控制与决策
Control and Decision

基于改进交叉熵的模仿学习鲁棒性增强方法

李晓豪, 郑海斌, 王雪柯, 张京京, 陈晋音, 王巍, 赵文红

引用本文:
李晓豪,郑海斌,王雪柯,张京京,陈晋音,王巍,赵文红. 基于改进交叉熵的模仿学习鲁棒性增强方法[J]. 控制与决策, 2024, 39(3): 768−776.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1181

基于改进交叉熵的模仿学习鲁棒性增强方法

李晓豪[1,2], 郑海斌[1,2], 王雪柯[1,2], 张京京[3], 陈晋音[1,2†], 王 巍[4], 赵文红[5]

(1. 浙江工业大学 网络空间安全研究院，杭州 310023；2. 浙江工业大学 信息工程学院，杭州 310023;
3. 信息安全国家重点实验室，北京 100039；4. 中国电子科技集团公司第三十六研究所，浙江 嘉兴 314001;
5. 嘉兴南湖学院 信息工程学院，浙江 嘉兴 314001)

摘 要: 模仿学习是一种模仿专家示例的学习模式,需要大量数据样本进行监督训练,如果专家示例掺杂恶意样本或探索数据受到噪声干扰,则影响学徒学习并累积学习误差;另一方面,模仿学习使用的深度模型容易受到对抗攻击. 针对模仿学习的模型安全问题,从模型损失以及模型结构两个方面分别进行防御. 在模型损失方面,提出基于改进交叉熵的模仿学习鲁棒性增强方法;在模型结构方面,利用噪声网络模型提高模仿学习的鲁棒性,并结合改进交叉熵提高模型对对抗样本的抵御能力. 使用3种白盒攻击及1种黑盒攻击方法进行防御性能验证,以生成对抗模仿学习为例,通过各种攻击策略验证所提出的鲁棒性增强方法的可行性以及模仿学习的脆弱性,并对模型的鲁棒性增强效果进行评估.

关键词: 模仿学习；鲁棒性增强；改进交叉熵；噪声网络；对抗攻击

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1181

引用格式: 李晓豪,郑海斌,王雪柯,等. 基于改进交叉熵的模仿学习鲁棒性增强方法[J]. 控制与决策, 2024, 39(3): 768-776.

Imitation learning robustness enhancement based on modified cross entropy

LI Xiao-hao[1,2], *ZHENG Hai-bin*[1,2], *WANG Xue-ke*[1,2], *ZHANG Jing-jing*[3], *CHEN Jin-yin*[1,2†], *WANG Wei*[4], *ZHAO Wen-hong*[5]

(1. Institute of Cyberspace Security，Zhejiang University of Technology，Hangzhou 310023，China；2. College of Information Engineering，Zhejiang University of Technology，Hangzhou 310023，China；3. National Key Laboratory of Science and Technology on Information System Security，Beijing 100039，China；4. The 36th Research Institute of China Electronics Technology Group Corporation，Jiaxing 314001，China；5. School of Information Engineering，Jiaxing Nanhu University，Jiaxing 314001，China)

Abstract: Imitation learning is a learning mode characterized by the way of imitating expert examples, which requires many data samples for supervised learning. Once the expert examples are mixed with malicious examples or the exploration data is disturbed, it may affect the students' learning and accumulate learning errors. On the other hand, the deep learning model used by the imitation learning is vulnerable to adversarial attacks. Addressing to the security threat of imitation learning, this paper defends it from two aspects, including model loss and model structure. In terms of model loss, a robust enhancement method for imitation learning based on improved cross-entropy is proposed. In terms of model structure, the existing robust enhancement method for a noise network is applied to verify the robustness enhancement effect. The noise network is also combined with improved cross entropy to improve the model's robustness. Three white box attacks and one black box attack methods in deep learning are applied to imitation learning to verify the defense performance of the proposed method. Specifically, generative adversarial imitation learning (GAIL) is selected as an example. The feasibility of the robustness enhancement method and the fragility of the imitation learning model are verified by various attack strategies, and the robustness enhancement effect of the model is evaluated.

Keywords: imitation learning；robustness enhancement；improved cross entropy；noise network；adversarial attack

收稿日期: 2022-07-04；**录用日期:** 2022-11-10.

