Some New Ideas in RL

传统算法回顾

Two posts:

- https://lilianweng.github.io/lil-log/2018/02/19/a-long-peek-into-reinforcement-learning.html
- https://lilianweng.github.io/lil-log/2018/04/08/policy-gradient-algorithms.html

传统算法回顾

- What is Reinforcement Learning?
  - Key Concepts
    - Model: Transition and Reward
    - Policy
    - Value Function
    - Optimal Value and Policy
  - Markov Decision Processes
  - Bellman Equations
    - Bellman Expectation Equations
    - Bellman Optimality Equations
- Common Approaches
  - Dynamic Programming
    - Policy Evaluation
    - Policy Improvement
    - Policy Iteration
  - Monte-Carlo Methods
  - Temporal-Difference Learning
    - Bootstrapping
    - Value Estimation
    - SARSA: On-Policy TD control
    - Q-Learning: Off-policy TD control
    - Deep Q-Network
  - Combining TD and MC Learning
  - Policy Gradient
    - Policy Gradient Theorem
    - REINFORCE
    - Actor-Critic
    - A3C
  - Evolution Strategies
- Known Problems
  - Exploration-Exploitation Dilemma
  - Deadly Triad Issue
- Case Study: AlphaGo Zero
- References

Two pc

- https o/lil-log/2018/02/19/a-long-peek-into-reinfc ml
- https o/lil-log/2018/04/08/policy-gradient-algori

