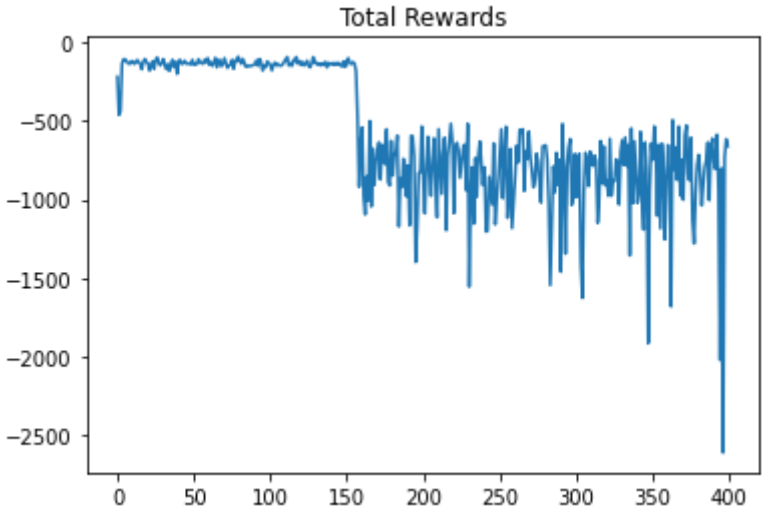
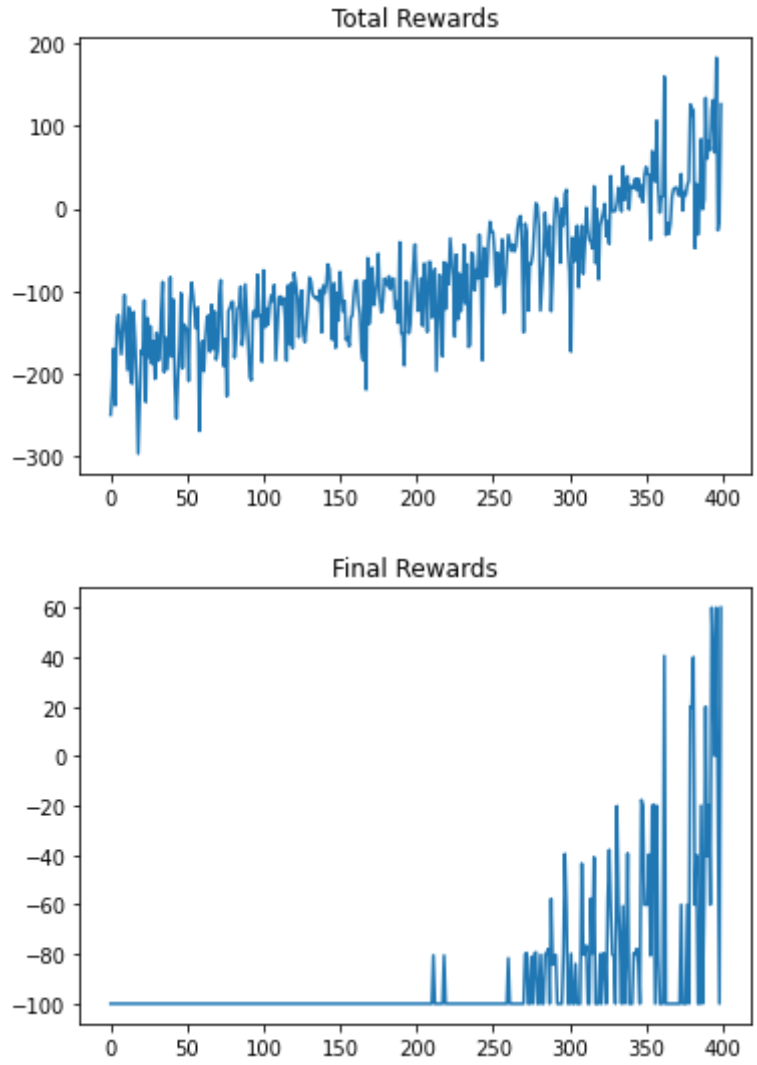
1.accumulated reward（去掉标准化）: total reward在训练150epoch后突然下降




