Optimizing Agent Behavior over Long Time Scales by Transporting Value —— hoho

论文试图解决什么问题？

长时间尺度的credit assignment问题。“How to evaluate the utility of the actions within a long-duration behavioral sequence leading to success or failure in the task”

这是否是一个新的问题？

不

这篇文章要验证一个什么科学假设？

使用神经网络的中的注意力机制去credit过去久远的动作的reward。（uses neural network attentional memory mechanisms to credit distant past actions for future rewards.）

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？

总体流程：

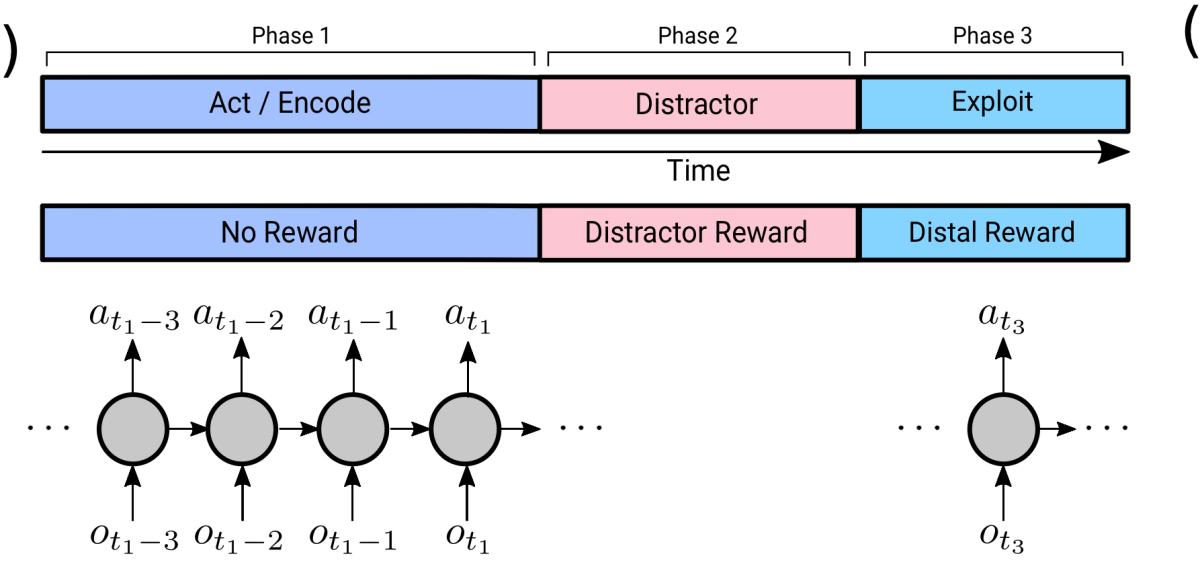
1. agents编码和存储感知和事件记忆

2. agents必须通过识别和访问过去事件的记忆来预测味蕾的奖励
3. agents必须基于对未来回报的贡献来重新评估这些过去的事件

首先，定义两类task：

- type1 (信息获取任务)，有3个阶段
    1. agent与环境交互获取信息，但没有即时奖励
    2. agent在一段长时间内参与了一个无相关的任务，但有获得大量的附带奖励
    3. agent需要利用第1阶段获取的信息，最终使得任务顺利完成并获得最终奖励
- type2（因果任务），也有3个阶段
    1. agent主动触发某些事件，而这些事件只会产生长期的因果后果
    2. 同样是一个分散注意力的任务
    3. agent必须利用它在第1阶段的活动引起的环境变化来获得成功。

(The three phase task structure. event. In In phase 1 (P1), there is no reward, but the agent must seek information or trigger an the agent phase 2 (P2), the agent performs a distractor task that delivers reward. In phase 3 (P3), can acquire a distal reward, depending on its behavior in P1. At each time step, the RL agent takes in observations ot and produces actions at, and passes memory state to the next time step.)



