Synthetic Returns for Long-Term Credit Assignment —— hoho

论文试图解决什么问题？

Credit assignment问题

这是否是一个新的问题？

否

这篇文章要验证一个什么科学假设？

是否可以建立任意时间步的状态与未来奖励的联系，让agent学习这种联系？

有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

hoho_todo

论文中提到的解决方案之关键是什么？

提出SA learning方法（state-associations learning）, 用历史的状态预测未来状态的reward。当建立了历史与未来的这种联系后，可直接跳过中间的事件，将信用（预测到的reward）直接分配给先前（历史）的状态。

定义损失函数：

$$Loss_{c,g,b} = \|r_t - g(s_t)\sum_{k=0}^{t-1} c(s_k) - b(s_t)\|$$

本质是用神经网络拟合立即回报：r_t是立即回报，$g(s_t)\sum_{k=0}^{t-1} c(s_k) + b(s_t)$是神经网络的部分

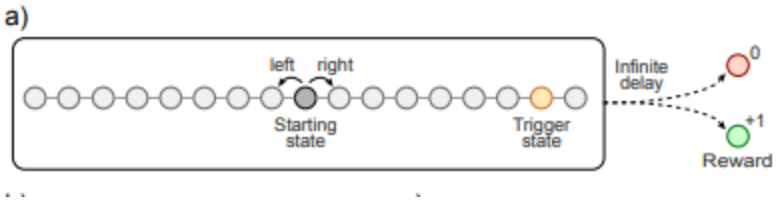
其中：

- $g(s_t)$是sigmoid函数，输出值在$[0, 1]$，其可看作当前回报跟历史状态回报的权重
- $c(s_t)$是其中一个神经网络，是整个架构的核心，其输出对应状态下的reward，可以把它看作是一种优势值的估计：当经历的当前状态后，对于未来状态的影响是好还是坏。另外，它和$g(s_t)$的线性组合把reward的预测建模为线性回归问题。
- $b(s_t)$也是一个神经网络，也是计算当前状态对于当前回报的贡献度，目的是鼓励模型更倾向于使用当前状态预测当前回报（what？）

神经网络架构可用CNN编码状态，LSTM作为策略网络

论文中的实验是如何设计的？

设计了一个Chain Task的实验：智能体可以左右移动，直到移动到目标点结束，才获得奖励+1，否则奖励一直为0

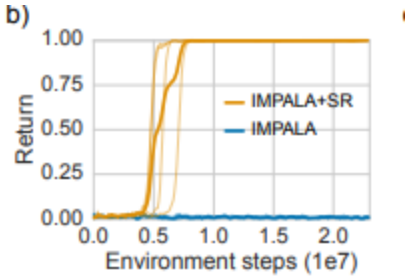


用于定量评估的数据集是什么？代码有没有开源？

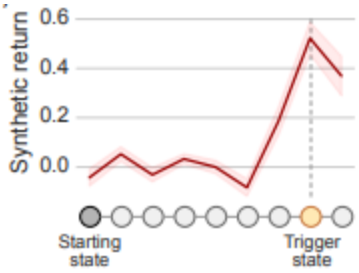
无

论文中的实验及结果有没有很好地支持需要验证的科学假设？

下图表明，使用了SA方法的智能体可以比无使用SA的算法更快的完成任务。



作者还展示了使用SA模型生成的reward分布情况，进一步说明，SA实现了credit assignment 回报状态（the rewarding state）到相应动作的关联：



作者还使用的真实的游戏环境进行测试：

- Catch with Delayed Rewards （Catch 的改进）：接球游戏
- Key-to-Door：其中加入一些跟最终目标无关的干扰动作，考察第一步寻找Key的动作是否与最后开门的动作的重要相关性
- Pong：双方接发球游戏，球来回弹，直到一方没接
- Atari Skiing

以上游戏也显示SA方法优越性的结果。

这篇论文到底有什么贡献？

SA方法显示了可以提升深度强化学习效率

下一步呢？有什么工作可以继续深入？

提到SA方法的一些局限性：

1. we expect the performance of the gated SA architecture to suffer when the environment is not sparse.
2. the current method is not sensitive to the number of times a predicted state occurs.
3. because of our use of a multiplicative gate, we cannot offer convergence guarantees as to the semantics of $c(s_t)$. Additionally, because of our use of an additive regression model, we cannot
   offer a rigorous guarantee of optimal credit assignment,
4. indeed when multiple states predict the same reward, which state “gets credit” is unconstrained.