·第二阶段总结：  
我们使用了 Deep Q-learning，也称 Deep Q-network（简称 DQN）作为强化学习的底层算法。从数学原理上看，DQN使用了随机梯度下降的方式，近似求解贝尔曼最优方程。从模型特点上看，DQN使用了两个神经网络（main network与target network）异步更新，并从经验池（replay buffer）中抽取样本进行训练，能够使学习过程较为稳定，收敛较快。

DQN技术主要有两个特点：双神经网络与经验回放。第一项技术是使用主网络和目标网络，并进行异步更新。主网络用于给出对 ˆQ(s, a; w) 的估计值，并且频繁更新；而目标网络则在短期内保持不变，用于给出r + γ maxa∈A(s′) ˆQ(s’, a; w) - ˆQ(s’, a; w)的估计值。在每次迭代中，我们从经验池（replay buffer）中按照均匀分布随机抽样出一批样本 {(st, at, rt, st+1)} 输入主网络。主网络给出估计值ˆQ(s, a; w)，目标网络给出目标值r + γ maxa∈A(s′) ˆQ(s’, a; w) - ˆQ(s’, a; w)，然后使用随机梯度下降的方式，最小化损失函数的期望。在一定轮次之后，将主网络的权重覆盖到目标网络上。

第二项技术是经验回放（experience replay）。当我们收集一些经验样本后，我们不按照收集的 顺序来使用它们，而是将它们储存在回放缓冲区里。每次更新主网络时，我们按照均匀分布从回放 缓冲区中提取一组样本（batch）。尽管样本是贯序收集的，相邻样本之间具有显著的相关性，但是使用抽样的方式，则可以打破样本间的前后相关性，让神经网络从样本中学习到更普适性的知识。此外，通过将经验保存在经验池中，每个经验都可以被多次学习，也提高了经验利用率。

在神经网络的特征提取过程中，我们决定提取7个战场特征：天然矿石数量、战舰规模、舰队分布、已采矿石、船厂分布、敌方舰队动向、我方舰队动向，构成21 × 21 × 7的张量，输入神经网络对Q(s, ai)进行拟合。

在建立强化学习模型的过程中，我们发现，直接让 AI 学会制定各个船厂的具体指令十分困难。 比如，AI 基本无法学会制定出“N10E5C”“N5E6S5W”这类精巧复杂的舰队指令。于是我们转变了思路，为 AI 提供一些封装成模块的动作。我们提取并总结出了 5 套“组合拳法”，将基本动作采用不同顺序组合在一起，分别应对优先进攻、优先挖矿、优先扩张、优先防御、紧急绝处逢生的战略部署，让 AI 根据战场状况，选择最适合的组合动作。

在训练的过程中，我们为AI提供了并且采用ϵ-greedy的策略，使AI保持较高的探索性。

我们小组重构了奖励函数，使用了根据战局进程而调整各要素权重的综合评分奖励机制，并着重考虑了四种要素的价值：已经开采并且运回大本营的矿石、已经开采但仍然由舰队携带的矿石、舰队和船厂。