传统算法回顾

Two po

- https
  reinfo
  /lil-log/201
  ml
- https
  algori
  /lil-log/201

挑战

- Sparse supervision
- Severe partial observability
- Sample efficiency
- ……

New Ideas

- 分层强化学习
- 记忆和注意力
- 世界模型和想象

分层强化学习：HRL

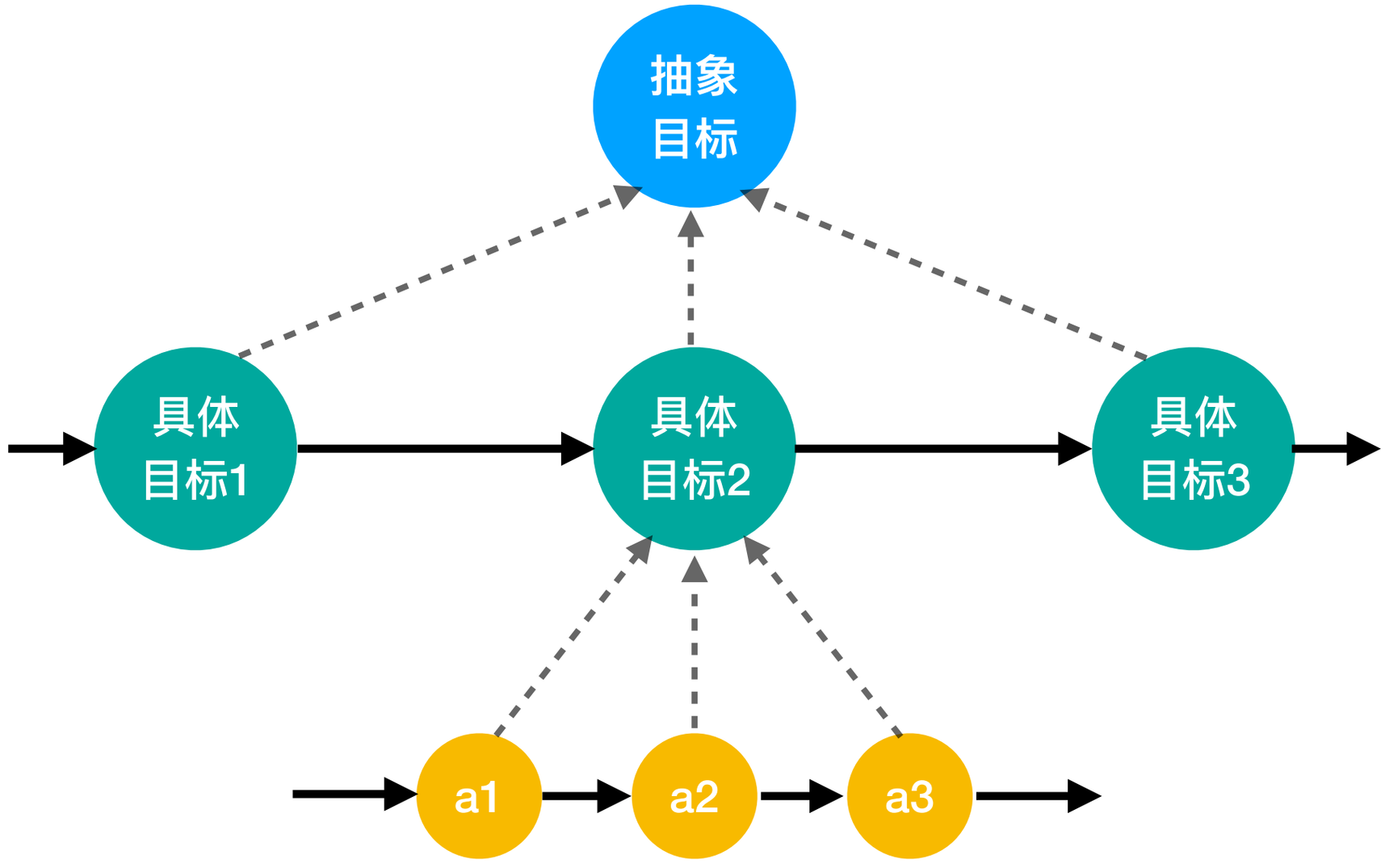
