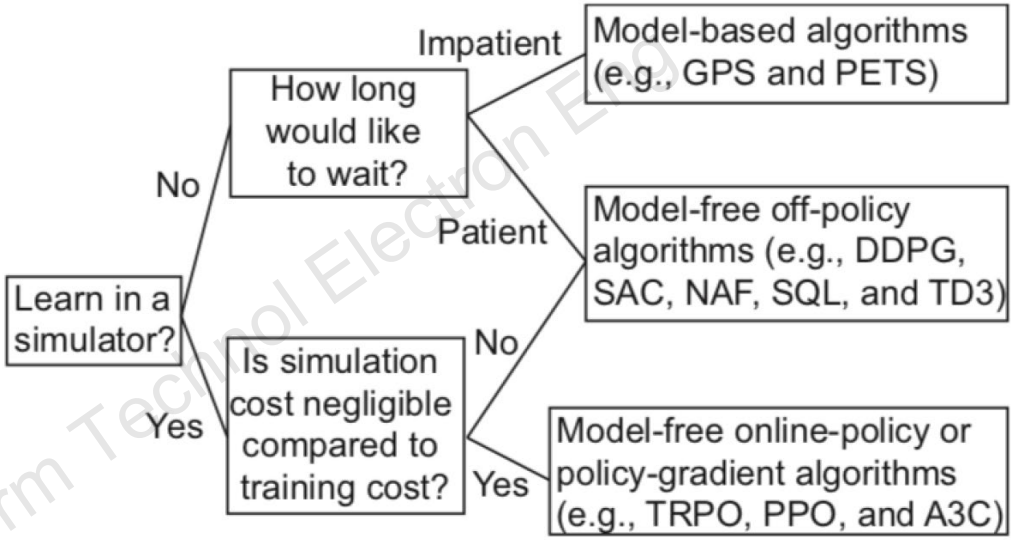
   There is a performance gap between pure-model-based and model-free methods. The uncertainty-aware model is an effective approach for solving this problem.

3. Model for complex observations

   Model-based RL is difficult to use in partially observable Markov decision processes (MDPs) with complex observations because agents have to make a decision based on the observation rather than the accurate state of the environment.

Model-based reinforcement learning

4. Different usage scenarios of methods



5. Comparison of different algorithms

Table 1 Comparison of different reinforcement learning algorithms

	Method	Reference	Number of steps	Number of episodes	Time
Model-free	Fully online (e.g., A3C)	Wang ZY et al. (2016)	100 000 000	100 000	∼15 d
	Policy gradient (e.g., TROPO)	Schulman et al. (2017)	10 000 000	10 000	∼1.5 d
	Value estimation (e.g., DDPG and SAG)	Gu SX et al. (2016)	1 000 000	1000	∼3 h
Model-based	E.g., PETS and GPS	Chua et al. (2018)	30 000	30	∼5 min

Advanced reinforcement learning

1. Exploration

   In many complex RL tasks, agents face the challenge of balancing exploration and exploitation when interacting with unknown dynamics. With the rapid development of RL, various effective and scalable approaches have been proposed to overcome the drawback of lacking exploration. We will introduce the typical algorithms of each category and analyze their advantages and disadvantages

   - Optimistic exploration
   - Posterior sampling exploration
   - Information gain exploration

Advanced reinforcement learning

2. Inverse RL

Inverse RL is introduced to learn a proper reward function from the observed expert examples. However, there are some challenges in IRL:

(1) The problem is under-defined and lacks prior knowledge;

(2) It is difficult to evaluate a learned reward;

(3) Demonstrations may not be precisely optimal.

We discuss the solutions based on maximum margin and maximum entropy.

- IRL based on maximum margin
- IRL based on maximum entropy

Advanced reinforcement learning

3. Transfer RL

- **Forward transfer**

The simplest method of forward transfer is just to try with the best hope. Policies trained for one set of circumstances might just work and could deal with new tasks successfully with good luck because sometimes there is enough variability during training to generalize.

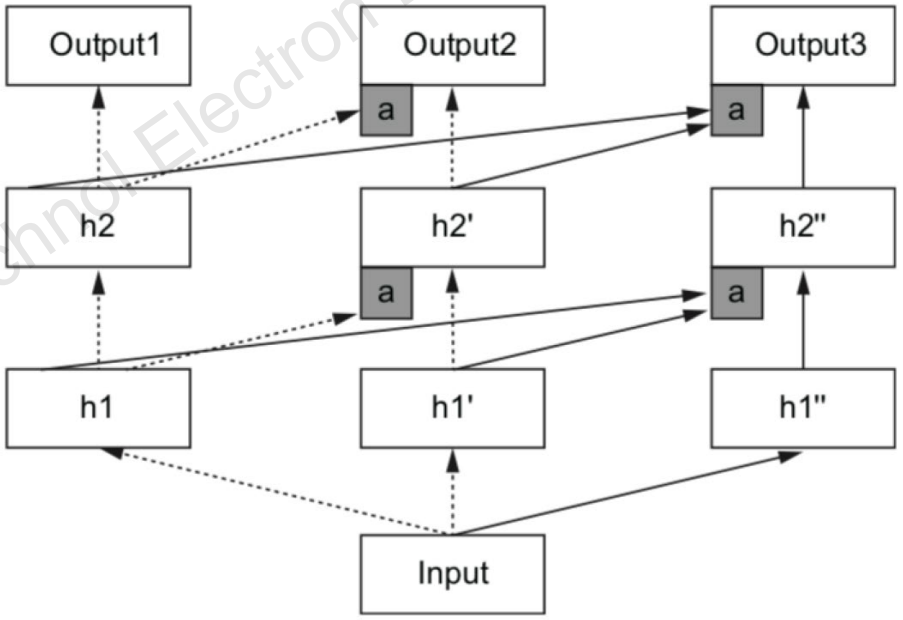


Fig. 4 A simple progressive neural network

Advanced reinforcement learning

3. Transfer RL

- **Multi-task transfer**

Typical applications of RL focus more on mastery than on one-shot learning and require a substantial number of training episodes. Multi-task transfer provides a way to address these challenges and is closer to what people do—build a lifetime of experience. One of the simplest solutions is to learn a model that can simultaneously perform many tasks.