基金项目: 国家自然科学基金项目(62072406)；浙江省自然科学基金项目(LY19F020025)；宁波市“科技创新2025”重大专项项目(2018B10063)；科技创新2030——“新一代人工智能”重大项目(2018AAA0100801)；浙江省重点研发计划项目(2021C01117)；浙江省“万人计划”科技创新领军人才项目(2020R52011).

责任编委: 侯忠生.

†通讯作者. E-mail: chenjinyin@zjut.edu.cn.

0 引 言

模仿学习(imitation learning, IL)[1]是指从专家提供的范例中进行学习,已有研究表明模仿学习能较好地求解多步决策问题,在自动驾驶[2-3]、机器人[4]、自然语言处理[5]等领域有众的多应用. 模仿学习方法通过模仿专家演示的样本实现高效策略的学习,实现智能决策. 它无需从学习中获得值函数反馈,其反馈信息来自于专家策略经判别器的对比反馈. 在许多实际问题中,相较于设置合适的值函数反馈,获取专家样本相对容易且代价更小. 模仿学习类似于监督学习,目前应用广泛,其安全性问题必将成为研究重点之一.

模仿学习的安全性研究是其进一步应用的保障,一旦模仿学习模型受到攻击,其学习结果可能产生严重危害. 以实际的自动驾驶场景[3]为例,如果环境路况受到恶意干扰(虚假标识、特制路障等),则目标车辆就可能会受到攻击并执行错误动作,从而越线甚至撞向其他车辆等. 因此,可以通过多种攻击方法挖掘模仿学习模型存在的安全漏洞,然后寻找安全漏洞的解决方案,探究有效的鲁棒增强方法,从而提升模仿学习的安全性.

挖掘模仿学习方法的脆弱性并对其进行鲁棒增强是有效的解决途径之一[6],因此针对模仿学习模型存在的安全漏洞问题,即容易受到恶意样本攻击,本文提出一种基于改进交叉熵的模仿学习鲁棒性增强防御方法,通过对判别器模型损失函数添加惩罚项,增加专家数据与学习者数据的边界距离,从而使得判别器能更好地识别专家数据与非专家数据. 本文将噪声网络[7]应用到模仿学习模型中,并在训练阶段添加模型参数噪声以增强模型鲁棒性. 最后,利用多种攻击方法分别对加固前后的模型进行鲁棒性评估,从而验证本文所提出方法的有效性.

本文的主要贡献包括以下3个方面:

1) 针对IL模型的安全性问题提出基于改进交叉熵的模型鲁棒性增强方法,并将噪声网络迁移到模仿学习模型,验证噪声网络对模仿学习鲁棒性增强效果,同时作为本文的对比方法;

2) 使用多种对抗攻击方法挖掘模仿学习存在的安全漏洞,通过攻击鲁棒性增强前后的模型,验证改进交叉熵的鲁棒性增强方法的可行性;

3) 通过大量实验(实验场景选用Gym[8]平台上的box2d场景)验证了基于改进交叉熵的IL鲁棒增强方法的鲁棒性增强效果.

1 相关工作

本节介绍现有的主流IL模型以及相关的攻击和防御方法.

1.1 模仿学习

模仿学习一般提供人类专家的决策数据,每个决策包含状态和动作序列,将所有“状态-动作”对抽取出来构造新的集合. 模型的训练目标是使模型生成的策略轨迹分布与输入的轨迹分布相匹配. 根据策略优化的方式不同,模仿学习方法可分为:行为克隆(behavioral cloning, BC)[9-11]、逆强化学习(inverse reinforcement learning , IRL)[12-13]、生成对抗模仿学习(generative adversarial imitation learning, GAIL)[6].

行为克隆[9-11]直接从专家的经验数据中学习采样状态下的最优动作,而不构建奖励函数,学习后预测新的状态下对应的最优动作. 但当受到噪声干扰时,执行动作将有微小误差,后续状态也会有微小误差,之后的时序学习将不断累积误差,导致学习误差逐渐加剧. 由于BC没有考虑长远影响,在序贯的决策过程中会将细微的误差逐步放大,即产生级联误差问题.

