

# 从框架到算法 强化学习玩合成大西瓜



强化学习中的策略可以按目标策略和行为策略进行分类：

- 目标策略 (target policy)：智能体要学习的策略
- 行为策略 (behavior policy)：智能体与环境交互的策略，即用于生成行为的策略

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[ R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

## Q-learning: An off-policy TD control algorithm

Initialize  $Q(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$ , arbitrarily, and  $Q(\text{terminal-state}, \cdot) = 0$

Repeat (for each episode):

    Initialize  $S$

    Repeat (for each step of episode):

        Choose  $A$  from  $S$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

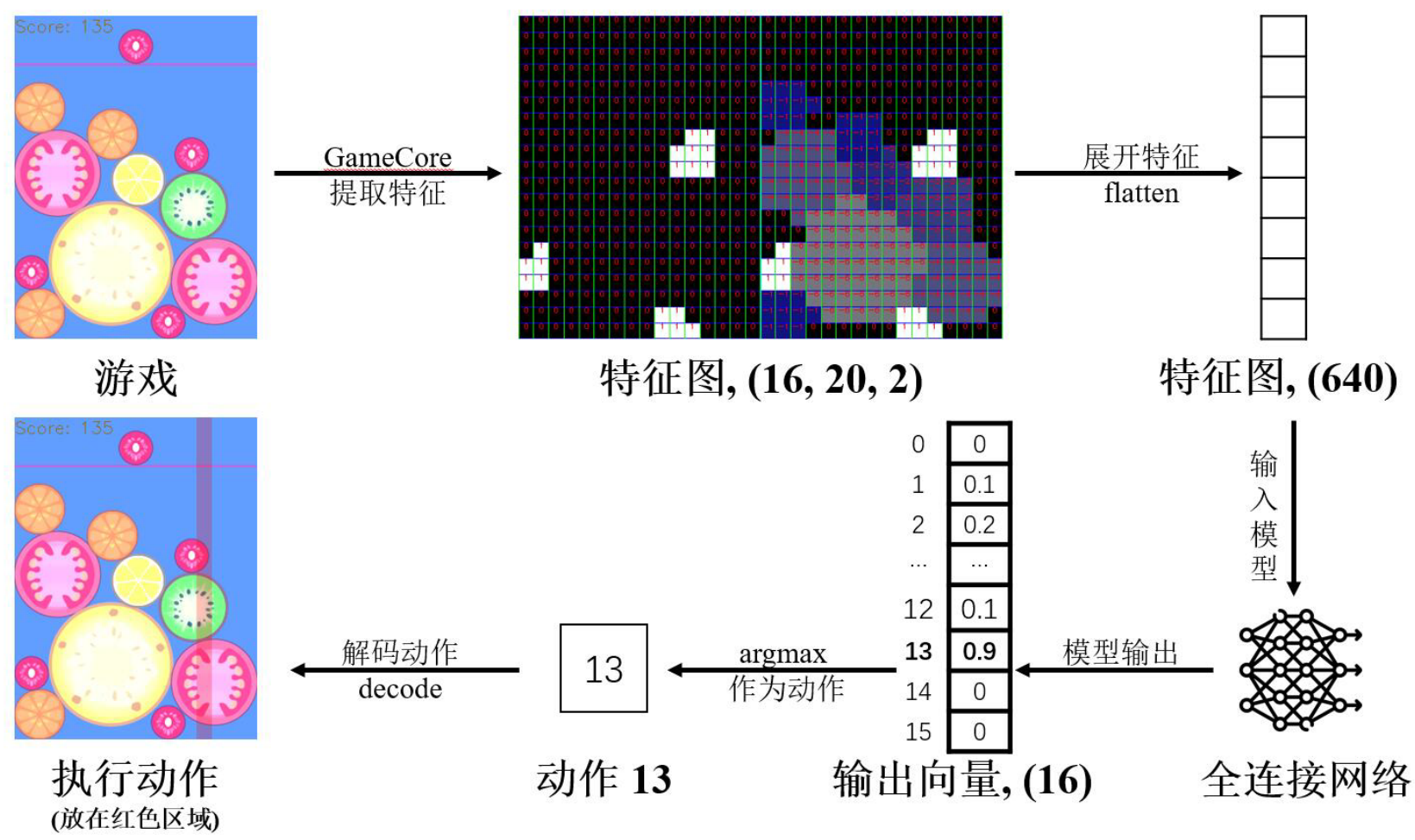
        Take action  $A$ , observe  $R, S'$

$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$

$S \leftarrow S'$

    until  $S$  is terminal

# 数据流



如果使用超验知识来指导强化学习，效果会非常好

