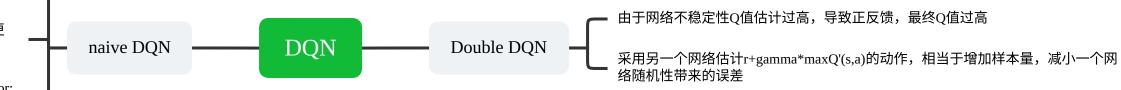
用NN估计Q(s,a)值并依据epsilon-greedy决策 Q网络有两个,一个目标网络,一个当前网络。为什么这样做:保证一个方向更新,保证稳定性。

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



作者: @ry664663 | 来自: 知犀思维导图