Actor-Critic Algorithm

AAIS in PKU　陈伟杰 1901111420

April 24, 2020

1 Problem Setting

已知足球环境 google football，采用 Simple115 的向量表示状态，行动空间采取默认的 19 维行动空间。运用 Actor-Crtic 方法对 academy_empty_goal 问题进行求解：其中 Actor 和 Critic 的损失函数 $\mathcal{L}_A$,$\mathcal{L}_C$ 分别如下：

$$\begin{aligned} &\mathcal{L}_C(\theta) = \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}\left[\delta^2\right] \\ &\mathcal{L}_A(\phi) = \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}[\delta logP(s,a;\phi)] \\ &\text{where} \quad \delta = r + \gamma Q\left(s',a';\theta\right) - Q(s,a;\theta) \end{aligned} \tag{1}$$

2 Experiment Result

实验中参数的设置可以查看 data.py，Actor 和 Critic 函数都采用 2 层全链接网络，Actor 网络最后引入 softmax 函数进行归一化使其成为动作策略的概率分布，并且均采用 Adam 算法来优化损失函数。

对于 academy_empty_goal,为了减少无用操作从而减轻训练难度,因此将行动域限定为 [3,4,5,6,7,12,13,14,15,17,18]，即 top，top-right，right，bottom-right，bottom，shot，sprint，release-direction，releas-sprint，dribble，release-dribble。

数值结果如下图，其中 entropy 使用环境的初态计算动作策略的香农熵，即

$$\mathcal{S}[P] = -\sum_{i=1}^{n=19} P_i logP_i \tag{2}$$

reward 采用滑动平均的处理方法使得图线更直观（即 $\hat{r}[i] = \frac{1}{k}\sum_{j=i}^{i+k} r[j]$）

从图线可以看出，随着 episode 的增加，滑动平均的 reward 逐渐上升，entropy 明显下降。entropy 的下降意味着从无序趋于有序，即动作策略的概率分布信息量增加，具有实际场景的特征。

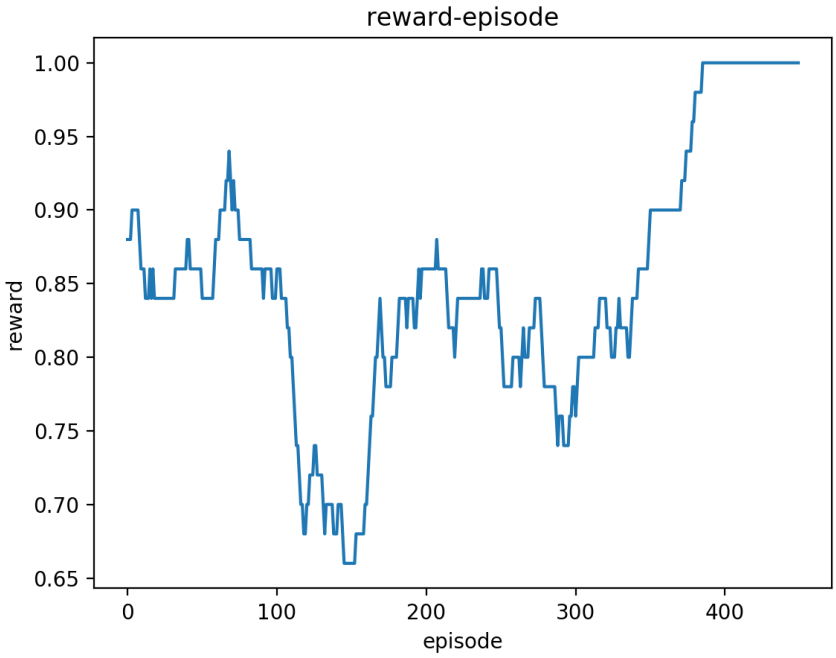

Figure 1: reward 与迭代次数的关系图

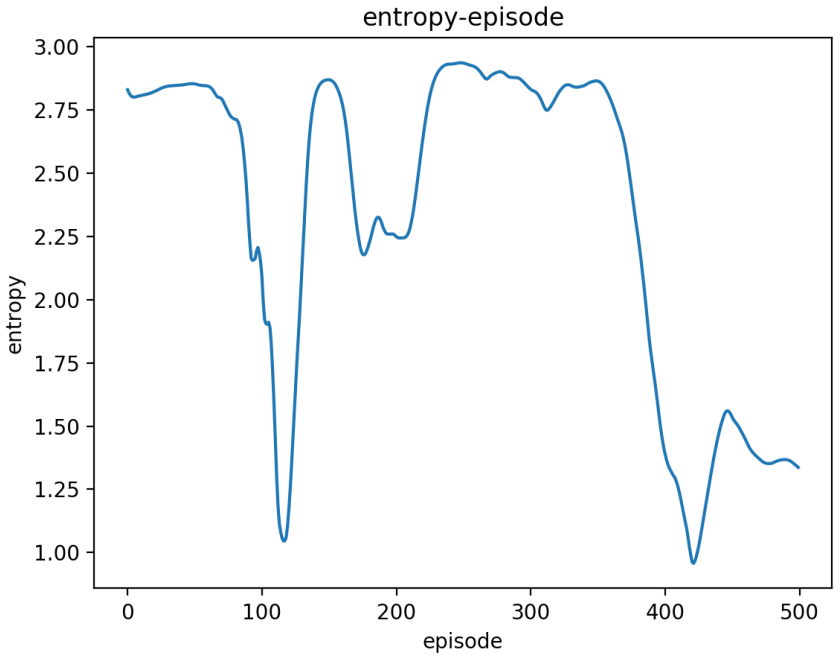

Figure 2: entropy 与迭代次数的关系图

3 README

代码包括 data.py，model.py 和 train.py 三个程序，需要 google football 环境。直接运行 train.py 可以得到上述结果

- 电脑：MacBook Pro(15-inch, 2019)
- 处理器：2.6 GHz 6-Core Intel Core i7
- 内存：16 GB 2400 MHz DDR4
- 操作系统：macOS Catalina version 10.15.1
- 语言：python3.6
- 实验环境：docker+ubuntu18.04