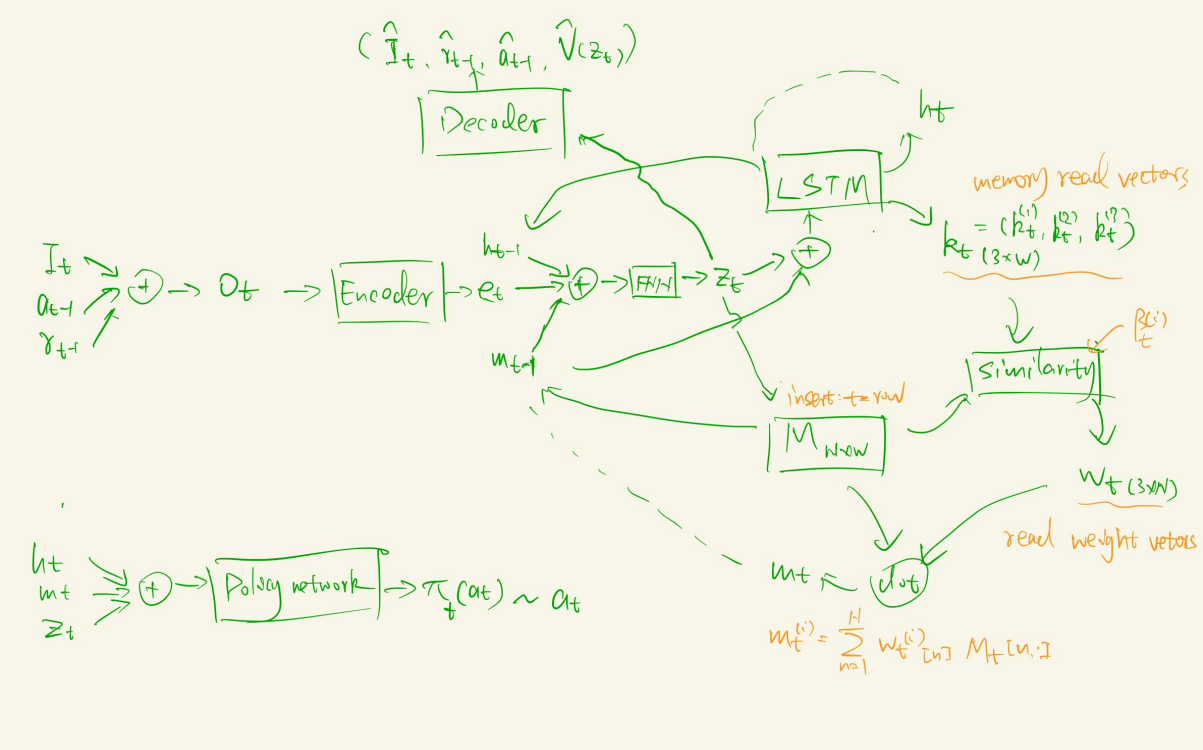
定义signal-to-noise ratio（SNR）：说明为何第2阶段（distractor phase）会破会长期 credi assignment

$$\mathrm{SNR} \approx \frac{\| \mathbb{E}_\pi[\Delta\theta] \|^2}{\mathrm{Var}_\pi\Big[\sum_{t\in P2} r_t\Big] \times C(\theta) + \mathrm{Var}_\pi[\Delta\theta|\text{no P2}]},$$

说明如下：

where $C(\theta)$ is a reward-independent term, and $\mathrm{Var}_\pi[\Delta\theta|\text{no P2}]$ is the (trace of the) policy gradient variance in an equivalent problem without a distractor interval. $\mathrm{Var}_\pi\big[\sum_{t\in P2} r_t\big]$ is the reward variance in P2. When P2 reward variance is large, the policy gradient SNR is inversely proportional to it. Reduced SNR is known to adversely affect the convergence of stochastic gradient

