Hierarchical Deep Reinforcement Learning

Seminar DRL
Lucas Brunner 17.03.2020
Goals

- What problem arise when doing HRL
- How can one solve them
- How are methods connected
4 stochastic primitive actions

Fail 33% of the time

8 multi-step options
(to each room's 2 hallways)
(OC) The option-critic architecture
Bacon, Pierre-Luc, Jean Harb, and Doina Precup. 2017
Advantage Function

\[ A(s,w) = Q(s,w) - V(s) \]

(“mesures how much better w is compared to alternative actions)

0 if w is optimal action
<0 if w is suboptimal
Figure 3: Termination probabilities for the option-critic agent learning with 4 options. The darkest color represents the walls in the environment while lighter colors encode higher termination probabilities.
Option Framework (first hour) → Learn options → Option Critic
FeUdal Networks for Hierarchical Reinforcement Learning

Vezhnevets, Alexander Sasha, et al. 2017
Option Framework (first hour) → Learn options → Option Critic → Decoupling Manager and Worker → FeUdal Net
Data-Efficient Hierarchical Reinforcement Learning (HIRO) Nachum, Ofir, et al. 2017
Does it help?

Ant Maze

Ant Push
Option Framework (first hour) → Learn options → Option Critic → Off Policy correction (keep learning while part of model changes) → HIRO Data-Efficient

Decoupling Manager and Worker → FeUdal Net

Full-state to worker
Option Framework (first hour) → Learn options → Option Critic → Off Policy correction (keep learning while part of model changes) → HIRO Data-Efficient

Option Critic → Decoupling Manager and Worker → FeUdal Net

Option Critic → HAAR

HIRO → Full-state to worker
Hierarchical Reinforcement Learning with Advantage-Based Auxiliary Rewards (HAAR)
Li, Siyuan, et al. 2019
Does it help?
Option Framework (first hour) → Learn options → Option Critic

- DAC Double Actor-Critic
- Fixed length option
- Off Policy correction (keep learning while part of model changes)

Option Critic → Decoupling Manager and Worker → HIRO Data-Efficient

- HAAR
- Auxiliary rewards from ENV

FeUdal Net

- Full-state to worker
Double Actor-Critic (DAC)
Shangtong Zhang, Shimon Whiteson 2019

Algorithm 1: Pseudocode of DAC

Input:
Parameterized $\pi, \{\pi_o, \beta_o\}_{o \in \mathcal{O}}$
Policy optimization algorithms $A_1, A_2$

Get an initial state $S_0$
$t \leftarrow 0$

while True do
    Sample $O_t$ from $\pi^H(\cdot | (O_{t-1}, S_t))$
    Sample $A_t$ from $\pi^L(\cdot | (S_t, O_t))$
    Execute $A_t$, get $R_{t+1}, S_{t+1}$

    Optimize $\pi^H$ with $(S_t^H, A_t^H, R_{t+1}, S_{t+1}^H)$ and $A_1$
    Optimize $\pi^L$ with $(S_t^L, A_t^L, R_{t+1}, S_{t+1}^L)$ and $A_2$

    $t \leftarrow t + 1$
end
Thank You
Advantage Function

- Do we really need to compute both