Efficient Deep Reinforcement Learning via Adaptive Policy Transfer

Published as a conference paper in IJCAI 2020

Tianpei Yang, Jianye Hao, Zhaopeng Meng, Zongzhang Zhang, Yujing Hu,
Yingfeng Chen, Changjie Fan, Weixun Wang, Wulong Liu, Zhaodong Wang, Jiajie Peng

College of Intelligence and Computing, Tianjin University
Noah's Ark Lab, Huawei
Nanjing University
Fuxi AI Lab in Netease

Introduction

In summary, the main contributions of our work are: 1) Policy Transfer Framework (PTF) learns when and which source policy is the best to reuse for the target policy and when to terminate it by modelling multi-policy transfer as the option learning problem; 2) we propose an adaptive and heuristic mechanism to ensure the efficient reuse of source policies and avoid negative transfer; and 3) both existing value-based and policy-based Deep Reinforcement Learning (DRL) approaches can be incorporated and experimental results show PTF significantly boosts the performance of existing DRL approaches, and outperforms state-of-the-art policy transfer methods both in discrete and continuous action spaces.

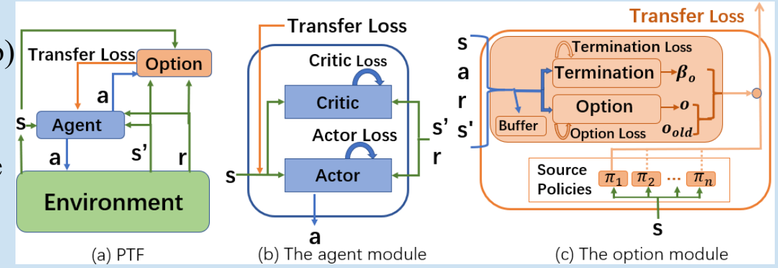
Previous studies

One major direction of previous works focuses on transferring value functions directly according to the similarity between two tasks. However, this way often assumes a well-estimated model for measurement which causes computational complexity and is infeasible in complex scenarios. Another direction of policy transfer methods focuses on selecting appropriate source policies based on the performance of source policies on the target task to provide guided explorations during each episode. However, most of these works are faced with the challenge of how to select a suitable source policy, since each source policy may only be partially useful for the target task. Furthermore, some of them assume source policies to be optimal and deterministic which restricts the generality.

How to directly optimize the target policy by alternatively utilizing knowledge from appropriate source policies without explicitly measuring the similarity is currently missing in previous work.

Proposed method

The figure (a) illustrates PTF which contains two main components, figure (b) is the agent module, which is used to learn the target policy with guidance from the option module. Figure (c) is the option module, which is used to learn when and which source policy is useful for the agent module.



For the update, PTF introduces a complementary loss, which transfers knowledge from the intra-option policy through imitation, weighted by an adaptive adjustment factor. The reuse terminates according to the termination probability and then another option is selected for reuse.

Results