解决方法：调lr/网络参数/random seed/加标准化

2.accumulated reward+标准化+network结构调整（8->16->32->4）:




test total reward: 55.36

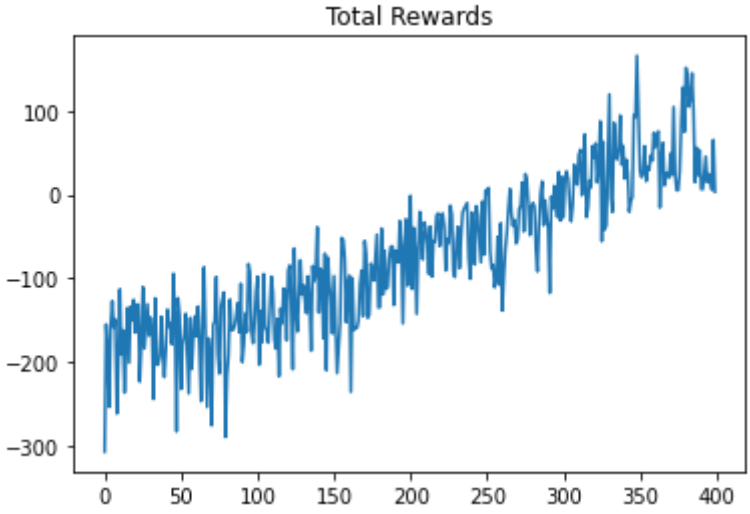
3.actor-critic:

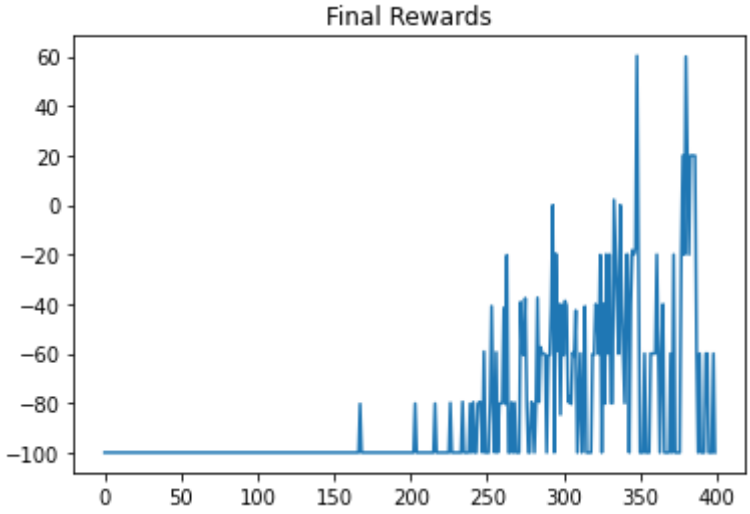
关键：在rewards-baseline需要对reward标准化，否则训不起来

critic network：8->16->32->1

3.1

loss func: smooth_l1_loss

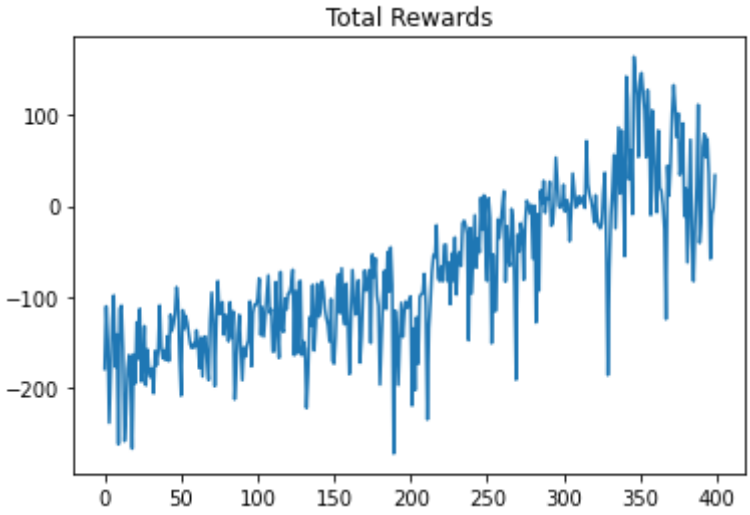


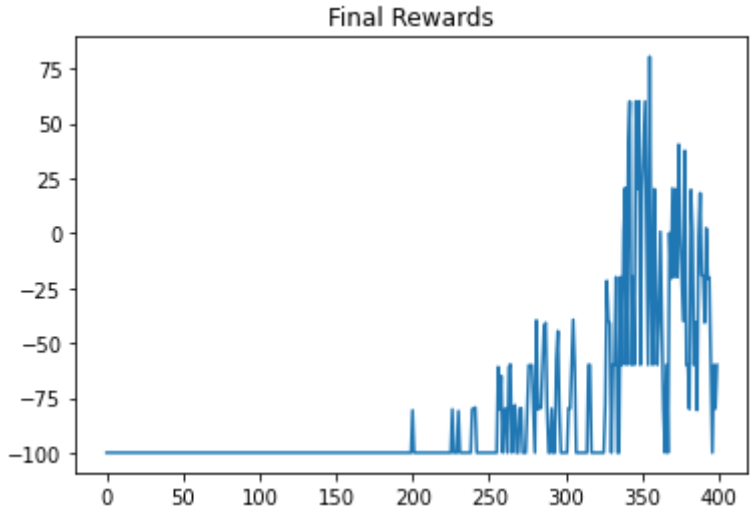




test total reward: 88.58

3.2

loss func: MSELoss






test total reward: 77.53