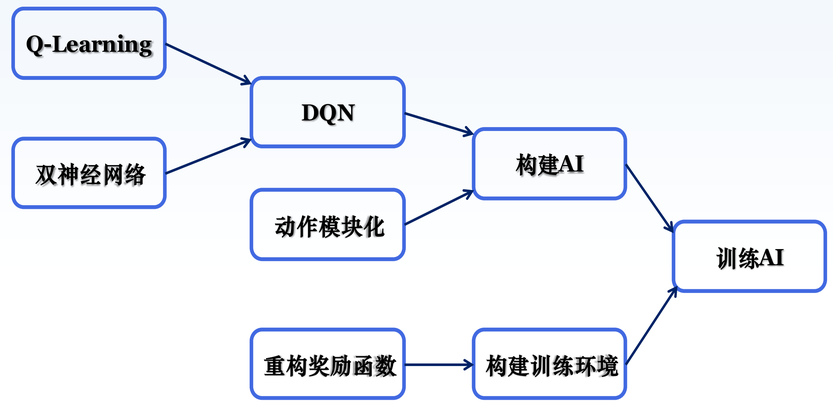
人工智能·太空大战
——基于DQN的强化学习

小组成员：林子开 鞠扬 苏宇驰

项目开发流程

在强化学习算法的设计上，本小组以Q-Learning算法为基础，引入异步更新的双神经网络搭建**DQN模型**，切断经验序列的相关性，提高经验利用率，加速收敛。在神经网络设计上，本小组基于先验知识，提取多个战局特征作为输入层，搭建**卷积神经网络**，提高模型泛化能力。在动作设计上，巧妙避开繁琐的船厂生产指令与舰队飞行指令，为AI提供**模块化封装**的战斗策略，优化了动作空间。在模型训练阶段，根据人类观察经验**重构奖励函数**，促使AI根据战局变化，自动调整各类战场要素的权重。

关键词：太空采矿，DQN，神经网络，模块化设计，重构奖励



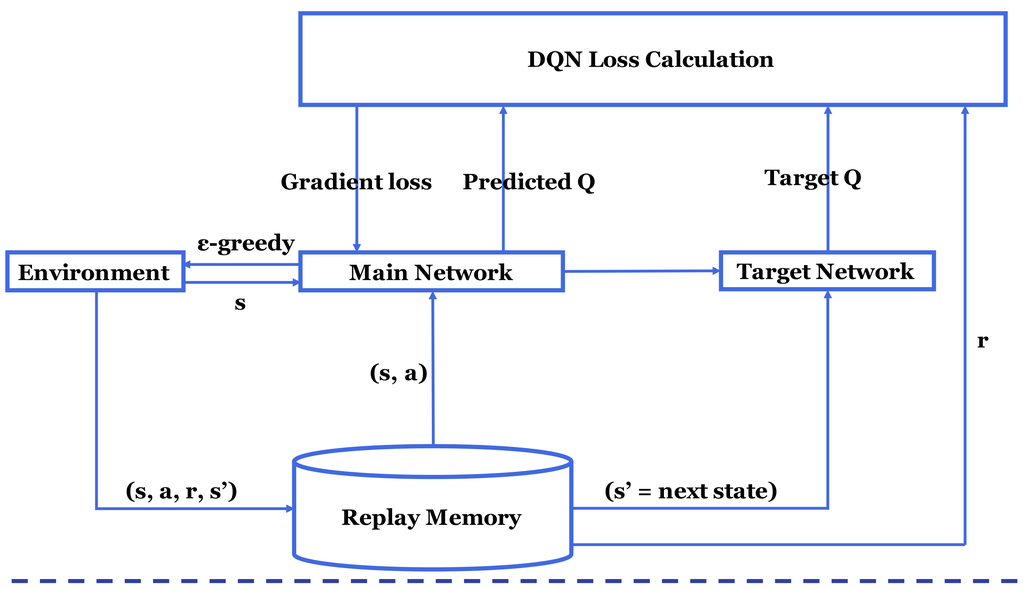
DQN原理

在DQN中，我们引入**两个神经网络**：主网络与目标网络。训练时，从经验池中**等概率抽取样本**，对主网络进行训练，目标网络暂时不动。训练的目标为最小化损失函数的期望：

$$J = \mathbb{E}\left[L\left(r + \gamma \max_{a \in \mathcal{A}(s')} \hat{Q}(s', a; w), \hat{Q}(s, a; w)\right)\right]$$