逆强化学习[12-13]通过给定最优策略或最优行为轨迹,寻找可解释这些策略或行为的奖赏函数. 基于逆强化学习的模仿学习(imitation learning via IRL, IRL-IL)[14],通过强化学习(reinforcement learning, RL)[15]求解最优策略,间接还原专家策略,能进行长远规划,解决BC的级联误差问题,具有更强的泛化性、鲁棒性. 然而,大多数IRL-IL方法的线性奖励函数有一定的局限性[16],而且迭代求解RL子过程需要消耗大量计算资源.

基于IRL-IL,Ho等[6]结合生成对抗网络[17]提出了生成对抗模仿学习,通过专家数据与构建的智能体网络产生的数据进行对比并反向优化模型,使得智能体能够获得与专家类似的数据,从而达到学习的目的,缓解了IRL-IL的训练问题. 然而,GAIL仍面临模态崩塌、环境交互样本利用率低的问题. 因此,Fei等[18]提出了一种基于GAIL的多模态模仿学习算法框架,称为Triple-GAIL. 通过模态选择和行为模仿联合学习,并利用模态选择器增量式生成数据促进模态区分,从而优化了模仿效果. 另外,文献[19]提出了基于多模态假设的改进方法;文献[20]提出了生成模型的改进方法;文献[21]提出了一种简单且易于实现的模仿学习算法(SQIL),可用于高维、连续、动态环境中,能很好地克服行为克隆模仿学习中的分布漂移问题.

针对生成样本利用率低的问题,Baram等[22]提出了基于模型的对抗模仿学习(MAIL)算法,展示了如何使用前向模型使系统完全可微,从而能够利用判别器的梯度训练策略. 此外,文献[22]的方法需要相对较少的环境交互作用,以及较少的可调超参数. 针对样本利用率低的问题,Jeon等[23]提出了基于贝叶斯的改进方案,Blondé等[24]提出了动态模型的改进方案. 除了算法层面的改进,Song等[25]也将GAIL迁移到多智能体场景,提出了MAGAIL算法,解决了多智能体场景下的模仿学习问题.

1.2 防御实验

针对IL的防御方法目前还没有相关工作;而针对DRL的防御方法已有相关工作,一些防御方法用于提高DRL的安全性. 根据防御的入手点不同,可将现有的针对DRL的防御方法分为如下3类.

1) 修改防御模型的输入. Gu等[26]提出了一种对抗A3C学习框架,这种对抗学习框架在学习过程中引入一个敌对智能体,以此模拟环境中可能存在的不稳定因素. 目标智能体通过与该敌对智能体博弈训练,最终达到纳什均衡,从而使A3C模型具有更强的鲁棒性,对环境具有更强的自适应能力. Lin等[27]提出了一种动作条件帧预测模型,通过比较目标策略对预测帧与当前帧的动作分布差异来判断当前帧是否为对抗样本,如果当前帧被判断为对抗样本,则智能体使用预测帧作为输入并执行动作. 这一类防御实现了基于Atari2600博弈环境对对抗样本的检测和防御. 此外,Fischer等[28]提出了一个鲁棒性学生-DQN (robust student-DQN, RS-DQN)的网络架构,该架构允许在线鲁棒性训练与Q网络并行,同时保持竞争性能;将RS-DQN与对抗训练和鲁棒训练相结合,在训练和测试过程中能够抵御攻击.

2) 修改目标函数. Lütjens等[29]提出了一种在线认证的防御机制,智能体在执行过程中能够保证状态动作值的下界,以保证在输入空间可能存在对抗扰动的情况下选择最优动作. Lütjens等[29]的方案没有对智能体进行操作,而Pinto等[30]基于零和博弈思想在环境中引入其他智能体,将建模误差与训练及测试场景下的差异都视为系统中的额外干扰,添加智能体来干扰目标智能体,并将策略的学习建模为零和极大极小值目标函数,目标智能体在学习过程中一边以完成原任务为目标,一边使自己在面对对抗智能体的干扰时变得更加鲁棒.

