Deep Q Network(DQN)学习计划

一、学习规划

1、机器学习、深度学习、强化学习、深度强化学习

(1) 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

人工神经网络是由**大量处理单元**互联组成的非线性、自适应信息处理系统。

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)

(2) 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。

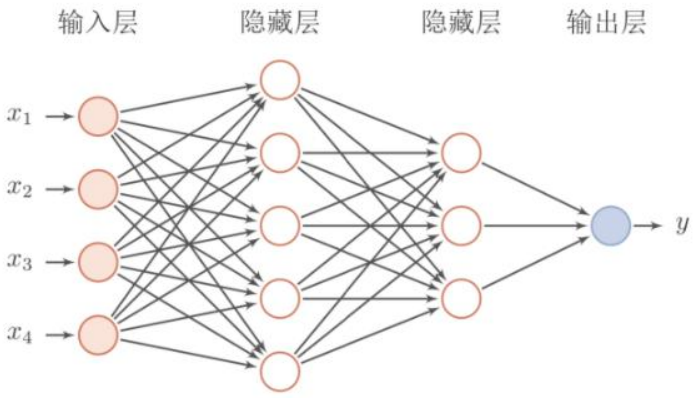
典型模型：

A,卷积神经网络模型(Convolutional Neural Networks, CNN)

定义：是一类包含**卷积计算**且具有**深度结构**的**前馈**神经网络。具有**表征学习**(representation learning)能力，能够按其阶层结构对输入信息进行平移不变分类(shift-invariant classification)，因此也被称为“**平移不变人工神经网络**(Shift-Invariant Artificial Neural Networks,**SIANN**)”。

卷积计算：在**泛函分析**中，通过两个函数 f 和 g 生成第三个函数的一种**数学算子**，表征函数 f 与 g 经过翻转和平移的重叠部分函数值乘积对重叠长度的积分。

前馈神经网络：是一种**最简单**的神经网络，各神经元分层排列。每个神经元只与前一层的神经元相连。接收前一层的输出，并输出给下一层．各层间没有反馈。



表征学习(特征学习)：是学习一个特征的技术的集合，即将原始数据**转换**成为能够被机器学习来有效开发的一种形式。它**避免**了**手动**提取特征的麻烦，允许计算机学习使用特征的同时，也学习如何提取特征(学习如何学习)。

CNN **仿造**生物的视知觉机制构建，可以进行**监督学习**和**非监督学习**，其**隐含层内的卷积核参数共享**和**层间连接的稀疏性**使得卷积神经网络能够以**较小的计算量**对**格点化**特征，例如像素和音频进行学习，有稳定的效果且对数据没有额外的特征工程要求。

结构：

输入层：可以处理多维数据。

一维卷积神经网络的输入层接收一维或二维数组，其中一维数组通常为时间或频谱采样；

二维数组可能包含多个通道，二维卷积神经网络的输入层接收二维或三维数组；

三维卷积神经网络的输入层接收四维数组。

由于在计算机视觉领域应用较广，因此许多研究在介绍其结构时**预先假设**了**三维输入数据**，即**平面上的二维像素点**和 **RGB 通道**。

RGB 通道：是工业界的一种**颜色标准**，是通过对**红(R)、绿(G)、蓝(B)**三个颜色通道的变化以及它们相互之间的叠加来得到各式各样的颜色的，RGB 即是代表红、绿、蓝三个通道的颜色，这个标准几乎包括了人类视力所能感知的所有颜色，是运用最广的颜色系统之一。

由于使用**梯度下降算法**进行学习，卷积神经网络的**输入特征**需要进行**标准化处理**。具体地，在将学习数据输入卷积神经网络前，需在通道或时间/频率维对输入数据进行**归一化**，若输入数据为像素，也可将分布于 $[0,255]$ 的原始像素值归一化至 $[0,1]$ 区间。输入特征的标准化有利于提升卷积神经网络的学习效率和表现。

梯度下降算法:是**迭代法**的一种,可以用于求解**最小二乘问题**(线性和非线性都可以)。在**求解**机器学习算法的**模型参数**，即无约束优化问题时，梯度下降是**最常采用的方法**之一，**另一种常用的方法**是**最小二乘法**。

在**求解损失函数的最小值**时，可以通过梯度下降法来一步步的迭代求解，得到最小化的损失函数和模型参数值。**反过来**，如果我们需要**求解损失函数的最大值**，这时就需要用**梯度上升法**来迭代了。

