Offline Reinforcement Learning with Uncertainty Critic Regularization Based on Density Estimation

李超、吴凤鸽、赵军锁

邮箱：lichao2022@iscas.ac.cn

电话：15665825817

背景：在离线强化学习中，已有的研究提出了策略约束、值函数正则化和不确定性估计等方法促进学习策略与行为策略相似，解决分布偏移问题。然而策略约束方法约束严格导致学习到次优策略；值函数正则化方法使值函数的估计过于保守；不确定性估计方法的估计结果可能存在偏差，导致学习到的策略不准确。

解决的科学难题：离线强化学习策略学习过程中存在的分布偏移问题。

创新点：（1）利用变分自编码器对离线数据集的状态动作分布进行建模，对样本进行更准确的不确定性估计；（2）引入分布外动作的采样作为额外的训练样本，并将对分布外动作的惩罚作为额外的训练目标；（3）在线微调阶段，将不确定性作为优先回放缓冲区的权重，以减少采样高不确定性样本的可能性。

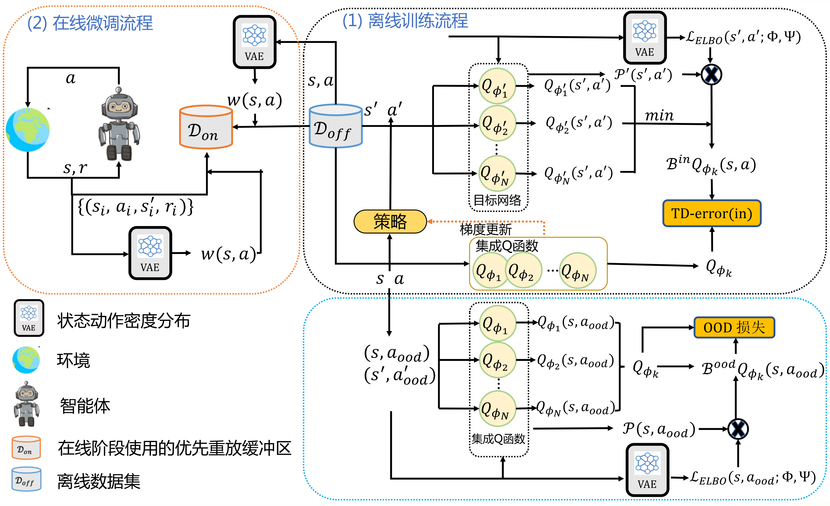

图1：模型结构

成果1：有效缓解分布偏移问题，效果在某些环境下超过先前的SOTA方法

Task Name	10%BC	CQL	AWAC	TD3+BC	IQL	PBRL	EDAC	Ours
halfcheetah-random	2.2 ± 2.1	24.6 ± 0.5	18.1 ± 0.3	15.0 ± 0.3	10.9 ± 1.5	33.1 ± 13.5	25.7 ± 0.3	**35.0 ± 0.4**
halfcheetah-medium	42.8 ± 0.9	46.9 ± 0.5	50.1 ± 0.7	48.7 ± 0.6	44.9 ± 1.3	58.5 ± 0.9	**67.4 ± 0.7**	**67.7 ± 0.6**
halfcheetah-medium-replay	40.1 ± 0.9	46.2 ± 1.1	45.6 ± 1.0	45.2 ± 1.0	43.5 ± 2.4	48.1 ± 0.8	**66.6 ± 0.8**	65.2 ± 1.1
halfcheetah-medium-expert	90.1 ± 8.9	96.4 ± 0.9	95.3 ± 0.7	95.1 ± 1.4	93.1 ± 3.9	92.4 ± 1.8	**104.6 ± 1.6**	102.2 ± 3.4
halfcheetah-expert	90.6 ± 1.8	96.9 ± 0.7	97.9 ± 1.2	100.7 ± 1.9	95.8 ± 1.9	97.2 ± 2.3	**106.4 ± 1.4**	**105.7 ± 1.2**