3) 改变模型的结构. Behzadan等[31]证明了在DQN模型参数中加入噪声并对原始模型进行再训练可以抵抗攻击. 实验中他们使用等价模型方法建立目标网络的副本,以副本为基础制造FGSM对抗扰动. 此外,Havens等[32]提出了一种元学习优势层次框架,通过衡量一定时间内子策略的回报来决定是否继续执行当前子策略. 在此框架的基础上,Lee等[33]进一步探索了将其作为防御框架的可行性,并实现了对DRL的防御.

以上防御方法是基于敌对智能体或等价模型,而模仿学习是基于专家数据的过程学习,也是一种等价模型的思想. 因此,以上防御方法通过加入对抗智能体相当于在没有固定专家数据的探索过程中的博弈,并不完全适用于模仿学习.

2 方 法

2.1 总体框架

深度强化学习离不开模型、数据以及硬件基础,模型网络结构直接影响学习效果,而损失函数用来度量模型,指导模型训练,所以损失函数的设计与模型结构都很重要. 因此,本文从模型损失以及网络结构两个角度提高模型鲁棒性. 首先在模型损失方面,本文提出改进交叉熵惩罚学习者学到的示例与专家示例的边界距离,实现在专家示例的预测置信度变大、损失函数梯度变小的情况下,通过惩罚增大梯度,并通过模型训练提高鲁棒性,同时达到一定防御效果. 其次,在模型结构层面首次将噪声网络[7]迁移到模仿学习中,验证其是否也在模仿学习模型上适用,并作为对比基线方法.

图1是模仿学习鲁棒性增强的整体框图,以GAIL为例,在模型训练过程中,通过改进交叉熵增强模型鲁棒性. 图1中给出了模型训练步骤:首先执行步骤①,采样专家数据;然后执行步骤②,actor网络根据专家状态选择动作α';接着执行步骤③,用discriminator模型判别专家状态动作与学习者状态动作对,并在步骤④反馈更新actor网络. 在改进交叉熵的鲁棒性提升方法中,通过操纵判别器损失函数改进交叉熵,然后在训练中增强模型的鲁棒性;加入改进的交叉熵之后,判别器对专家数据与非专家数据的识别能力会有所增强. 从图1中右侧专家与非专家数据的分类结果可以看出,专家和非专家数据都与决策边界的距离更远了. 本文改进交叉熵对专家数据的惩罚较大,相应地,它距离决策边界也较远,聚合效果较好.

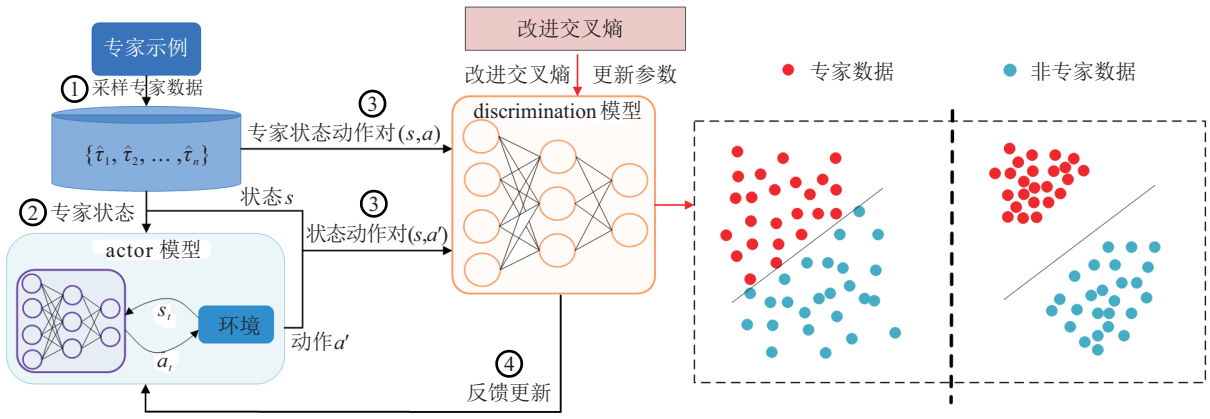

图 1　模型鲁棒性增强整体框图

2.2　网络结构

基于逆强化学习的模仿学习代表性工作 GAIL[6] 的模型主要由两部分组成,包括 actor 网络和 discriminator 判别器网络. 判别器用于区分智能体探索样本与专家样本,进而利用判别器描述的奖惩函数探索环境,并通过强化学习训练智能体策略. 提供专家样本为

