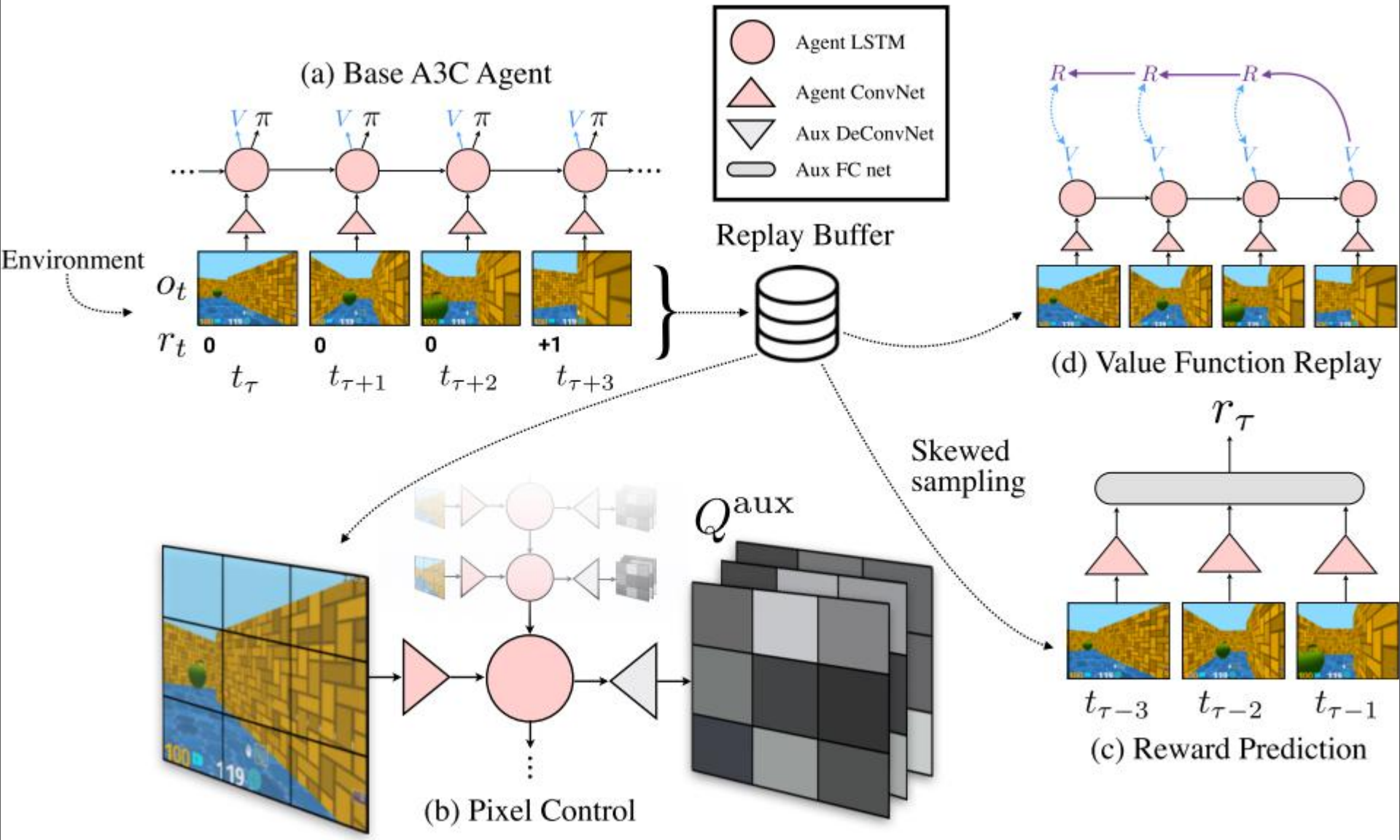
UNREAL

presented by Jason TOKO

算法介绍

- 算法全称是：UNsupervised REinforcement and Auxiliary Learning(UNREAL)
- 核心思想：UNREAL算法在A3C算法的基础上进行改善，通过在训练A3C的同时，训练多个**辅助任务（AUXILIARY TASKS**）来改进算法性能
- 算法背景：Asynchronous Advantage Actor-Critic(A3C) 和 Long Short-Term Memory (LSTM)