然后，使用一个AI智能体来解决这个任务，将其命名为RMA（Reconstructive Memory Agent）agent模型架构如下：

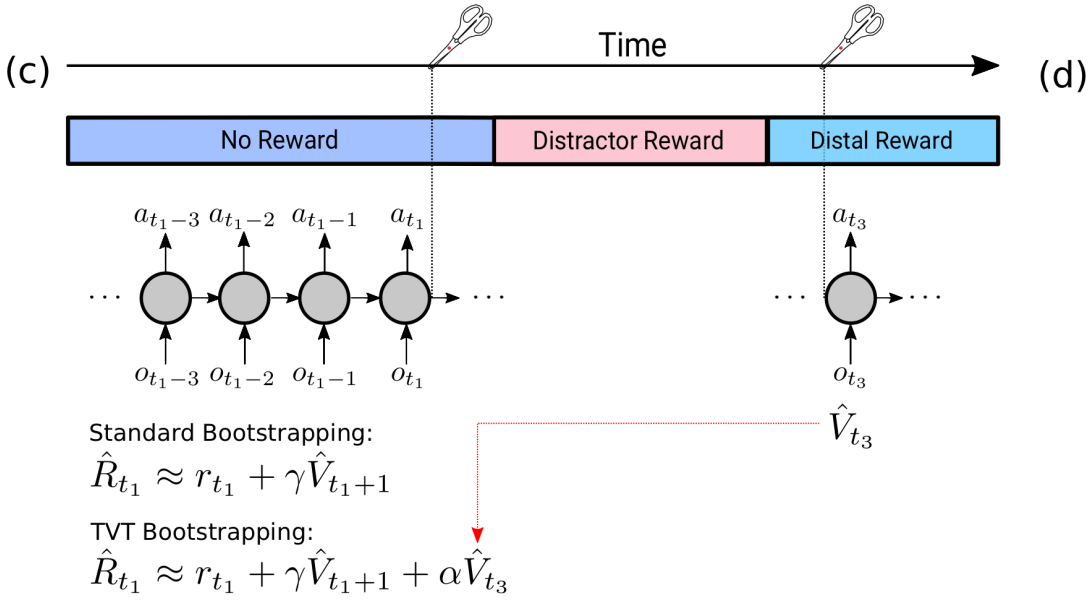


核心模块RNN，有LSTM和记忆矩阵M组成。

为了保证状态z_t包含有用的感知信息，需要用Decoder重构输入观测$\widehat{I_t}, \widehat{a}_{t-1}, \widehat{r}_{t-1}$，并且还要预测出价值函数$\widehat{V_t}$

RMA本身不具有支持LTCA（Long-term credit assignment）的专门功能，但是为TVT算法的操作提供了基础

Temporal Value Transport (TVT)算法，把第3阶段预测出的价值函数V_t“拼接”到第1阶段的回报计算上，如图：



算法流程：

step t_1.

Algorithm 1 Temporal Value Transport for One Read

input: $\{r_t\}_{t\in[1,T]}$, $\{\hat{V}_t\}_{t\in[1,T]}$, read strengths $\{\beta_t\}_{t\in[1,T]}$, read weights $\{w_t\}_{t\in[1,T]}$
splices : = []
for each crossing of read strength β_t above $\beta_{\text{threshold}}$ **do**
  $t_{\max} := \arg\max_t\{\beta_t | t \in$ crossing window$\}$
  Append $t_{\max}$ to splices
end for
for t in 1 to T **do**
  for t' in splices **do**
    if $t < t' - 1/(1-\gamma)$ **then**
      $r_t := r_t + \alpha w_{t'}[t]\hat{V}_{t'+1}$
      {The read based on $w_{t'}$ influences value prediction at next step, hence $\hat{V}_{t'+1}$.}
    end if
  end for
end for
return $\{r_t\}_{t\in[1,T]}$