$$D = (s_1, \alpha_1), (s_2, \alpha_2), \ldots, (s_T, \alpha_T), \tag{1}$$

其中 s 和 α 分别是状态和动作. 其优化目标可以表示为

$$\min_{\pi} D_{KL}(\rho_\pi(s,\alpha)||\rho_E(s,\alpha)). \tag{2}$$

其中: ρ_π 是学徒示例, ρ_E 是专家示例, s 和 α 分别是状态和动作. 详细参数更新公式如下:

$$\hat{E}_\pi[\bigtriangledown_w \log(D_w(s,\alpha))] + \hat{E}_{\tau E}[\bigtriangledown_w \log(1 - D_w(s,\alpha))]. \tag{3}$$

其中: w 是判别器网络参数; s 和 α 分别是当前状态和动作,而状态是指专家状态; $D(\cdot)$ 是判别器网络的输出. 模型的训练目标是使模型生成的状态-动作轨迹分布与专家轨迹分布相匹配.

2.3　鲁棒性增强

本文从两个角度提出模型鲁棒性增强方法: 在模型损失层面,提出基于改进交叉熵的模型鲁棒性增强方法,通过改进交叉熵实现对专家示例预测的惩罚,干扰训练过程专家与学者的判别边界,从而实现模型鲁棒性增强; 在模型结构层面,将噪声网络应用到模仿学习中以提高模型鲁棒性,并作为对比方法.

2.3.1　改进交叉熵

本文选择基于生成对抗网络的模仿学习 GAIL[7] 模型作为目标模型,它本质上是一个马尔可夫决策过程. 在模仿学习过程中,专家数据是提前存储的,已有专家的示例轨迹为 $\tau = (s_0, \alpha_0, s_1, \alpha_1, \ldots)$, 其中 (s,α) 状态动作对是基于专家的最优策略 π^* 生成的.

在 GAIL 中,学习者通过 actor 网络模型生成无限逼近专家策略的轨迹,同时通过判别器网络判断区分专家数据与学习者数据,经过两个网络模型的不断博弈,最终学习到最优学习者策略. 然而,现有的模仿学习不能有效区分学习者数据与专家数据. 对此,本文将改进交叉熵添加到判别器网络的训练过程中,将专家数据作为类别 1,即真实数据标签; 学习者数据作为类别 0,即虚假数据标签. 随后,对判别器网络提出新的交叉熵函数,以使判别器更好地区分专家数据和学习者数据. 目的是在训练过程中,扩大专家数据与分类边界的距离,并可以将专家数据以较高的置信度判别为专家数据; 而学习者数据将更偏向虚假数据一方,通过交叉熵惩罚实现. 改进的交叉熵损失函数定义如下:

$$L_i(w_i) = E_{\tau i}[\log(D_w(s,\alpha))] + E_{\tau E}[(y_i^j + (-1)^{y_i^{(j)}+1}\hat{y}_i^j)\log(1 - D_w(s,\alpha))]. \tag{4}$$

其中: s 和 α 分别是状态和动作; D 表示判别器网路; ω 表示判别器网络参数; y 表示当前数据标签; $y_i^{(j)}$ 表示真实标签经过 one-hot 编码后,第 i 个样本的第 j 维的值; $\hat{y}_i^j$ 表示预测输出的第 i 个样本的第 j 维的值.

训练过程中,每次判断专家数据与学习者数据之间的置信度差异时,会通过改进交叉熵自适应增大二者之间的距离,进而使判别器网络更好地区分专家与学者,同时 actor 模型又在不断逼近专家轨迹,由此不断博弈,提高了判别器的判别能力,同时也增强了 actor 网络的性能.