针对复杂任务、长程反馈——多层策略：

- 高层：分解高层目标为抽象的低层目标
- 底层：针对低层目标输出环境动作

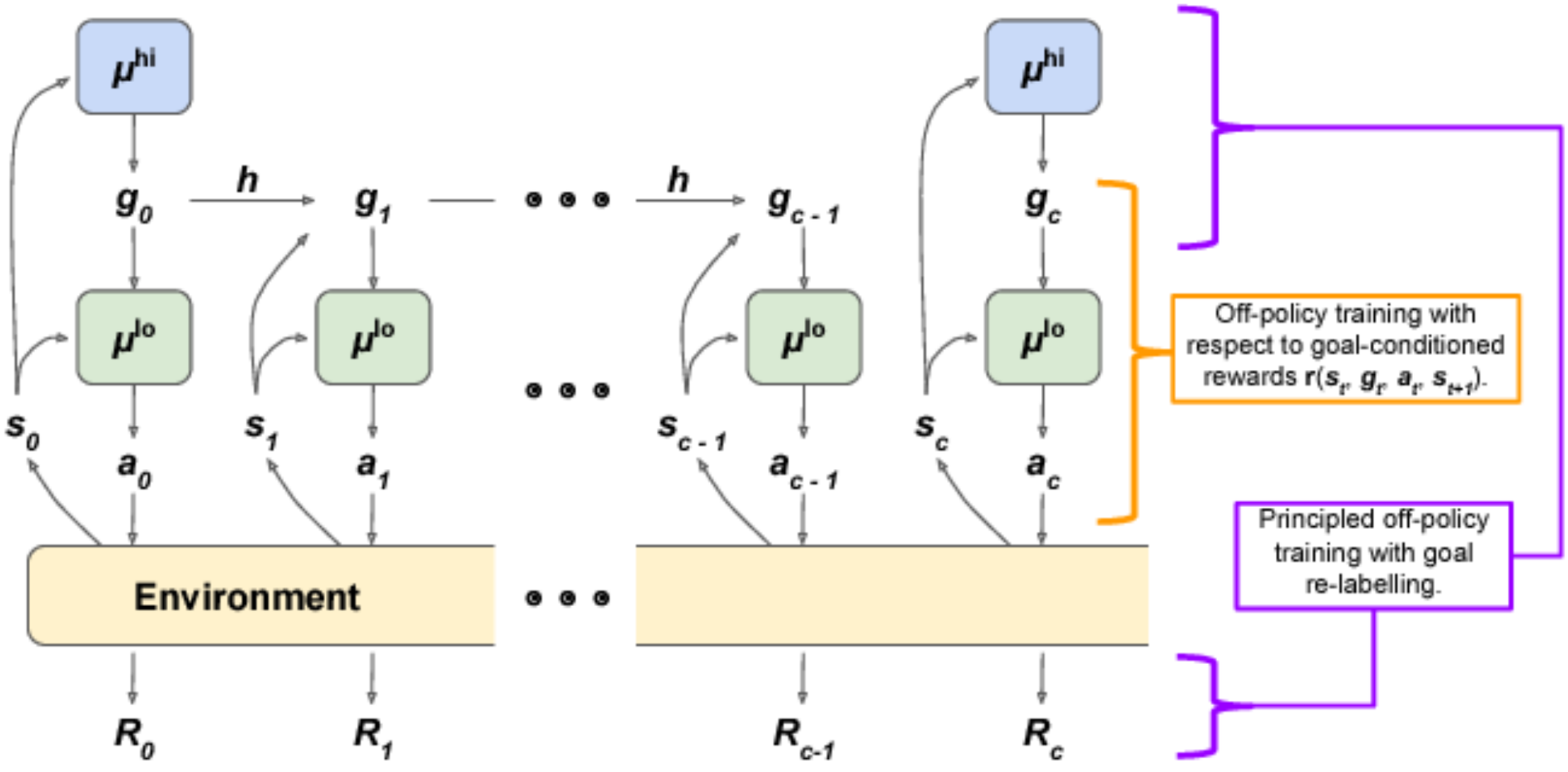
HRL

高层：决策者

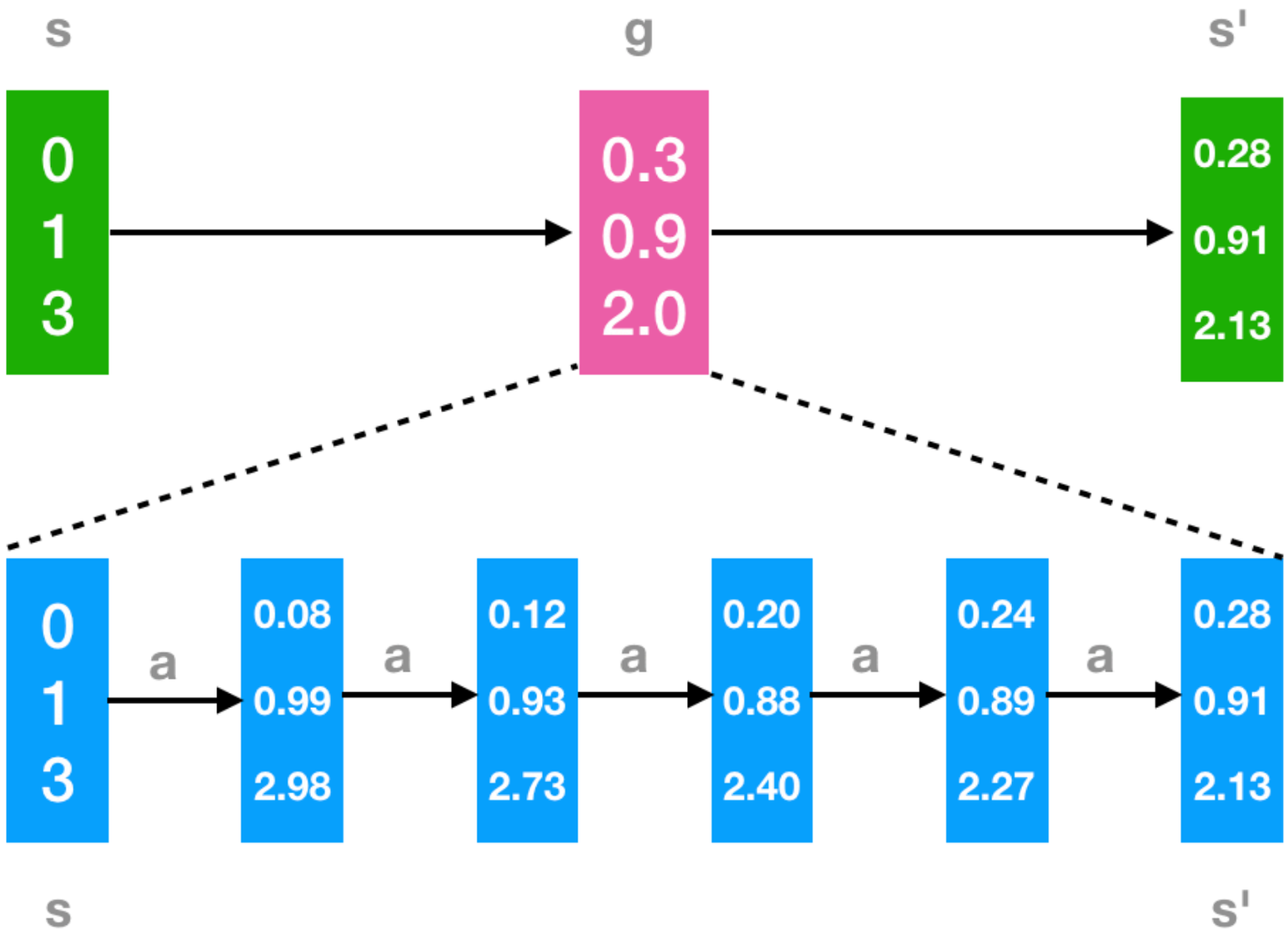
底层：执行者



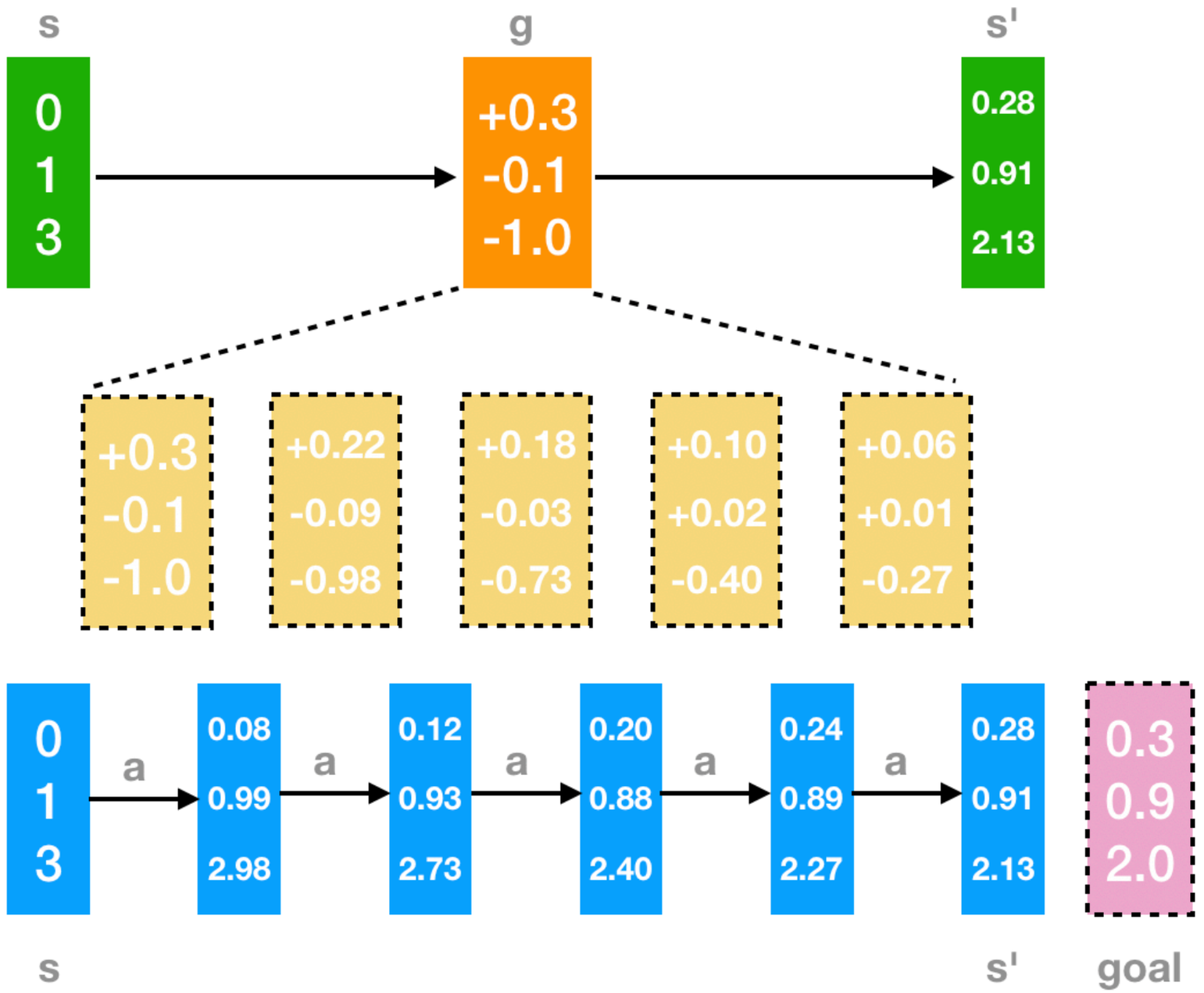
实现：HIRO


μ^{hi}
g_0
h
g_1
h
g_{c-1}
g_c
μ^{lo}
μ^{lo}
μ^{lo}
μ^{lo}
s_0
s_1
s_{c-1}
s_c
a_0
a_1
a_{c-1}
a_c
Environment
R_0
R_1
R_{c-1}
R_c
Off-policy training with respect to goal-conditioned rewards $\mathbf{r}(s_t, g_t, a_t, s_{t+1})$.
Principled off-policy training with goal re-labelling.