Agent LSTM
Agent ConvNet
Aux DeConvNet
Aux FC net
(a) Base A3C Agent
Environment
o_t
r_t
0
0
0
+1
t_τ
$t_{\tau+1}$
$t_{\tau+2}$
$t_{\tau+3}$
Replay Buffer
(d) Value Function Replay
Skewed sampling
r_τ
$t_{\tau-3}$
$t_{\tau-2}$
$t_{\tau-1}$
(c) Reward Prediction
Q^{aux}
(b) Pixel Control

辅助任务

- 1、辅助控制任务
  - 优化目标：

$$\arg\max_{\theta} \mathbb{E}_{\pi}[R_{1:\infty}] + \lambda_c \sum_{c \in \mathcal{C}} \mathbb{E}_{\pi_c}[R^{(c)}_{1:\infty}],$$

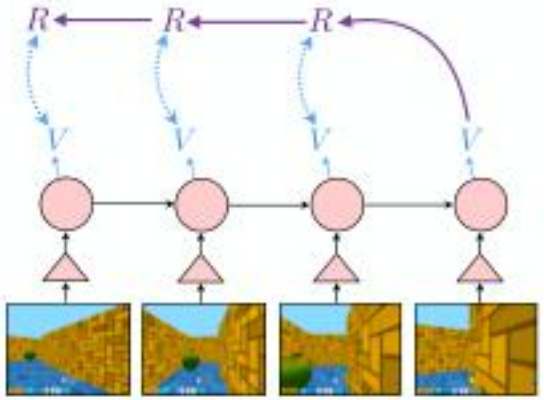
    - 其中 $R^{(c)}_{t:t+n} = \sum_{k=1}^{n} \gamma^k r^{(c)}_t$
  - n-step Q learning的损失函数:

$$\mathcal{L}^{(c)}_Q = \mathbb{E}\left[\left(R_{t:t+n} + \gamma^n \max_{a'} Q^{(c)}(s', a', \theta^-) - Q^{(c)}(s, a, \theta)\right)^2\right]$$

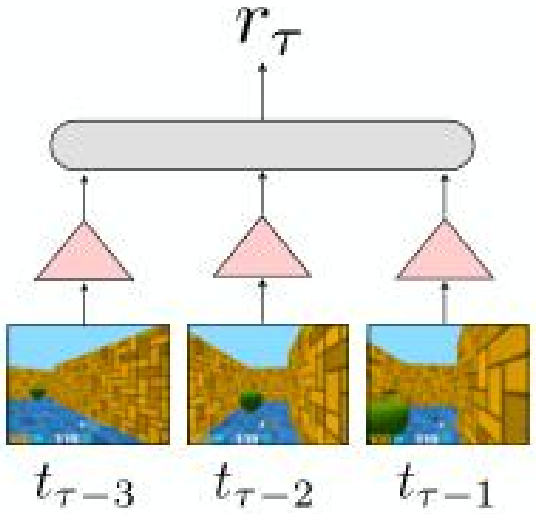
  - 两种辅助控制任务
    - 像素控制（pixel control）
    - 特征控制（feature control）

辅助任务

- 2、奖励预测（Reward Prediction）
  - 目的：消除奖励的稀疏性辅助学习，同时不引入偏差
  - 做法：通过序列 $S_\tau = (s_{\tau-k}, s_{\tau-k+1}, \ldots, s_{\tau-1})$
    预测奖励 r_τ
- 3、经验回放（Experience Replay）



