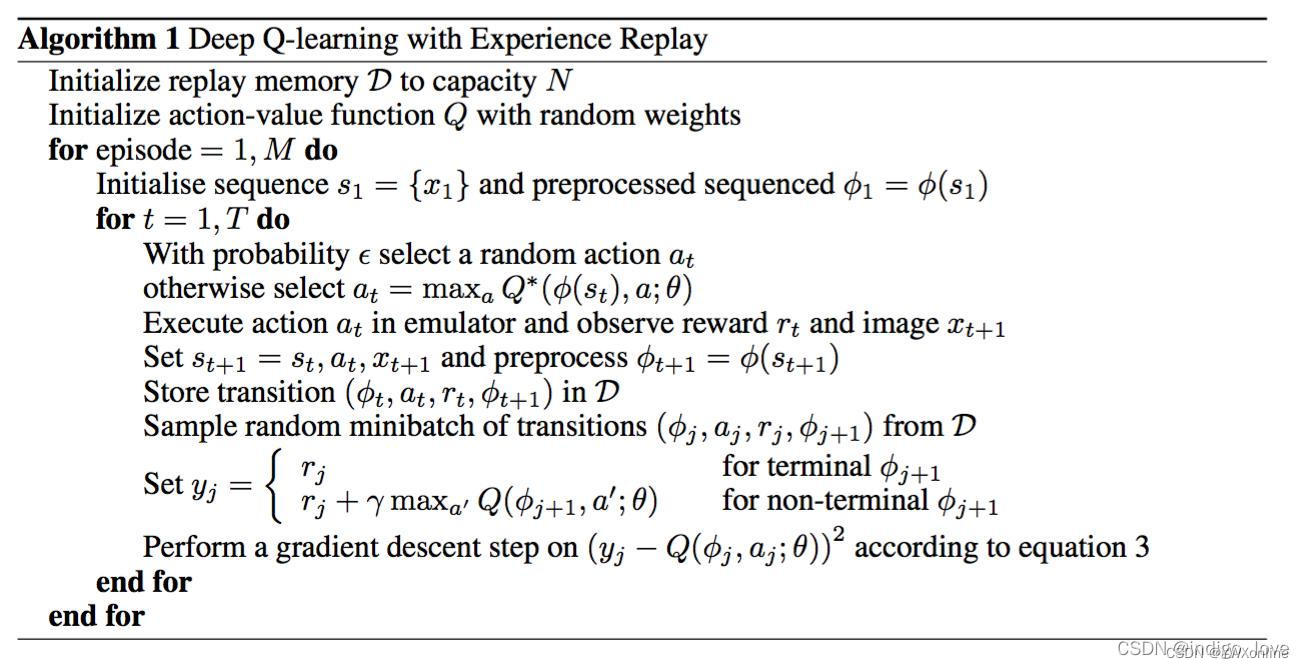
第三次编程作业报告

——李昭阳 2021013445

1. 问题建模
2. 状态空间：
   1. 31个普通状态二维坐标
   2. 4个陷阱状态二维坐标
   3. 1个奖励状态二维坐标
3. 行动集合：{上、下、左、右}
4. 状态转移概率：

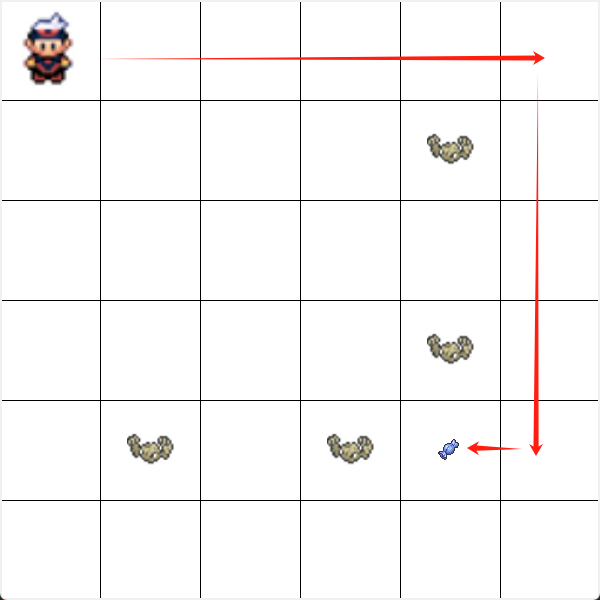
由于且，因此认为，即s状态下选取a动作的概率。

1. 回报设计：
   1. 行动鼓励：每次转移到不同的位置奖励1分
   2. 距离惩罚：惩罚动作后位置与奖励状态的曼哈顿距离
   3. 结束奖励：若掉入陷阱惩罚-100分，得到糖果奖励100分
2. DQN网络设计
3. 算法原理：

算法维护了一个策略网络，一个目标网络。目标网络通过神经网络模拟Q表，策略网络通过与目标网络贝尔曼方程所得结果进行均方误差Loss求解，不断更新自身参数，在一定步长后整体更新目标网络。

1. 算法效果：

算法采用-贪心策略，选取如下超参数，得到较好训练效果。

 训练后智能体采取如下策略进行移动，易知其选取了其中一条最优路径。