在机器学习中，基于基本的梯度下降法发展了两种梯度下降方法，分别为**随机梯度下降法**和**批量梯度下降法**。

最小二乘问题：也称**最小平差问题**，**数值逼近**的重要问题之一，是用**离散平方**逼近技术求**拟合曲线**的问题。

最小二乘法：是一种**数学优化**技术。它通过**最小化误差的平方**和寻找数据的最佳函数匹配。利用最小二乘法可以简便地**求得未知的数据**，并使得这些求得的数据与实际数据之间**误差的平方和为最小。**

隐含层：包含**卷积层、池化层和全连接层** 3 类常见构筑，也可能有 Inception 模块、残差块等复杂构筑。卷积层和池化层为卷积神经网络**特有**。卷积层中的**卷积核**包含**权重系数**，而池化层**不包含**权重系数。

卷积层：对输入数据进行特征提取，其内部包含多个卷积核，**组成卷积核的每个元素**都对应一个权重系数和一个偏差量。卷积核在工作时，会有规律地扫过输入特征，在**感受野**内对输入特征**做矩阵元素乘法求和**并**叠加偏差量。**

感受野：卷积层内每个神经元都与前一层中位置接近的区域的多个神经元**相连，区域的大小**取决于卷积核的大小。

卷积层参数：包括卷积核**大小、步长和填充**，三者共同决定了卷积层**输出特征图的尺寸**，是卷积神经网络的**超参数。**卷积核**越大**，可提取的输入特征越复杂。卷积**步长**定义了卷积核相邻两次扫过特征图时位置的距离。

随着卷积层的堆叠，特征图的尺寸会逐步减小。为此，**填充**是在特征图通过卷积核之前人为增大其尺寸以抵消计算中尺寸收缩影响的方法。**常见的填充方法**为**按 0 填充**和**重复边界值填充。**

激励函数：协助表达复杂特征，操作通常在卷积核之后。

池化层：在卷积层进行特征提取后，输出的特征图会被传递至池化层进行**特征选择和信息过滤**。池化层包含**预设定的池化函数**，其功能是将特征图中**单个点**的结果**替换为**其**相邻区域**的特征图统计量。池化层选取池化区域与卷积核扫描特征图步骤相同，由池化大小、步长和填充控制。

全连接层：等价于传统前馈神经网络中的**隐含层**。全连接层**位于**卷积神经网络隐含层的**最后部分**，并**只向其它全连接层**传递信号。特征图在全连接层中会**失去空间拓扑结构，被展开为向量**并通过激励函数。

输出层：结构和工作原理与传统前馈神经网络中的输出层相同。

优化：正则化、分批归一化、跳跃连接

加速：通用加速技术、FFT 卷积、权重稀疏化

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(**用于从**没有标记**响应的输入数据组成的数据集中进行推断。最常见的无监督学习方法是聚类分析。**)**

B,自顶向下的监督学习

(3) 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。

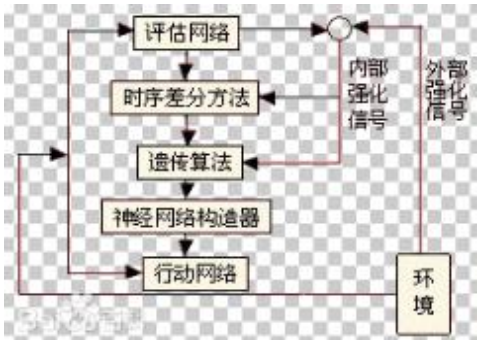
智能体：agent，进行机器学习的代理，可以感知外界环境的状态进行决策、对环境做出动作并通过环境的反馈调整决策。

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;

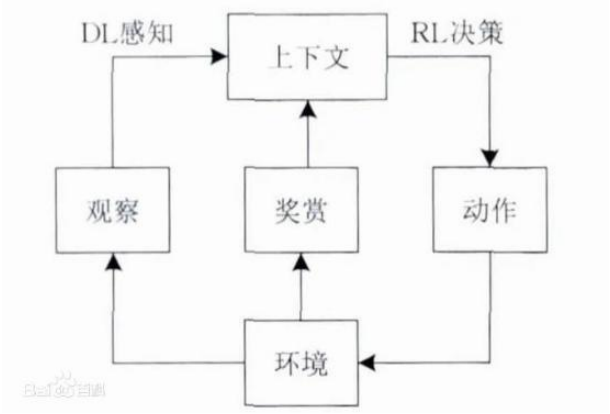
B,如何**选择**建立信号以及如何通过学习来**修正**不同状态－动作对的**值**;

C,如何**根据**这些值来选择适合的动作;

(4) 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。

Q learning 是一种 **off-policy 离线学习法。**

- On-policy: The agent learned and the agent interacting with the environment is the same.
- Off-policy: The agent learned and the agent interacting with the environment is different.

