Keras深度强化学习--Actor-Critic实现

洛荷 (/u/51ee4397ee8b) +关注
0.1 2019.01.04 13:18* 字数 728 阅读 84 评论 0 喜欢 1
(/u/51ee4397ee8b)

AC算法（Actor-Critic）架构可以追溯到三、四十年前， 其概念最早由Witten在1977年提出，然后Barto, Sutton和Anderson等在1983年左右引入了actor-critic架构。AC算法结合了value-based和policy-based方法，value-based可以在游戏的每一步都进行更新，但是只能对离散值进行处理；policy-based可以处理离散值和连续值，但是必须等到每一回合游戏结束才可以进行处理。而AC算法结合两者的优点，既可以处理连续值又可以单步更新。

Paper:
Witten（1977）: An adaptive optimal controller for discrete-time Markov environments (https://www.sciencedirect.com/science/article/pii/S0019995877903540)
Barto（1983）: Neuronlike adaptive elements that can solve difficult learning control problems (http://dl.acm.org/citation.cfm?id=104432)
Advantage Actor Critic (A2C): Actor-Critic Algorithms (https://papers.nips.cc/paper/1786-actor-critic-algorithms.pdf)

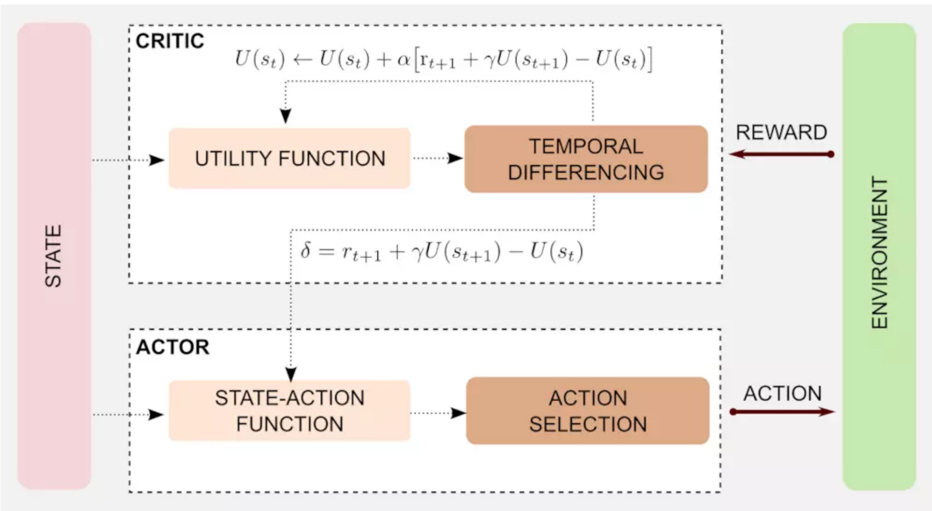
Github: https://github.com/xiaochus/Deep-Reinforcement-Learning-Practice (https://github.com/xiaochus/Deep-Reinforcement-Learning-Practice)

环境

- Python 3.6
- Tensorflow-gpu 1.8.0
- Keras 2.2.2
- Gym 0.10.8

算法原理

AC算法的结构如下图所示。在AC中，policy网络是actor（行动者），输出动作（action-selection）。value网络是critic（评价者），用来评价actor网络所选动作的好坏（action value estimated），并生成TD_error信号同时指导actor网络的更新。 在这里我们引入DNN模型作为函数近似。



Actor-Critic

Actor-Critic的实现流程如下：

Actor看到游戏目前的state，做出一个action。
Critic根据state和action两者，对actor刚才的表现打一个分数。
Actor依据critic（评委）的打分，调整自己的策略（actor神经网络参数），争取下次做得更好。
Critic根据系统给出的reward（相当于ground truth）和其他评委的打分（critic target）来调整自己的打分策略（critic神经网络参数）。
一开始actor随机表演，critic随机打分。但是由于reward的存在，critic评分越来越准，actor表现越来越好。