（hoho_todo: β是什么？）

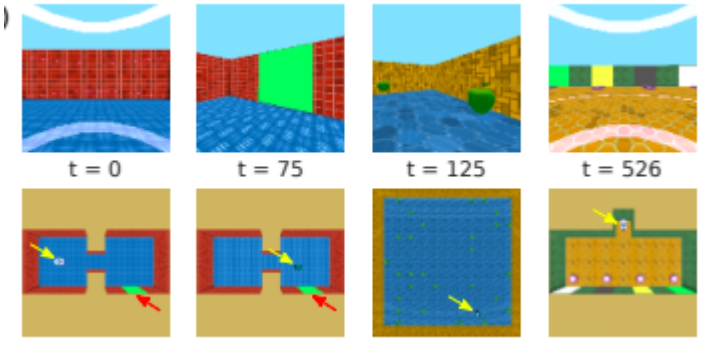
论文中的实验是如何设计的？

设计一个type1类任务：Active Visual Match

在P1阶段，agent必须主动地在一个两个房间的迷宫中随机找到一个彩色正方形。如果一个agent在P1中偶然发现了视觉线索，那么它可以在P3中使用这个信息，但这只能是随机成功的。

在P2阶段，agent执行一个30秒的收集苹果干扰任务。

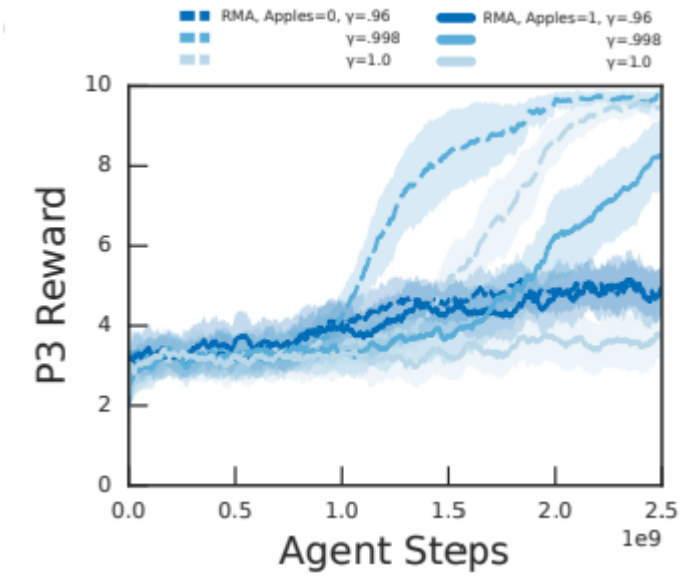
在P3阶段中，agent需要找到P1阶段中的匹配项。

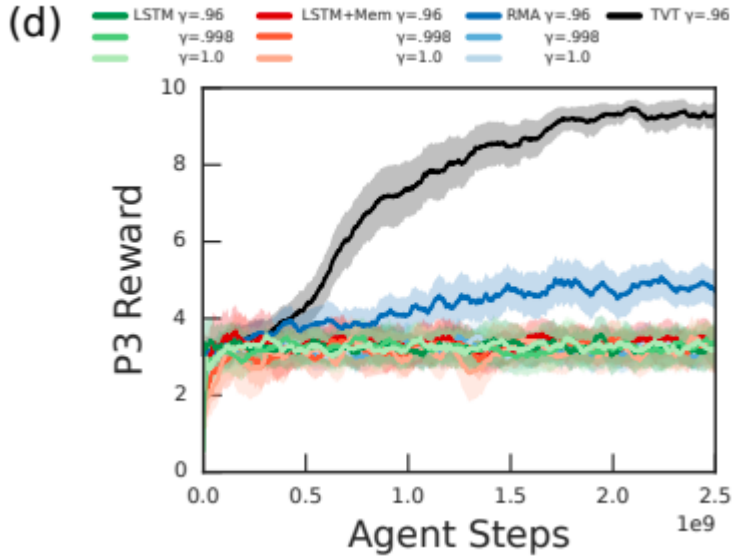


用于定量评估的数据集是什么？代码有没有开源？

- 无数据集
- 代码：https://github.com/deepmind/deepmind-research/tree/master/tvt

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？





这篇论文到底有什么贡献？

用大量的证据支持了神经记忆系统和奖励系统具有高度互相依赖的特点

下一步呢？有什么工作可以继续深入？

hoho_todo