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 $L_i(\theta_i) = \mathrm{E}_{(s,a,r,s')}[(y_i^{\mathrm{DQN}} - Q(s,a;\theta_i))^2]$，其中 $y_i^{\mathrm{DQN}} = r + \gamma \max_{a'} Q(s',a';\theta^-)$，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：

$$\nabla_{\theta_i} L_i(\theta_i) = \mathrm{E}_{(s,a,r,s')}[(r + \gamma \max_{a'} Q(s',a';\theta^-) - Q(s,a;\theta_i))\nabla_{\theta_i} Q(s,a;\theta_i)].$$

。

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。

2、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。

$$\mathcal{S}=\{s_1,s_2,\cdots,s_\tau\}$$

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。 $\mathcal{A}=\{a_1,a_2,\cdots,a_\tau\}$

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。 $\pi(a|s)=p(a|s)$

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。 $R=R(s_t,a_t,s_{t+1})$

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。 $G=\sum_{t=0}^{\tau-1}R_{t+1}$

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 Reward State 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 算法使用随机梯度下降求解（可微分的）策略函数的参数使得目标函数最小。

3、Q-learning 算法

主要思想：基于**值**的算法，**Q** 即为 Q（s,a）就是在某一时刻的 s 状态下(s∈S)，采取动作 a (a∈A)动作能够**获得收益的期望**，环境会根据 agent 的动作反馈相应的回报 reward。**将 State 与 Action 构建成一张 Q-table 来存储 Q 值**，然后**根据 Q 值来选取**能够获得最大的收益的动作。

4、损失函数(代价函数)

定义：将随机事件或其有关随机变量的取值**映射为**非负实数以表示该随机事件的**“风险”或“损失”**的函数。

在应用中，通常作为**学习准则与优化问题**相联系，即通过**最小化**损失函数求解和评估模型。

5、Ubuntu, Cuda, Cudnn, Tensorflow, OpenAI Gym

Ubuntu：是一个以桌面应用为主的 **Linux 操作系统**。

Cuda：是一种由 NVIDIA 推出的通用并行**计算**架构，使 GPU 能够解决复杂的计算问题。

Cudnn：是 NVIDIA CUDA 的一个 GPU 加速的**深层神经网络原语库**。

Tensorflow：是一个基于数据流编程的符号数学系统，被广泛应用于各类机器学习算法的**编程实现**，其前身是谷歌的神经网络算法库 DistBelief。

OpenAI Gym：是一款用于研发和比较强化学习算法的工具包，它支持训练智能体（agent）做任何事。与其他的数值计算库**兼容**，如 tensorflow 或 theano 库。现在主要支持的是 python 语言。

6、DQN 算法步骤

以下操作执行N个Episode：

- 将游戏画面s进行预处理，送入CNN，输出每个动作的Q值
- 根据$\epsilon - greedy$算法计算应该采取的动作$a = argmax(Q(s, a, \theta))$
- 执行动作a，转移到下一状态s'，收到环境给出的回报
- 将序列$< s, a, r, s' >$存入经验池
- 从经验池采样一个batch的数据，训练Q Network，实际上是一个由输入预测输出的回归问题

$$loss = (r + \gamma max_{a'} Q(s', a'; \theta') - Q(s, a; \theta))^2$$

- CNN（当前网络）训练若干代之后，将其参数拷贝给目标网络。

7、改进方法(Nature DQN)

8、Double DQN、Prioritied Replay、Dueling Network

9、DQN 拓展到连续控制的算法-NAF

二、基础知识学习

1、超参数：在机器学习的上下文中，**是在开始学习过程之前设置值的参数**，而不是通过训练得到的参数数据。通常情况下，需要对超参数进行优化，给学习机选择一组最优超参数，以提高学习的性能和效果。

2、DEEPMIND 成功案例:TD-gammon(自我对弈)(了解)

3、DQN 与神经拟合 Q 学习(NFQ)区别

除 DQN 每 N 个步骤更新目标网络外，二者工作差不多。