One-step Actor–Critic (episodic), for estimating $\pi_{\boldsymbol\theta} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol\theta)$
Input: a differentiable state-value function parameterization $\hat v(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol\theta} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol\theta \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^d$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
  Initialize S (first state of episode)
  $I \leftarrow 1$
  Loop while S is not terminal (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol\theta)$
    Take action A, observe S', R
    $\delta \leftarrow R + \gamma \hat v(S',\mathbf{w}) - \hat v(S,\mathbf{w})$ (if S' is terminal, then $\hat v(S',\mathbf{w}) \doteq 0$)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} I \delta \nabla \hat v(S,\mathbf{w})$
    $\boldsymbol\theta \leftarrow \boldsymbol\theta + \alpha^{\boldsymbol\theta} I \delta \nabla \ln \pi(A|S,\boldsymbol\theta)$
    $I \leftarrow \gamma I$
    $S \leftarrow S'$

Algorithm

AC算法的关键问题在于使用critic引导actor的更新。在Policy Network中，我们使用每一轮游戏的discount reward来引导策略模型的更新方向；在AC中，discount reward被替换为critic的Q值。在AC中critic的学习率要高于actor的学习率，因为我们需要让critic学习的比actor快，以此指导actor的更新方向。

算法实现

keras实现的的AC如下所示：

```
# -*- coding: utf-8 -*-
import os

import numpy as np

from keras.layers import Input, Dense
from keras.models import Model
from keras.optimizers import Adam
import keras.backend as K

from DRL import DRL


class AC(DRL):
    """Actor Critic Algorithms with sparse action.
    """
    def __init__(self):
        super(AC, self).__init__()

        self.actor = self._build_actor()
        self.critic = self._build_critic()

        if os.path.exists('model/actor_acs.h5') and os.path.exists('model/critic_acs.
            self.actor.load_weights('model/actor_acs.h5')
            self.critic.load_weights('model/critic_acs.h5')

        self.gamma = 0.9

    def _build_actor(self):
        """actor model.
        """
        inputs = Input(shape=(4,))
        x = Dense(20, activation='relu')(inputs)
        x = Dense(20, activation='relu')(x)
        x = Dense(1, activation='sigmoid')(x)

        model = Model(inputs=inputs, outputs=x)

        return model

    def _build_critic(self):
        """critic model.
        """
        inputs = Input(shape=(4,))
        x = Dense(20, activation='relu')(inputs)
        x = Dense(20, activation='relu')(x)
        x = Dense(1, activation='linear')(x)

        model = Model(inputs=inputs, outputs=x)

        return model

    def _actor_loss(self, y_true, y_pred):
        """actor loss function.

        Arguments:
            y_true: (action, reward)
            y_pred: action_prob

        Returns:
            loss: reward loss
        """
        action_pred = y_pred
        action_true, td_error = y_true[:, 0], y_true[:, 1]
        action_true = K.reshape(action_true, (-1, 1))

        loss = K.binary_crossentropy(action_true, action_pred)
        loss = loss * K.flatten(td_error)

        return loss

    def discount_reward(self, next_states, reward, done):
        """Discount reward for Critic

        Arguments:
            next_states: next_states
            rewards: reward of last action.
            done: if game done.
        """
        q = self.critic.predict(next_states)[0][0]

        target = reward
        if not done:
            target = reward + self.gamma * q

        return target
```

```
    def train(self, episode):
        """training model.

        Arguments:
            episode: ganme episode

        Returns:
            history: training history
        """
        self.actor.compile(loss=self._actor_loss, optimizer=Adam(lr=0.001))
        self.critic.compile(loss='mse', optimizer=Adam(lr=0.01))

        history = {'episode': [], 'Episode_reward': [],
                   'actor_loss': [], 'critic_loss': []}

        for i in range(episode):
            observation = self.env.reset()
            rewards = []
            alosses = []
            closses = []

            while True:
                x = observation.reshape(-1, 4)
                # choice action with prob.
                prob = self.actor.predict(x)[0][0]
                action = np.random.choice(np.array(range(2)), p=[1 - prob, prob])

                next_observation, reward, done, _ = self.env.step(action)
                next_observation = next_observation.reshape(-1, 4)
                rewards.append(reward)

                target = self.discount_reward(next_observation, reward, done)
                y = np.array([target])

                # loss1 = mse((r + gamma * next_q), current_q)
                loss1 = self.critic.train_on_batch(x, y)
                # TD_error = (r + gamma * next_q) - current_q
                td_error = target - self.critic.predict(x)[0][0]

                y = np.array([[action, td_error]])
                loss2 = self.actor.train_on_batch(x, y)

                observation = next_observation[0]

                alosses.append(loss2)
                closses.append(loss1)

                if done:
                    episode_reward = sum(rewards)
                    aloss = np.mean(alosses)
                    closs = np.mean(closses)

                    history['episode'].append(i)
                    history['Episode_reward'].append(episode_reward)
                    history['actor_loss'].append(aloss)
                    history['critic_loss'].append(closs)

                    print('Episode: {} | Episode reward: {} | actor_loss: {:.3f} | cr

                    break

        self.actor.save_weights('model/actor_acs.h5')
        self.critic.save_weights('model/critic_acs.h5')

        return history


if __name__ == '__main__':
    model = AC()

    history = model.train(300)
    model.save_history(history, 'ac_sparse.csv')

    model.play('ac')
```

