DQN模型关键技术：

* 样本池（Experience Reply）

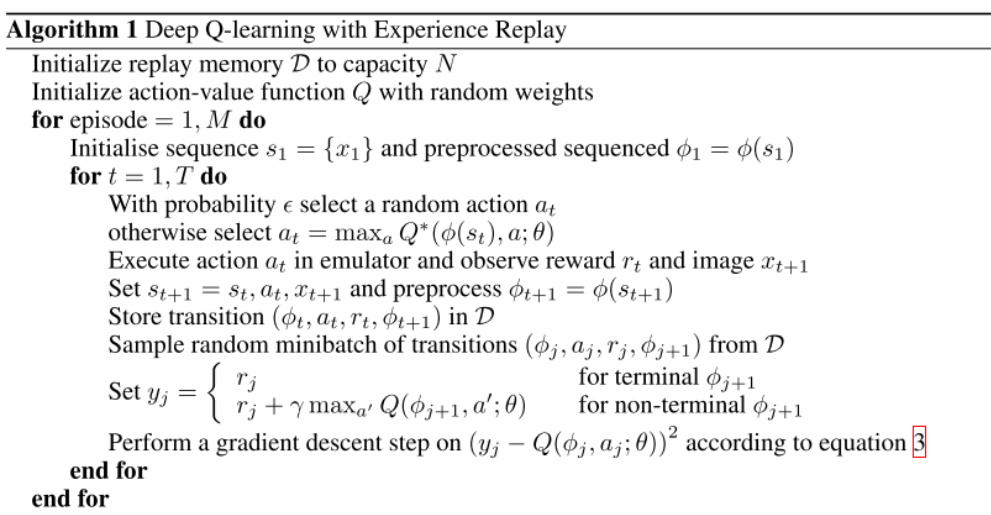
DQN 模型的目标是将强化学习和深层神经网络结合起来，只需要输入原始 RGB 图像并通过 SGD 对样本进行训练便可以输出最优 Q 网络。

DQN 模型中最重要的一个技巧是采用了经验回放机制。在每个时间步长，都会将经验et=(st,at,rt,st+1) 存储在数据集 D={e1,e2,…,eN} 中（其实终止标志 done 也要存进去，表示当前状态是否还有后继状态），并且会保留许多 episode 的经验，即新的 episode 开启时，经验回放池 D 不会重置。打破观察序列的相关性，满足深度网络使用要求；平滑数据分布中的变化，避免训练的振荡，提高采样数据利用效率

DQN 模型中采用 off-policy，与Q-learning是一样的，目标策略采用贪婪策略，行动策略采用ϵ-贪婪策略。

因为从经验回放池中抽取的历史经验具有不一样的长度（因为状态实际上是一个动作-观测值序列），因此定义函数ϕ来将长度不一致的历史经验输出为固定长度的历史经验表示。等式（3）就是上面的均方误差损失函数。

算法伪代码如下：



* 双网络结构

DQN除了使用深度卷积网络近似表示当前的值函数之外，还单独使用了另一个网络来产生目标Q值.具体地，Q(s,a)表示当前值网络的输出，用来评估当前状态动作对的值函数：Q(s,a)表示目标值网络的输出,一般采用, Q(s',a')近似表示值函数的优化目标，即目标Q值。当前值网络的参数θ是实时更新的，每经过N轮迭代，将当前值网络的参数复制给目标值网络.通过最小化当前Q值和目标Q值之间的均方误差来更新网络参数.误差函数为：

FBI$VCF1G90_$)X0P}_1NAK

对参数θ求偏导，得到以下梯度：

B4V@A0Q4A_A%@5S%2`(90IV

引入目标值网络后，在一段时间内目标Q值是保持不变的，一定程度上降低了当前Q值和目标Q值之间的相关性，提升了算法的稳定性。  
  DQN将奖赏值和误差项缩小到有限的区间内，保证了Q值和梯度值都处于合理的范围内，提高了算法的稳定性。实验表明,DQN在解决诸Atari 2600游戏等类真实环境的复杂问题时，表现出与人类玩家相媲美的竞技水平,甚至在一些难度较低的非战略性游戏中, DQN的表现超过了有经验的人类玩家。在解决各类基于视觉感知的DRL任务时，DQN使用了同一套网络模型、参数设置和训练算法，这充分说明DQN方法具有很强的适应性和通用性。