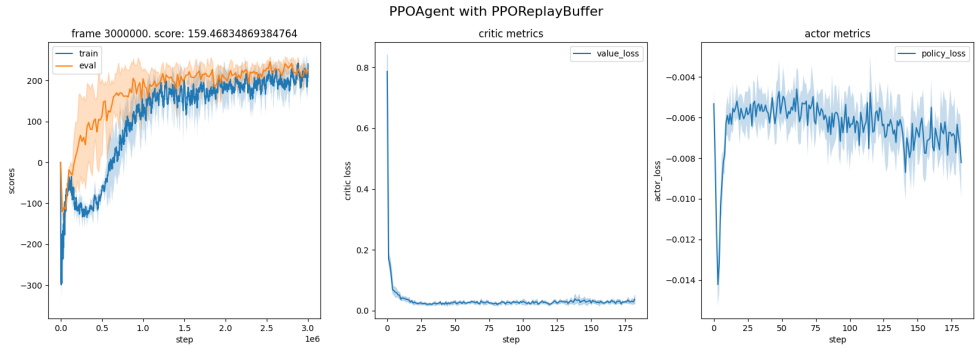
HW4

张瑞泽 2019011189



可以由图中看出PPO训练结果，return曲线最终到达200以上，value_loss不断下降，policy_loss接近于0。此外，值得一提的是，和之前off-policy算法结果相比，PPO会更加稳定。