主要成果

找到一个能运行的PPO算法，学习对照其与目前PPO的异同之处。
PS：由于本周忙于考试复习和课程大作业，因此花费时间不多。

不同之处分析

参考项目：https://github.com/seolhokim/Mujoco-Pytorch

- 基类和网络结构：
  PPOModel继承自object，使用PolicyNet和ValueNet作为策略网络和价值网络。
  PPO继承自nn.Module，使用Actor和Critic作为策略网络和价值网络。
- 优化器和损失函数：
  PPOModel直接在初始化方法中定义了优化器和损失函数，而PPO通过参数传递。
- 获取动作方法：
  PPOModel返回动作、对数概率和均值，而PPO只返回均值和标准差。
- 计算回报和优势方法：
  PPOModel使用循环计算回报和优势，并进行归一化处理，而PPO使用列表推导式。
- 训练网络方法：
  PPOModel从mem中获取数据，使用固定次数的优化迭代，策略损失计算中包含了熵项，价值损失计算使用均方误差损失。
  PPO从ReplayBuffer中采样数据，使用参数化的训练轮数，策略损失计算中包含了熵系数，价值损失计算使用PPO2技术。