一、 注意力机制与多智能体强化学习

注意力机制（Attention）是一种重要的深度学习方法，它最主要的用途是自然语言处理。

为什么要因为注意力机制在 Attention 诞生之前，已经有 CNN 和 RNN 及其变体模型了，那为什么还要引入 Attention 机制？主要有两个方面的原因，如下：

1. 计算能力的限制：当要记住很多“信息“，模型就要变得更复杂，然而目前计算能力依然是限制神经网络发展的瓶颈。
2. 优化算法的限制：LSTM只能在一定程度上缓解RNN中的长距离依赖问题，且信息“记忆”能力并不高。

1.1 自注意力在中心化训练中的应用

自注意力机制（Self-Attention）是改进多智能体强化学习的一种有效技巧。

自注意力机制在非合作关系的 MARL 中普遍适用。如果系统架构使用中心化训练，那么 m 个价值网络可以用一个神经网络实现，其中使用自注意力层。如果系统架构使用中心化决策，那么 m 个策略网络也可以实现成一个神经网络，其中使用自注意力层。在 m 较大的情况下，使用自注意力层对效果有较大的提升。

二、 AlphaGo 与蒙特卡洛树搜索

之前章节的强化学习方法都是无模型的强化学习（Model-Free），包括价值学习（Value-Based）和策略学习（Policy-Based）。而蒙特卡洛树搜索（Monte Carlo Tree Search, MCTS）是一种基于模型的强化学习方法。

AlphaGo 依靠 MCTS 做决策，而决策的过程中需要策略网络和价值网络的辅助。

AlphaGo 的价值网络 $v(s;w)$ 用于对状态价值函数 $V_\pi(s)$ 做近似。

2.1 蒙特卡洛树搜索（MCTS）

若已经训练好了策略网络 $\pi(a|s;\theta)$ 和价值网络 $v(s;w)$。AlphaGo 真正跟人下棋时，做决策的不是策略网络或价值网络，而是蒙特卡洛树搜索，缩写 MCTS。MCTS 不需要训练，可以直接做决策，策略网络和价值网络的目的时辅助 MCTS。

包括四个步骤:

- 选择 (Selection)
  - 三类节点
    - 未访问: 还没有评估过当前局面
    - 未完全展开: 被评估过至少一次, 但是子节点 (下一步的局面) 没有被全部访问过, **可以进一步扩展**
    - 完全展开: 子节点被全部访问过
  - 我们找到 **目前认为**「最有可能会走到的」一个未被评估的局面 (双方都很聪明的情况下), 并且 **选择** 它
  - $\text{score}(a) \triangleq Q(a) + \dfrac{\eta}{1+N(a)} \cdot \pi(a|s;\theta)$
- 扩展 (expansion)
  - 将刚刚选择的节点加上一个统计信息为「0/0」的节点
- 模拟 (Simulation)
  - 快速走子 (Rollout)
- 回溯 (Backpropagation)
  - 回溯, 沿途更新各个父节点的统计信息