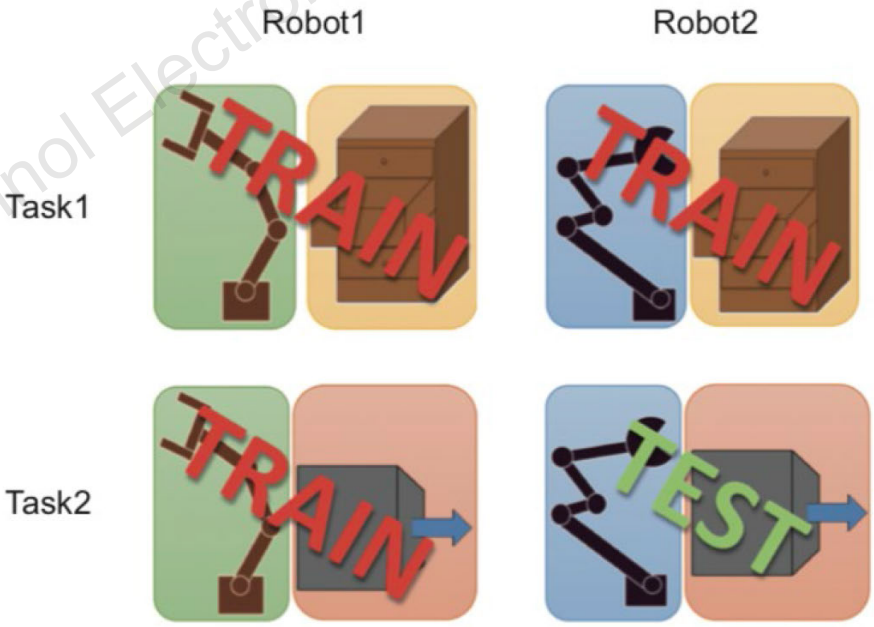


Fig. 5 Process of modular policies

Advanced reinforcement learning

3. Transfer RL

- **Meta-RL**

Meta-RL aims to improve the learning efficiency for novel subsequent tasks by learning a meta-item from the family of MDPs such as policy π_θ, model T_θ, and the reward function. We introduce two kinds of most representative meta-RL algorithms and then list some main advanced improvements.

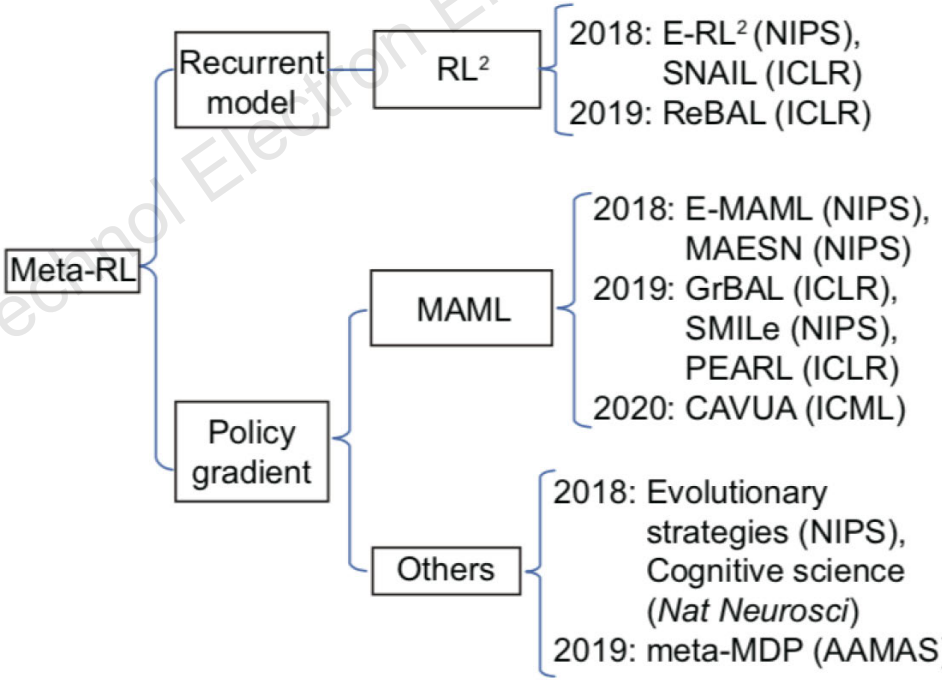


Fig. 7 Relationship between deep meta-RL algorithms

Recent applications, challenges and the future

1. Applications

   Games, robotics, natural language processing, computer systems

2. Challenges
   - Inefficient sample
   - A demanding reward function
   - Overfitting and instability

3. Future
   - Algorithms with favorable improvement and convergence are needed
   - Artificially add some supervision signals. An intrinsic reward or some auxiliary tasks can be added to increase the exploration ability
   - IRL can automatically learn the reward function, and imitation learning does not have high requirements for reward functions
   - Generalize from multi-task learning. Deep meta-RL is increasingly recognized as one of the most likely ways to implement AGI