其中$L(\cdot,\cdot)$为损失函数，本项目选择Huber损失函数

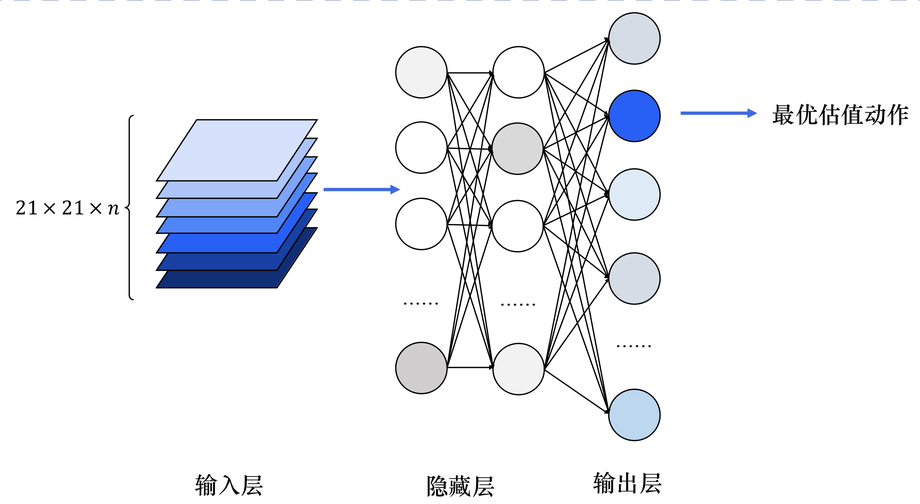
每间隔400回合，就**用主网络覆盖目标网络**。DQN算法使用双神经网络，能够**使收敛过程更稳定**；从经验池中均匀抽取样本，能够**切断**样本序列的**相关性**，并且**提高样本利用率**。以下是DQN算法的示意图。



神经网络结构

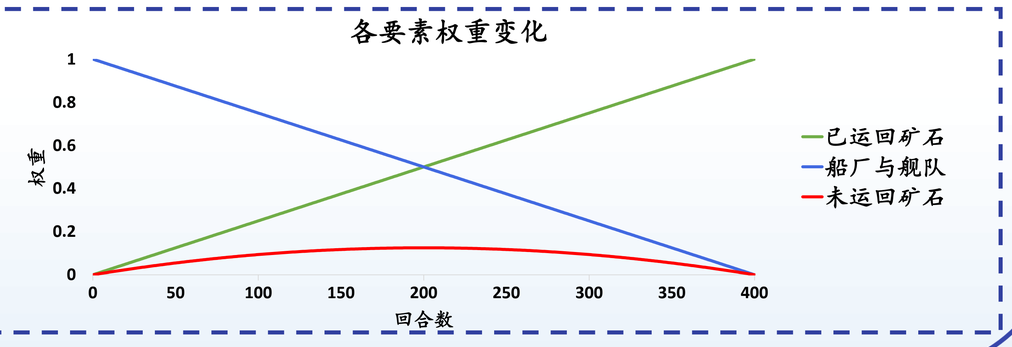
神经网络包括两个卷积层，一个展平层，两个全连接层。**输入为**基于先验知识提取的n个特征，为 $21 \times 21 \times n$的张量。**输出为**对在当前状态的所有可行动作a_i的估值$\hat{Q}(s, a_i; w)$，AI按照贪心策略，执行估值最高的动作。

本小组提取的特征包括：**天然矿石分布，敌我战舰规模，敌我战舰分布，已采未送回矿石分布，敌我船厂分布，敌方舰队动向，我方舰队动向，高威胁性敌方船厂**。



创新点1：重构奖励函数

综合考虑四种战场要素，包括：**①已开采且运回大本营的矿石，②已经开采且仍由舰队携带的矿石，③舰队数量，④船厂数量**。对四要素加权求和可得每回合评分。根据Kore的游戏规则，最终只有运回船厂的矿石计入总数，因此**①的权重逐渐上升，②③④的权重逐渐下降**。此外，基于人类的观察经验，对AI的扩张、攻击、采矿等行为引入**额外的奖惩机制**，有助于AI更快习得最优策略。



创新点2：动作模块化

本小组发现让AI学会直接编写各船厂生产指令与舰队飞行指令的难度较大，于是将多个基本动作按照具有特定倾向的固定策略进行**模块化封装**，再将模块化的动作“组合拳”提供AI，AI执行由神经网络估值后价值最高的“组合拳”。我们目前一共设计了五套模块化封装的“组合拳”，分别是：**优先进攻、优先防御、优先采矿、优先扩张、中庸之道**。五套“组合拳”执行策略如下所示：