Routing Algorithms

Model Architecture

对于每个液滴，其动作空间是离散的，动作数量为 5；其观测状态是一张$10 \times 15 \times 15$的图像，其中10代表通道数，15×15代表图像的大小，每个通道是一张局部观测的灰度图，代表不同的环境信息。在给定 SARS(s_t, a_t, r_t, s_{t+1}) 元组的情况下，我们的目标是学习一个策略$\pi(a_t|s_t)$，使得在给定当前状态s_t的情况下，选择一个动作a_t，使得期望的累积奖励最大化。具体地，我们使用了 Dueling DQN 算法来学习动作价值函数$Q(s_t, a_t)$，其中s_t是当前状态，a_t是当前动作，Dueling DQN 算法将动作价值函数分解为状态价值函数$V(s_t)$和优势函数$A(s_t, a_t)$，其中$V(s_t)$是状态s_t的价值，$A(s_t, a_t)$是状态s_t下采取动作a_t相对于其他动作的优势。

$$Q(s_t, a_t) = V(s_t) + A(s_t, a_t) - \frac{1}{|A|} \sum_{a'} A(s_t, a')$$

在模型结构中，我们使用了卷积神经网络来提取局部观测的特征，然后将特征输入到全连接层中，最后分别输出动作价值函数$Q(s_t, a_t)$和状态价值函数$V(s_t)$。在训练过程中，我们使用了经验回放机制来减小样本之间的相关性，使用了目标网络来减小训练过程中的不稳定性，使用了 Huber Loss 来减小异常值的影响。

$$L = f_{(s_t, a_t, r_t, s_{t+1}) \sim B} \left[r_t + \gamma \max_{a'} Q_{\text{target}}(s_{t+1}, a') - Q_{\text{policy}}(s_t, a_t) \right]$$

其中f是 Huber Loss 函数，B是 Batch，γ是折扣因子，Q_{target}是目标网络，Q_{policy}是策略网络。

Contrastive Learning

尽管我们的观测是局部的，但我们能够获得更全面的环境信息，其中最重要的一种信息是液滴在t时刻可以采取的合法动作集A_t^+，与之对应的是会导致任务失败的动作集A_t^-。离线强化学习算法对采集数据的质量要求较高，我们在调用 Gym 环境迭代时，违法动作的出现会导致环境的重置，因此一个 episode 中只有一个违法动作，这就导致了数据分布的不平衡，agent 很难学习到违法动作的价值，而如果我们想收集更多的违法动作又会增加环境初始化的次数，这涉及到每次对布局的重新规划，会导致时间的浪费。

因此我们想到了利用无监督学习的方法来解决对违法动作的学习问题。具体来说，我们沿用了对比学习的思想，将合法动作看作正样本，将违法动作看作负样本，将观测状态视作 query，然后将动作价值Q

视为分数，将正样本和负样本的分数进行对比，从而在不需要数据增强或重采样的情况下，让 agent 学习到违法动作的价值。我们参考了 InfoNCE Loss 来设计对比学习损失：

$$L_{\text{contrastive}} = -\log \frac{\sum_{a \in A_t^+} e^{Q(s_t,a)}}{\sum_{a \in A_t^+} e^{Q(s_t,a)} + \sum_{a \in A_t^-} e^{Q(s_t,a)}}$$