PolicyGradient: 网络结构

Trade\_env:

1. 环境TradeEnv:
   1. 通过QuandlEnvSrc更新读取getStock中的获取的数据，得到客观股票的状态
   2. Investor通过自己的买入卖出状态更新自己的时候的现金cash和仓位stock，并计算每次操作的净资产，作为回报reward

Run：

1. 跑模型，获得对应动作action的observation；
   1. 通过observation选择action：observation输入到网络中，输出softmax得到对应动作的概率，然后得到其动作的多项式分布，按照概率选取action
   2. Action作用在environment中，更新observation：如此处，相对应的买卖动作action施加后，会更新股民Investor的现金cash和仓位状态，从而得到新的observation\_
   3. 在b发生的同时，得到在每个observation下，对应action作用的reward，一一对应
2. 更新模型参数；
   1. 利用一系列取得的reward,根据该环境，制定一个奖励函数Vt
   2. 将在某observation选择该动作的概率的log 与其对应的reward 相乘得到每一步的loss值，即
   3. Loss值求和，得到总loss, 后向传播求导，更新模型