# Deep Q Network(DQN)学习计划

**一、学习规划(从入门到强化)**

1. **机器学习、深度学习、强化学习、深度强化学习**
2. **机器学习**

**定义：多领域交叉**学科，专门研究**计算机**怎样**模拟或实现人类的学习行为**，以**获取新的知识或技能**，**重新组织**已有的知识结构使之不断**改善**自身的性能。

**分类：**

A,基于学习策略(模拟人脑、直接采用数学方法)

B,基于学习方法(归纳、演绎、类比、分析)

C,基于学习方式(监督、无监督、**强化学习**)

D,基于数据形式(结构化、非结构化)

E,基于学习目标(概念、规则、函数、类别、贝叶斯网络)

**常见算法：**

A,决策树算法

B,朴素贝叶斯算法

C,支持向量机算法

D,随机森林算法

**E,人工神经网络算法**

F,关联规则算法

G,EM(期望最大化)算法

**H,深度学习(与人工神经网络算法的区别)**

1. **深度学习**

**定义：**深度学习是机器学习的一种。深度学习的概念**源于人工神经网络**的研究，含**多个隐藏层的多层感知器**就是一种深度学习结构。深度学习通过组合**低层特征**形成更加抽象的**高层表示**属性类别或特征，以发现数据的分布式特征表示。研究深度学习的动机在于建立**模拟人脑**进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本等。

**典型模型：**

**A,卷积神经网络模型**

B,深度信任网络模型

C,堆栈自编码网络模型

**训练过程：**

A,自下向上的非监督学习

B,自顶向下的监督学习

1. **强化学习**

**定义：**是机器学习的**范式和方法论**之一，用于描述和解决**智能体**在与**环境**的**交互**过程中，通过**学习**策略以达成**回报最大化**或**实现特定目标**的问题。

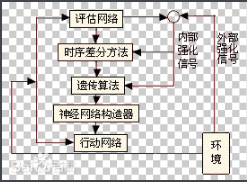
**智能体：agent，**进行机器学习的代理，可以感知外界环境的状态进行决策、对环境做出动作并通过环境的反馈调整决策。

**环境：**智能体**外部**所有事物的**集合**，其状态会受智能体动作的影响而改变，且上述改变可以完全或部分地被智能体感知。环境在每次决策后可能会反馈给智能体相应的奖励。

**基本原理：**如果Agent的某个行为策略导致环境正的奖赏(强化信号)，那么Agent以后产生这个行为策略的趋势便会加强。Agent的目标是在每个离散状态发现最优策略以使期望的折扣奖赏和最大。

强化学习系统**学习的目标**是**动态地调整参数**，以达到强化信号最大。

**网络模型设计：**



**设计考虑：**

A,如何**表示**状态空间和动作空间;

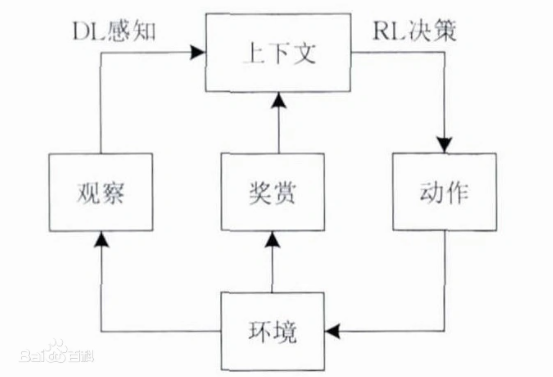
B,如何**选择**建立信号以及如何通过学习来**修正**不同状态－动作对的**值**;

C,如何**根据**这些值来选择适合的动作;

1. **深度强化学习**

**定义：**将**深度学习的感知能力**和**强化学习的决策能力**相结合，可以**直接根据**输入的图像进行控制，是一种**更接近人类思维方式**的人工智能方法。

**原理框架：**



**典型算法：**

**深度Q网络方法(DQN)：**融合了**神经网络**和**Q learning**的方法。

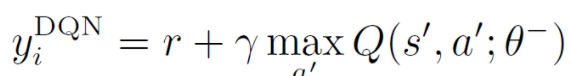
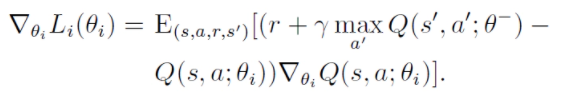
DQN有一个记忆库用于**学习之前的经历**。Q learning 是一种 **off-policy 离线学习法**，它能学习**当前**经历着的，也能学习**过去**经历过的，甚至是**学习别人**的经历。所以每次DQN**更新**的时候，我们都可以**随机抽取**一些**之前**的经历进行学习。随机抽取这种做法**打乱**了经历之间的**相关性**，也使得神经网络**更新更有效率**。**Fixed Q-targets** 也是一种**打乱相关性**的机理，如果使用 fixed Q-targets，我们就会在DQN中使用到**两个结构相同但参数不同的神经网络**，预测**Q估计**的神经网络具备**最新**的参数，而预测**Q现实**的神经网络使用的参数则是很久**以前**的。有了这**两种提升手段**，DQN才能在一些游戏中超越人类。