游戏结果如下：

```
play...
Reward for this episode was: 137.0
Reward for this episode was: 132.0
Reward for this episode was: 144.0
Reward for this episode was: 118.0
Reward for this episode was: 124.0
Reward for this episode was: 113.0
Reward for this episode was: 117.0
Reward for this episode was: 131.0
Reward for this episode was: 154.0
Reward for this episode was: 139.0
```

从上述实验可以看出，AC算法能够对这个问题进行优化但是模型收敛的并不稳定，效果也无法达到最优。这是因为单纯的AC算法属于on-policy方法，Actor部分的效果取决于Critic部分得到的td_error。在没有采取任何优化措施的情况下，DQN很难收敛由此导致整个AC算法无法收敛。

被以下专题收入，发现更多相似内容

人工智能/模式... (/c/257bcc1383e2?utm_source=desktop&utm_medium=notes-included-collection)

机器学习与数据挖掘 (/c/9ca077f0fae8?utm_source=desktop&utm_medium=notes-included-collection)

推荐阅读 更多精彩内容 > (/)

祝大家新年快乐，万事如意！ (/p/e2b7e22f460a?utm... (/p/e2b7e22f460a?utm_campaign=maleskine&utm_content=note&utm

我的新年贺词！ 原创： 晨风 早上从4:32分的第一阵鞭炮声里开始，爆竹声中一岁除，此起彼伏的噼里啪啦惊醒着沉睡中的黎明。 春的气息闻声而至，年的...

晨风_d76b (/u/b89ab5bbd9b2?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

大宋第一古惑仔，后来怎样了？（上） (/p/43ff4d57a0... (/p/43ff4d57a09d?utm_campaign=maleskine&utm_content=note&utm

文/麦大人 01 少年不识愁滋味，爱上层楼，爱上层楼，为赋新词强说愁。而今识尽愁滋味，欲说还休，欲说还休，却道天凉好个秋。 这阙词名叫《丑奴儿·书...

麦大人 (/u/2b3ad4f2a058?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

知否？知否？喝酒、投壶、打马球、带领粉丝团嗨飞...... (/p/9063cc0a3a32?utm_campaign=maleskine&utm_content=note&utm

原创：宁许砍柴书院1月22日 " 阅读和写作是一种力量 不限于表达自我 也不止于赚钱养家 " ——砍柴书院 电视剧《知否？知否？应是绿肥红瘦》正在热播...

李砍柴 (/u/f3092432a535?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

过年好 (/p/6d403c950c40?utm_campaign=maleskin... (/p/6d403c950c40?utm_campaign=maleskine&utm_content=note&utm

回到老家。 首先，跃入眼帘的是门前的这一条正在重新兴修之中的水渠。小时候，每逢到了夏天，这一条水渠沟就是我们儿时的乐园。 我赶紧放下行李。 ...

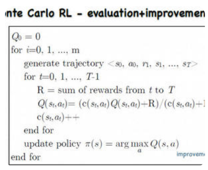
静夜风 (/u/2d20271ea0df?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

这些话，给受委屈的你 (/p/b51486410a82?utm_campaign=maleskine&ut...

转自：许莫私人音乐厅作品 其实再坚强的人都有脆弱的时候 我们总是去安慰别人 不要难过不要伤心 却忘了有时候我们也需要别人的安慰 人这一辈子会遇见一些人 遭遇一些事 这些人和事会让你心烦难过 让你受尽...

月宸 (/u/5aa13414a08c?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

(/p/acb90892c288?



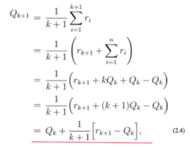
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

深度强化学习（理论篇）—— 从 Critic-only、Actor-only 到 Actor-... (/p/a...

