Meta Reinforcement Learning

Ruoheng Ma
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: ukeyt@student.kit.edu

Meng Zhang
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: @student.kit.edu

***Abstract*—Meta reinforcement learning is a research area of reinforcement learning. It applies meta learning approaches on reinforcement learning tasks to generalize the training result of reinforcement learning and adapt it to unseen tasks. Currently, multiple meta reinforcement learning methodologies are being researched, such as optimization-based approach, hyperparameter-based approach and so on. In this report, we first provide some fundamental knowledge about meta reinforcement learning. Then we give an overview of state-of-the-art meta reinforcement learning algorithms. Finally, a couple of algorithms are discussed in detail and their experimental results are also presented.**

Disclaimer: Section...by Ruoeheng Ma and Section...by Meng Zhang.

I. Introduction

This demo file is intended to serve as a “starter file” for IEEE conference papers produced under LaTeX using IEEEtran.cls version 1.8b and later. I wish you the best of success.

II. Fundamental Knowledge

In this section, we provide some fundamental knowledge for ease of understanding the algorithms discussed in the following sections.

A. Formulation of Reinforcement Learning and Meta Learning

In reinforcement learning, a value function

$$G_t =$$

no predefined value function is provided to tell the score of an action. In practice, a proxy of the true value function, known as a *return*, is used for estimation and optimization for the value function.

B. Hyperparameter

C. empty

III. Meta-Gradient Reinforcement Learning

One of the algorithm in meta reinforcement learning is the meta-gradient reinforcement learning introduced in [1] by Google DeepMind. In deep reinforcement learning, the environment model is often unknown to the agent. Therefore, the true value function has to be approximated by a function known as *return* with parameter θ. This return function plays an important role in determining the characteristics of the reinforcement learning algorithm. Usually, the return function is parameterised by a discount factor γ and the bootstrapping parameter λ. In [1], it is defined as shown below:

$$G_\eta^{(n)}(\tau_t) = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n v_\theta(s_{t+n})$$

$$G_\eta^\lambda(\tau_t) = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_\eta^{(n)}$$

The two parameters are hand-selected and held fixed throughout training. In [1], these parameters are denoted by $\eta = \{\gamma, \lambda\}$ and θ is updated by the following formula, taking η into account:

$$\theta' = \theta + f(\tau, \theta, \eta)$$

where τ is the sample of experience being considered.

In order to achieve better performance, η is also updated in the training process. A meta-objective defined as follows is used to define the rule to update η:

$$\bar{J}(\tau', \theta', \bar{\eta}) = (g_{\bar{\eta}}(\tau') - v'_\theta(S'))^2$$

where τ' is the next sample and θ' is the updated parameter θ.

$\Delta\eta$ is defined below:

$$\Delta\eta = -\beta \frac{\partial \bar{J}(\tau', \theta', \bar{\eta})}{\partial \theta'} \frac{\partial f(\tau, \theta, \eta)}{d\eta}$$

where β is the learning rate for updating η.

The validation of meta-gradient algorithm is executed on Arcade Learning Environment with an agent built with the IMPALA framework. The agents are evaluated on 57 different Atari games. The result is shown below:

	η	Human starts		No-op starts	
		$\gamma = 0.99$	$\gamma = 0.995$	$\gamma = 0.99$	$\gamma = 0.995$
IMPALA	$\{\}$	144.4%	211.9%	191.8%	257.1%
Meta-gradient	$\{\lambda\}$	156.6%	214.2%	185.5%	246.5%
		$\bar{\gamma} = 0.99$	$\bar{\gamma} = 0.995$	$\bar{\gamma} = 0.99$	$\bar{\gamma} = 0.995$
Meta-gradient	$\{\gamma\}$	233.2%	267.9%	280.9%	275.5%
Meta-gradient	$\{\gamma, \lambda\}$	221.6%	292.9%	242.6%	287.6%

Fig. 1. Experiment result of meta-gradient algorithm. ”Human starts” means episodes are initialized to a state that is randomly sampled from human play, while ”No-op starts” means each episode is initialized with a random sequence of no-op actions.

IV. MAML

V. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] Zhongwen Xu, Hado van Hasselt and David Silver, *Meta-Gradient Reinforcement Learning*. URL:https://proceedings.neurips.cc/paper/2018/file/2715518c875999308842e3455eda2fe3-Paper.pdf