- DQN
  - naive DQN
    - 用NN估计Q(s,a)值并依据epsilon-greedy决策
    - Q网络有两个，一个目标网络，一个当前网络。为什么这样做：保证一个方向更新，保证稳定性。
    - td_error=r+gamma*max Q'(s',a')-Q(s,a)如果不用Q'，则在两个方向minimize td_error: 减小gamma*max Q(s,a)和增大Q(s,a), 矛盾
  - Double DQN
    - 由于网络不稳定性Q值估计过高，导致正反馈，最终Q值过高
    - 采用另一个网络估计r+gamma*maxQ'(s,a)的动作，相当于增加样本量，减小一个网络随机性带来的误差

作者：@ry664663 | 来自：知犀思维导图