来源于 Tangowl 的系列文章 https://blog.csdn.net/lipengcn/article/details/81253033 自己第一篇 paper 就是用 MDP 解决资源优化问题，想来那时写个东西真是艰难啊。 彼时倒没想到这个数学工具，如今会这么火...

TangowL (/u/2f6ae386f03e?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/7a9f9225e2b2?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

增强学习（一） (/p/7a9f9225e2b2?utm_campaign=maleskine&utm_con…

一. 增强学习简介 1.1 什么是增强学习？ 机器学习的算法可以分为三类：监督学习，非监督学习和增强学习。 增强学习也称为强化学习。 增强学习就是将情况映射为行为，也就是去最大化收益。学习者并不是被…

阿阿阿阿毛 (/u/a18653721b40?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/0be451b84e06?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

迷雾探险3 | 强化学习入门 (/p/0be451b84e06?utm_campaign=maleskine…

看完《迷雾探险2》的深度学习入门，又发现了一些不错的文章： 通俗易懂的深度学习发展介绍：从最基本的神经网络算法（单个神经元模型的提出，到两层和三层的简单神经网络，到BP算法，到深度神经网络…

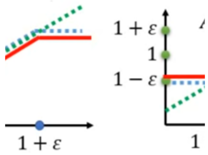
臻甄 (/u/81813c5cbb49?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

Reinforcement Learning (/p/ca94ca96fe0f?utm_campaign=maleskine&…

-- <center>tabular Q learning</center> [图片上传失败...(image-c02beb-1523787891898...

建怀 (/u/750352d3153a?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/dcfd927e7598?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

精简强化学习总结 (/p/dcfd927e7598?utm_campaign=maleskine&utm_c…

强化学习 元素：actor(我们可以控制，决策我们的行为)，Env，Reward（我们不能控制环境）主要方法：model-baed（对Env建模，actor可以理解环境），model-free(policy-based，value-based)；on-policy（学…

fada_away (/u/41ff6bc1633b?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

［笔记7］JavaScript DOM编程艺术_图片库改进版 (/p/875a7ed300ba?ut…

勤于思考是每位有创新精神的网页设计人员都应该具备的特质。 平稳退化 第一个问题，如果JS功能被禁用，会怎么样？把href属性设置为一个真实存在的值，能够让之前提到的图片库平稳退化。 JS与HTML标…

fumier (/u/d725f231c254?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

你说，结婚到底是为了什么！ (/p/631fdde71364?utm_campaign=maleski…

中午起床准备上班的时候，好同学兼好闺蜜发了条微信，说好烦躁！我说为什么呀，心里想不是还怀着2胎吗？心情不好可直接影响宝宝情况呢！闺蜜娓娓道来，他们两夫妻是在社交网站相识，交往了一段时间之…

一个有梦想的人 (/u/0bb4f1536502?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/c036dd71c622?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

爱站简报No.2 (/p/c036dd71c622?utm_campaign=maleskine&utm_cont…

每日一报 在这个难得的假期里，大家一定会想出去好好玩一玩，好好走一走吧！但出门前大家别忘了查一下天气预报喔！ 每日一看 放假了，宅在家里的小伙伴们是不是在苦恼该干些什么呢？别急，小编给你送电影…

PeerlessL (/u/b987465359bf?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/b768031c2123?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

马上封侯 (/p/b768031c2123?utm_campaign=maleskine&utm_content=…

这件料子贵气的紫色，浓烈的黄色。光是色彩就可以品玩很久。直到看到苍山上的五彩祥云，恍然大悟，正是这料子的色彩。 结合着这料子上浓重的黑点，邵师觉得做黑脸的猴，而彩色的部分雕成天马，会把料子…

阿荣雕玉 (/u/842ccd4ba463?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/b8f79caee6d4?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

“风情马套”山东诗人跨年诗会作品欣赏（8）------王强篇 (/p/b8f79caee6d4…

假如是真的 我将何去何从 是依然固执的走下去 还是迅速的逃离 假如是真的 你将又会向何方 是重回老路 还是背井离乡 假如是真的 上天将不会再给机会 我能抓住一根稻草 你将随风无遮无挡 假如是真的 我愿意在马…

山人周永 (/u/a474b05b0979?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)