**分类：**

**A,基于卷积神经网络的DRL**

卷积神经网络对图像处理拥有天然的优势。

**深度Q网络：**是深度强化学习领域的开创性工作。它采用时间上相邻的4帧游戏画面作为原始图像输入，经过深度卷积神经网络和全连接神经网络，输出状态动作Q函数，实现了端到端的学习控制。

深度Q网络使用带有参数θ的Q函数Q(s， a; θ)去逼近值函数。迭代次数为i 时，损失函数为，其中，θi代表学习过程中的网络参数。经过一段时间的学习后， 新的θi更新θ−。具体的学习过程根据：。

**B,基于递归神经网络的DRL**

DRL面临的问题往往具有很强的时间依赖性，而递归神经网络**适合处理和时间序列相关的问题**。

对于时间序列信息，深度Q网络的处理方法是加入**经验回放机制**。但是经验回放的记忆能力有限，每个决策点需要获取整个输入画面进行感知记忆。将**长短时记忆网络**与深度Q网络结合，提出**深度递归Q网络(deep recurrent Q network，DRQN)**，在**部分可观测马尔科夫决策过程(partially observable Markov decision process， POMDP)**中表现出了更好的鲁棒性，同时在缺失若干帧画面的情况下也能获得很好的实验结果。

1. **马尔可夫决策过程(MDP)**

**定义：**是**序贯决策**（sequential decision）的数学模型，用于在系统状态具有**马尔可夫性质**的环境中模拟**智能体**可实现的**随机性策略与回报**。

**序贯决策：**按**时间顺序**排列起来，以得到按**顺序的各种决策**，是用于**随机性或不确定性**动态系统**最优化**的决策方法。

**马尔可夫性质：**是概率论和数理统计中的一个概念。当一个随机过程在给定现在状态及所有过去状态情况下，其未来状态的条件概率分布仅依赖于当前状态。

**MDP理论基础：马尔科夫链，**即是具有**马尔可夫性质**且存在于**离散的指数集和状态空间**内的**随机工程。**

**分类：**离散时间马尔可夫决策过程、连续时间马尔可夫决策过程

**变体：**部分可观察马尔可夫决策过程、约束马尔可夫决策过程、模糊马尔可夫决策过程

**5个模型要素：**

A,状态：状态是对环境的描述，在智能体做出动作后，状态会发生变化，且演变具有马尔可夫性质。MDP所有状态的集合是状态空间。状态空间可以是离散或连续的。

B,动作：动作是对智能体行为的描述，是智能体决策的结果。MDP所有可能动作的集合是动作空间。动作空间可以是离散或连续的。

C,策略：MDP的策略是按状态给出的，动作的条件概率分布，在强化学习的语境下属于随机性策略。

D,奖励：智能体给出动作后环境对智能体的反馈。是当前时刻状态、动作和下个时刻状态的标量函数。

E,回报：回报是奖励随时间步的积累，在引入轨迹的概念后，回报也是轨迹上所有奖励的总和。

**值函数：**状态值函数、动作值函数(贝尔曼方程)

**MDP适用的强化学习算法分为两类：**

**值函数算法：**通过**迭代策略的值函数**求得全局最优

**分类：**

**动态规划：属于“基于模型的强化学习”。**作为贝尔曼最优化原理的推论，可求得有限时间步的MDP至少存在一个全局最优解，且该最优解是确定的。要求状态值函数和动作值函数的贝尔曼方程**已知**。**分为策略迭代、值迭代。**核心思想是最优化原理：最优策略的子策略在一次迭代中也是以该状态出发的最优策略，因此在迭代中不断选择该次迭代的最优子策略能够收敛至MDP的全局最优。

**随机模拟：属于“无模型的强化学习“。**以**蒙特卡罗方法**和**时序差分学习**为代表。求解MDP可用的时序差分学习算法包括SARSA算法（State Action RewardState Action, SARSA）和**Q学习（Q-Learning）算法。**二者都利用了MDP的马尔可夫性质，但前者的改进策略和采样策略是同一个策略，因此被称为**“同策略（on policy）”算法**，而后者采样与改进分别使用不同策略，因此被称为**“异策略（off policy）”算法。**

**策略搜索算法：**通过**搜索策略空间**得到全局最优。例子：REINFORCE算法使用随机梯度上升求解（可微分的）策略函数的参数使得目标函数最大。

**6、Q-learning算法**

**7、价值函数近似**

**8、卷积神经网络**

**10、损失函数**

**11、经验池、离线学习、在线学习**

**12、Ubuntu, Cuda, Cudnn, Tensorflow, OpenAI Gym**

**13、改进方法(Nature DQN)**

**14、Double DQN、Prioritied Replay、Dueling Network**

**15、DQN拓展到连续控制的算法-NAF**

**二、基础知识学习**

**1、随机梯度下降、微分损失函数**

**2、单一神经网络智能体(AGENT)**

**3、超参数、模拟器内部状态**

**4、函数逼近器**

**5、DEEPMIND、成功案例:TD-gammon、自我对弈**

**6、最相似的先前工作：神经拟合Q学习(NFQ)**

**7、可视化值函数**

**8、概率论、线性代数数学基础**

**9、Python**