(d) Value Function Replay



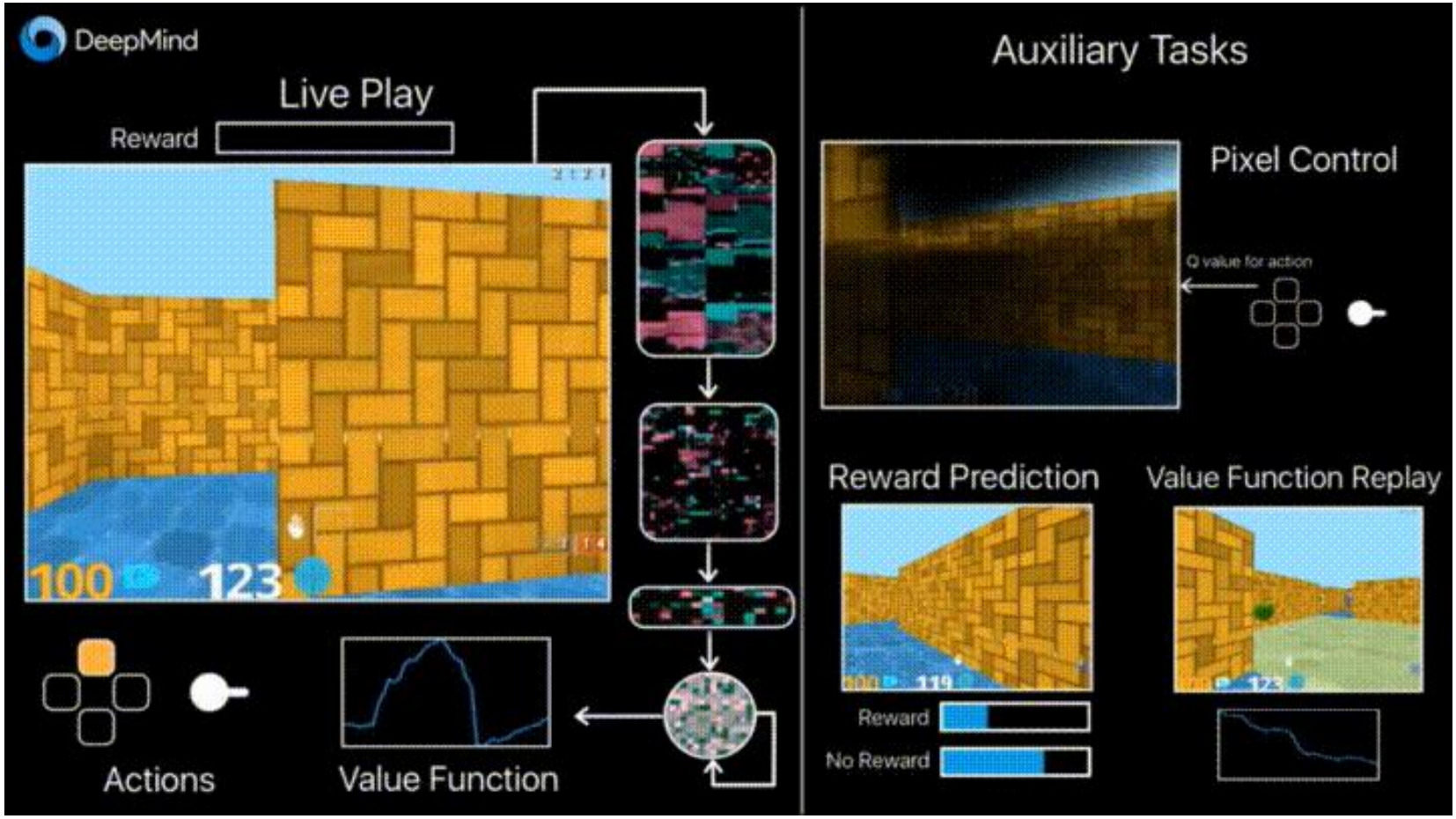
(c) Reward Prediction

- UNREAL 的损失函数：

$$\mathcal{L}_{UNREAL}(\theta) = \mathcal{L}_{\mathrm{A3C}} + \lambda_{\mathrm{VR}}\mathcal{L}_{\mathrm{VR}} + \lambda_{\mathrm{PC}} \sum_{c} \mathcal{L}_{Q}^{(c)} + \lambda_{\mathrm{RP}}\mathcal{L}_{\mathrm{RP}}$$

- 算法效果：

演示


DeepMind
Live Play
Reward
100
123
Actions
Value Function
Auxiliary Tasks
Pixel Control
Q value for action
Reward Prediction
Reward
No Reward
Value Function Replay