hopper-random	31.4 ± 0.1	12.5 ± 0.1	32.1 ± 0.3	31.6 ± 0.1	11.5 ± 0.3	31.5 ± 0.1	30.9 ± 1.0	**33.1 ± 0.1**
hopper-medium	47.5 ± 4.7	56.4 ± 9.9	61.6 ± 8.3	55.2 ± 8.6	50.9 ± 7.8	83.8 ± 19.1	**100.9 ± 0.7**	98.6 ± 2.4
hopper-medium-replay	42.8 ± 8.1	96.2 ± 9.6	99.7 ± 0.4	74.9 ± 16.1	89.1 ± 16.1	101.3 ± 0.3	103.6 ± 0.1	**104.2 ± 0.3**
hopper-medium-expert	106.4 ± 9.7	98.8 ± 15.2	106.3 ± 7.1	108.7 ± 2.6	106.2 ± 6.7	**111.1 ± 0.3**	110.0 ± 1.3	109.5 ± 0.9
hopper-expert	107.1 ± 7.2	110.1 ± 1.4	110.7 ± 1.2	**112.2 ± 0.2**	110.6 ± 1.4	110.4 ± 0.3	109.3 ± 0.7	109.4 ± 1.9
walker2d-random	4.8 ± 1.7	15.9 ± 5.1	17.3 ± 5.8	6.8 ± 1.5	18.7 ± 4.9	17.9 ± 3.5	21.9 ± 0.1	**22.1 ± 0.3**
walker2d-medium	79.7 ± 2.1	82.3 ± 3.2	87.6 ± 1.2	85.2 ± 2.2	85.3 ± 5.5	90.1 ± 0.7	**93.6 ± 0.9**	90.7 ± 0.9
walker2d-medium-replay	52.4 ± 7.3	85.6 ± 7.1	84.8 ± 1.3	84.1 ± 12.7	84.5 ± 9.5	85.2 ± 3.6	87.7 ± 1.7	**89.1 ± 1.5**
walker2d-medium-expert	107.9 ± 0.3	109.4 ± 0.5	113.1 ± 0.5	111.2 ± 0.1	111.7 ± 3.4	110.7 ± 0.4	112.4 ± 0.5	**117.3 ± 1.2**
walker2d-expert	106.3 ± 0.6	108.7 ± 0.3	111.3 ± 1.5	105.7 ± 2.7	112.4 ± 0.9	108.8 ± 0.2	115.1 ± 1.9	**116.8 ± 4.3**
Average	63.5	72.5	75.4	72.0	71.3	78.7	83.7	**84.4**

图2：离线训练结果

成果2：在线微调阶段可以获得更好的最终性能

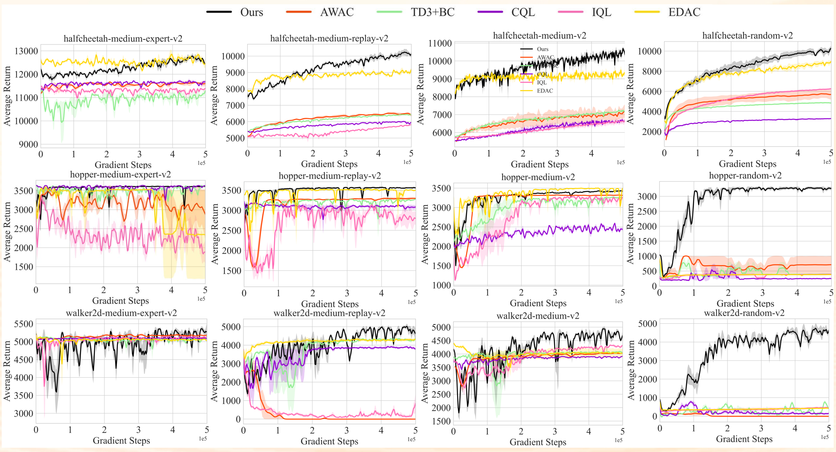

图3：在线微调结果