传统强化学习算法从特定的环境中进行学习，当智能体根据环境训练完成后，若环境发生内部变化，则难以适应于

元强化学习

元学习的目标是在各种学习任务上训练一个模型，这样它就可以只使用少量的训练样本来解决新的学习任务。

有监督的问题的训练过程中，损失函数如下式所示。



和分别是从任务中采样得到的输入和输出，从而获得了任务的损失，最小化损失，以对回归模型进行训练。

为每个任务更新过程的步长，每个函数的损失值为



元学习的目标是获得以下函数的最小化。



参数更新的过程中，



输入：元神经网络参数

输出：

1.随机初始化元神经网络参数

2.循环迭代一定次数

3. 从中采样任务

4. 对每个训练得到一个参数

5. 对采样的任务进行参数的更新

6.返回元学习参数

通过元学习获取到的参数由于最小化了多个任务的目标函数，因而最终得到的神经网络对各个任务更加敏感，通过少量训练即可拟合到单个任务中。

由于基于策略梯度的强化学习算法在计算价值函数的目标值时，采用从后向前的过程，得到的状态价值属于无偏估计，因此应用效果更佳。而基于值迭代的强化学习算法得到的价值估计，具有方差较小，偏差较大的特征，因此在应用元学习方法上需要进行改进。

由于交通环境模型的建立是为了最大化路口吞吐量，缓解交通阻塞。出于这样的目的，去掉了相位设置的不同分类，采用较为通用的8相位设计，使得智能体决策空间较为自由。而元强化学习在应用过程中，在各个任务之间的切换过程时，动作空间的表示一致性，使得最小化多任务之间的时序差分误差时，神经网络的设计和表达更合理。

相位设置：

|  |  |  |
| --- | --- | --- |
| 原始相位 | 拓展相位-1 | 拓展相位-2 |
| ①⑦ | ①② | ①⑥ |
| ③⑤ | ③④ | ③② |
| ②⑧ | ⑤⑥ | ⑤⑧ |
| ④⑥ | ⑦⑧ | ⑦④ |



强化学习应用时，环境搭建和模拟是一个不可或缺的部分。

代码进行了大量的清洗和重构。

分层次的结构调用

参数书写的一致性（形参书写顺序的一致性）

遵循的编程基本规范

1可读性

2可拓展性

Gym提供的强化学习环境表示出的马尔科夫决策过程中，至少需要包括以下函数，。因此在设计过程中，为了增加可读性，则需要将同样的函数接口提供，以降低对可读性的影响。

|  |  |  |
| --- | --- | --- |
| Reset | Step（action） |  |
|  |  |  |
|  |  |  |
|  |  |  |

通过类的继承关系，重新定义生成器、更新器。

保持灵活性

Cityflow对windows操作系统的支持性不佳。

部署方式：



在模拟数据和真实数据中进行过了实验。

最大压力方法：

pressure Pi

indicates the degree of disequilibrium

between the incoming and outgoing vehicle density

自适应交通信号控制方法

韦伯斯特配时法

在单个交叉口上，状态的更复杂表达不总是有用的【arxiv】，在多个交叉口。也可能没有必要：

. Diagnosing Reinforcement Learning for Traffic Signal Control

Q-Learning使用的

Baher Abdulhai, Rob Pringle, and Grigoris J Karakoulas. Reinforcement learning for true adaptive traffic signal control

Samah El-Tantawy and Baher Abdulhai. An agent-based learning towards decentralized and coordinated traffic signal control.

交通模型的状态定义：

全红时间

车道

道路

等。。。

公式化马尔科夫交通信号状态定义