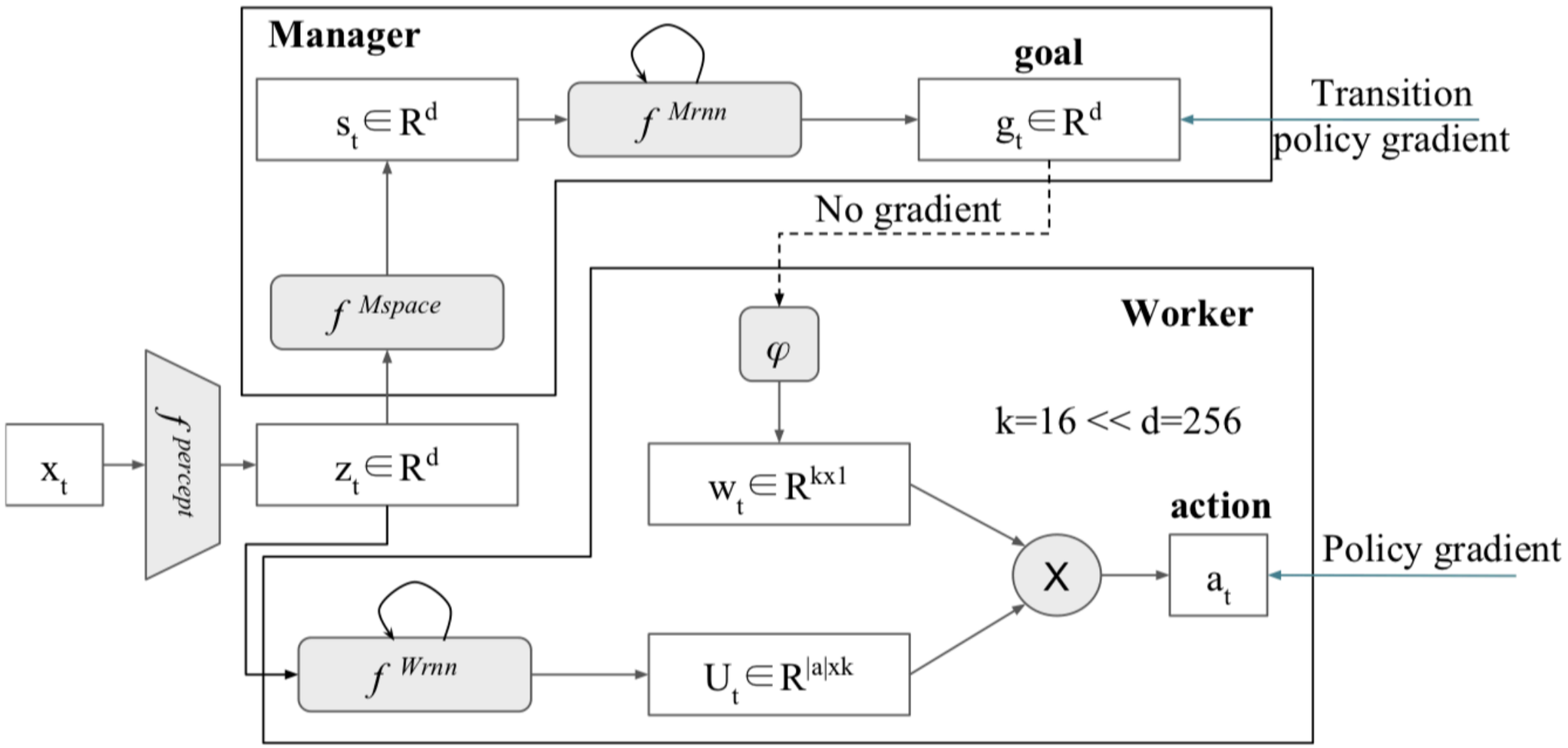
HIRO


s
g
s'
0
1
3
0.3
0.9
2.0
0.28
0.91
2.13
0
1
3
a
0.08
0.99
2.98
a
0.12
0.93
2.73
a
0.20
0.88
2.40
a
0.24
0.89
2.27
a
0.28
0.91
2.13
s
s'

HIRO


s
g
s'
0
1
3
+0.3
-0.1
-1.0
0.28
0.91
2.13
+0.3
-0.1
-1.0
+0.22
-0.09
-0.98
+0.18
-0.03
-0.73
+0.10
+0.02
-0.40
+0.06
+0.01
-0.27
0
1
3
a
0.08
0.99
2.98
a
0.12
0.93
2.73
a
0.20
0.88
2.40
a
0.24
0.89
2.27
a
0.28
0.91
2.13
0.3
0.9
2.0
s
s'
goal

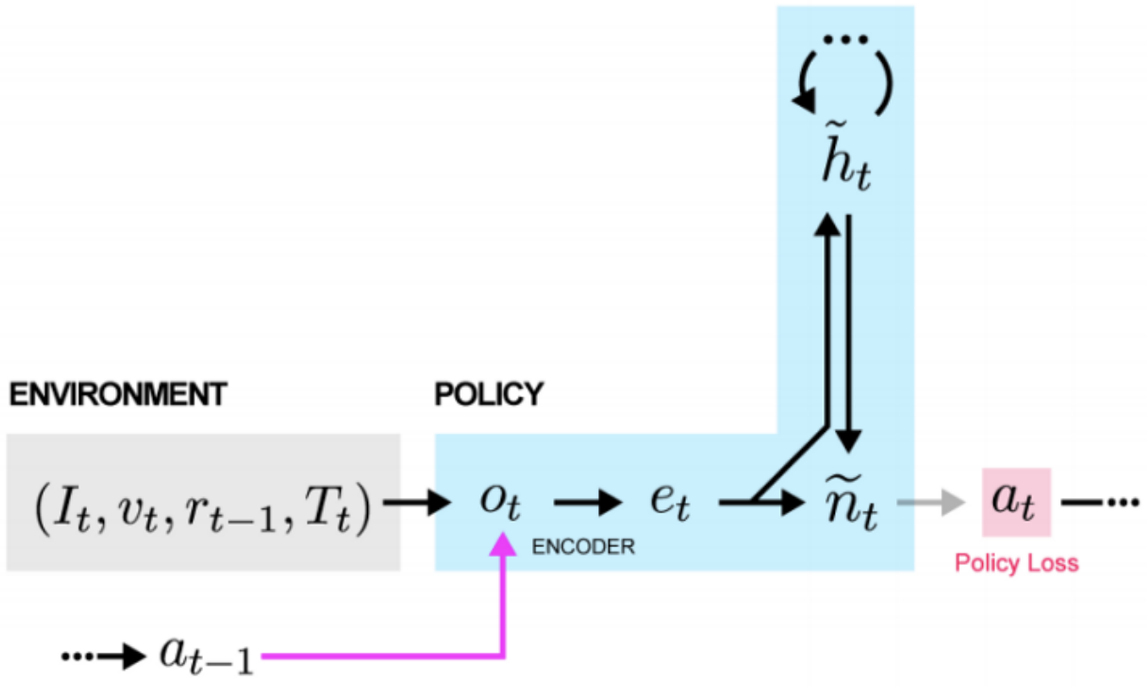
FeUdal Networks



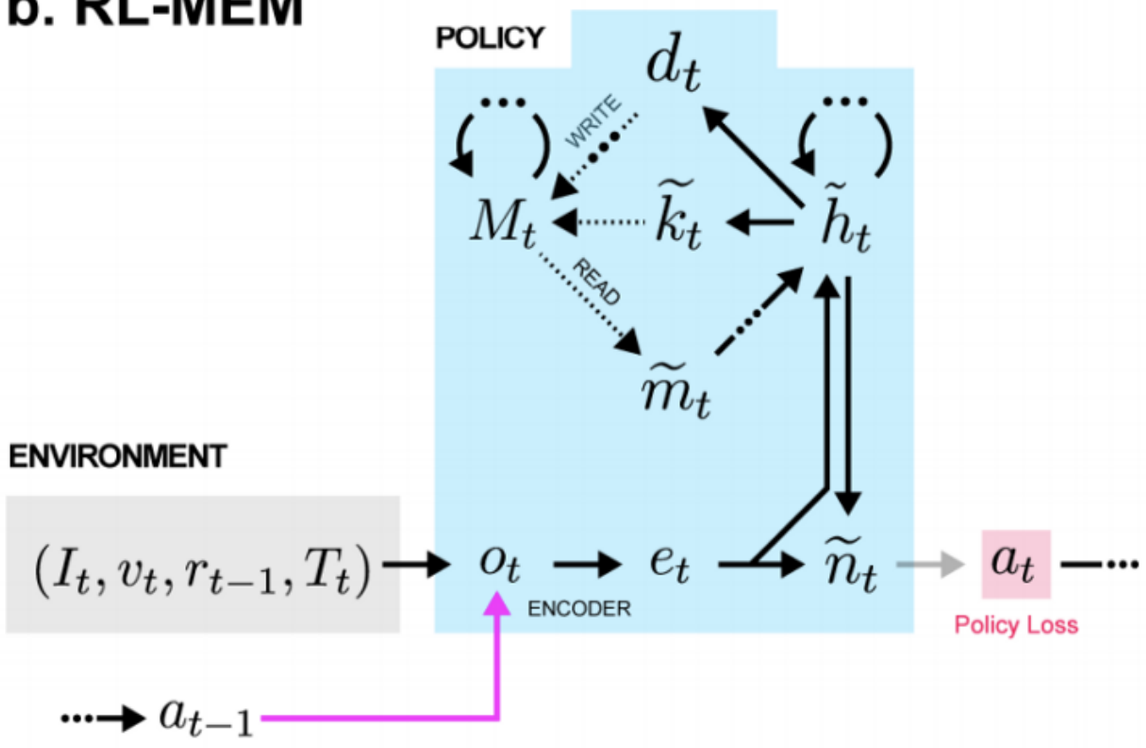
记忆和注意力

- 部分可见问题
- 将观察融入记忆
- 结合**观察**和**记忆**采取决策
- **相关**记忆：注意力机制

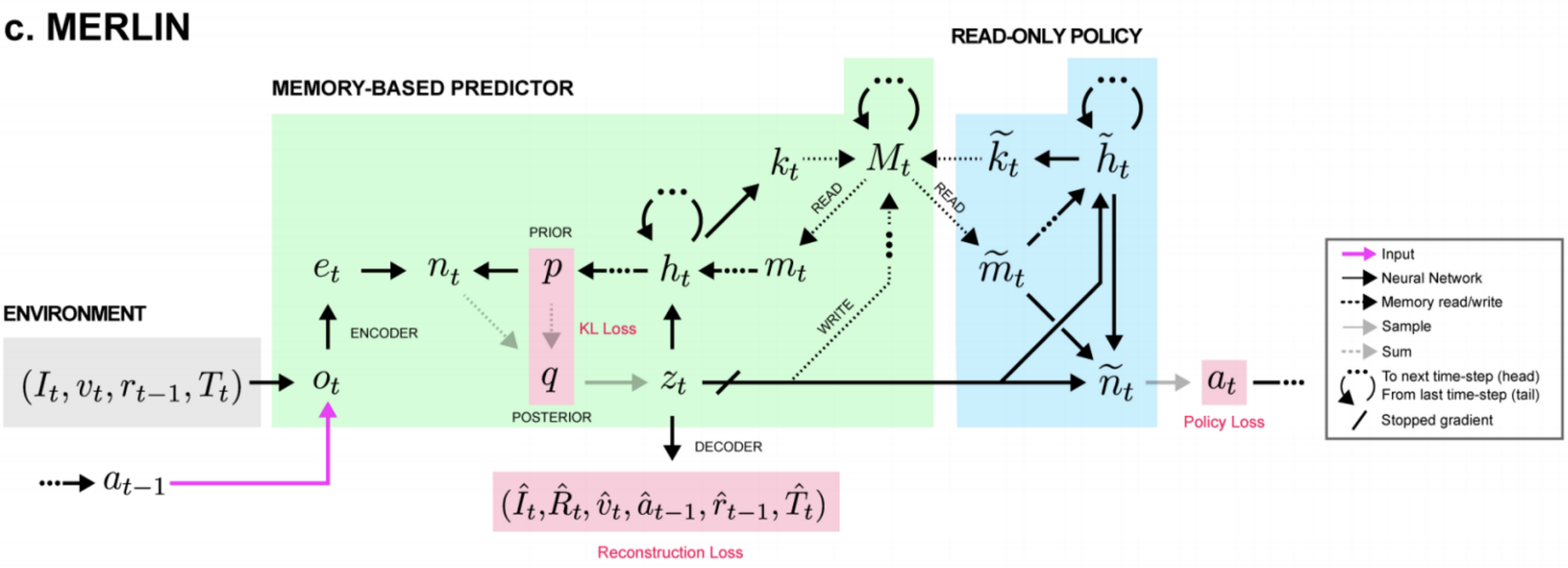
a. RL-LSTM



b. RL-MEM



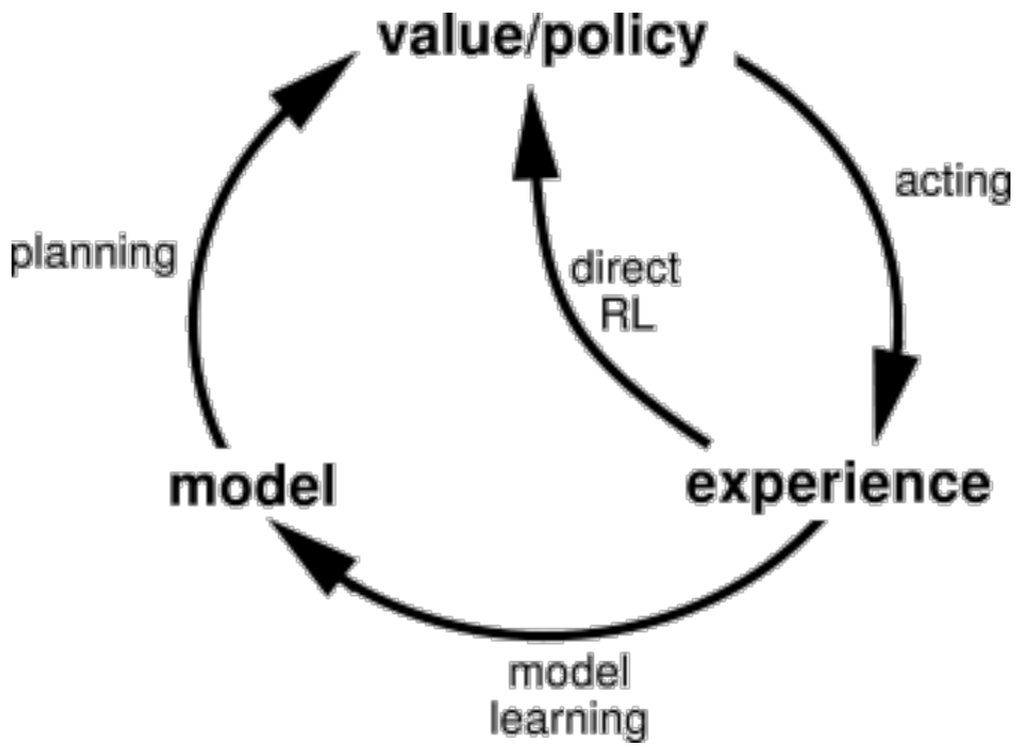
c. MERLIN



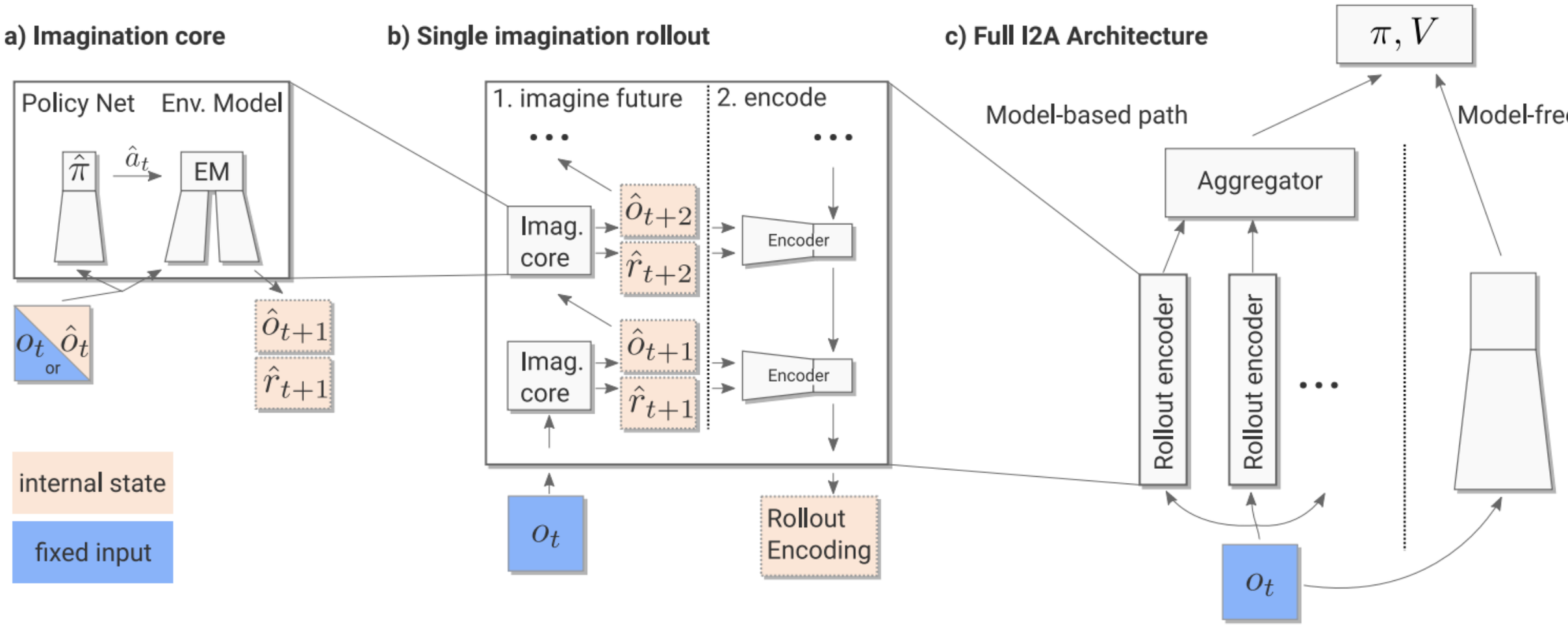
世界模型和想象

- 世界模型：对环境的理解（解释？）
- 想象：世界模型上的规划，improve sample efficiency
- 辅助决策

Framework


value/policy
acting
planning
direct
RL
model
experience
model
learning

I2A


a) Imagination core
Policy Net
Env. Model
$\hat{\pi}$
$\hat{a}_t$
EM
o_t $\hat{o}_t$ or
$\hat{o}_{t+1}$
$\hat{r}_{t+1}$
b) Single imagination rollout
1. imagine future
2. encode
Imag. core
$\hat{o}_{t+2}$
$\hat{r}_{t+2}$
Encoder
Imag. core
$\hat{o}_{t+1}$
$\hat{r}_{t+1}$
Encoder
o_t
Rollout Encoding
c) Full I2A Architecture
π, V
Model-based path
Model-free path
Aggregator
Rollout encoder
Rollout encoder
o_t
internal state
fixed input

Reference

- https://towardsdatascience.com/advanced-reinforcement-learning-6d769f529eb3
- http://karpathy.github.io/2016/05/31/rl/
- https://arxiv.org/abs/1703.01161
- https://arxiv.org/pdf/1805.08296.pdf
- https://arxiv.org/pdf/1803.10760.pdf
- https://arxiv.org/abs/1707.06203
- https://deepmind.com/blog/agents-imagine-and-plan/