2.3.2　噪声网络

噪声网络是由 Fortunato 等[7] 提出的,用于增强强化学习的鲁棒性,而噪声网络在模仿学习上是否有效还有待验证. 本文利用 Fortunato 等[7] 提出的噪声网络策略对模仿学习网络添加参数噪声,其权重和偏

置被噪声函数所干扰,用梯度下降法对这些参数进行调整. $y = f_{w^*}(s,\alpha)$ 是由噪声参数 w^* 的向量参数化的神经网络,同时将 w^* 表示为 $w^* = \mu + \sigma \otimes \varepsilon$. 其中: $\zeta = (\mu, \Sigma)$ 是一组可学习的参数向量, μ 是零均值噪声的向量, $\otimes$ 表示按元素相乘. 神经网络的损失函数由包含噪声的期望所得到,即 $\bar{L}(\zeta) = E[L(w^*)]$. 最后对参数的集合进行优化. 参数的更新公式为

$$\mathrm{d}w^* \leftarrow \mathrm{d}w^* + \bigtriangledown_w \bar{L}(\xi). \tag{5}$$

其中: $w^* \overset{\text{def}}{=} \mu^w + \sigma^w \otimes \varepsilon^w$, 用来替换 w; 参数 $\mu \in \mathbb{R}^{q\times p}$, $\sigma \in \mathbb{R}^{q\times p}$ 是噪声系数; $\varepsilon \in \mathbb{R}^{q\times p}$ 是随机噪声变量.

3 实验与结果

利用本文所提出的模型鲁棒性增强方法构建鲁棒模型,并使用多种攻击方法进行鲁棒性验证,同时对实验结果进行分析.

3.1 关键问题

本文围绕以下几个问题展开分析:

问题 1 模仿学习是否容易被对抗噪声攻击?

问题 2 改进交叉熵前后模型训练效果是否受影响?

问题 3 改进交叉熵能否增强模仿学习鲁棒性?

3.2 度量标准

在模仿学习中,模型鲁棒性判别主要根据模型训练后的长期累积奖励 $R_2 = \sum_{k=0}^{\infty} \gamma_k r_{t+k}$. 其中: γ 是折扣因子, r 是每步奖励值, t 是当前时刻, k 是当前时刻之后的未来期望步数.

扰动度量: 常用扰动计算方法有 0 范数、2 范数以及 ∞ 范数. 其中: 0 范数用来计算像素点改变的数量, 2 范数用来计算扰动像素点绝对值平方和的均方根, ∞ 范数用来计算像素的最大扰动量. 本文使用 2 范数度量扰动大小.

3.3 攻击防御方法

防御方法: 目前还没有针对模仿学习的防御方法,本文提出一种基于改进交叉熵的模型鲁棒性增强防御方法,同时也将噪声网络应用到模仿学习中,作为对比方法.

攻击方法: 在测试阶段进行对抗攻击以验证模型的鲁棒性. 为了验证本文方法对黑盒攻击与白盒攻击都有效,分别进行黑盒攻击和白盒攻击. 白盒攻击方法包括 FGSM[34]、MIFGSM[35]、PGD[36], 黑盒攻击采用简单且易于实现的随机噪声 RandomNoise[37] 和策略诱导攻击方法 (PIA)[38]. 初步验证面向黑盒攻击的防御能力.

3.4 实验场景

在 Gym[8] 游戏的 box2d 场景上测试本文提出的方法. 采用 box2d 里的 BipedalWalkerHardcorev2 连续动作双足机器人作为验证场景, 该游戏场景动作是 8 维,输入状态空间是 24 维. 控制两足机器人行走,机器人每走一步如果不摔倒则加分,摔倒则游戏结束,重新下一轮.

3.5 实验结果分析

本节首先挖掘模仿学习模型的漏洞,表明有安全问题的存在,整个实验部分都以 GAIL 模型为例进行验证; 其次进行模型鲁棒性训练,给出模型训练结果; 然后验证模型鲁棒性,分别验证面向白盒攻击以及黑盒攻击的模型鲁棒性; 最后进行消融实验,以验证结合了噪声网络的改进交叉熵防御方法与不加噪声网络的差异,并验证 GAIL 模型中不同网络加噪声的影响.

3.5.1 模型漏洞挖掘 (问题 1)

针对问题 1, 本节通过几种攻击方法挖掘模仿学习模型存在的安全漏洞,如果模型受到攻击后奖励值下降很多,则说明模型存在安全问题. 实验结果如图 2 所示,显示了 GAIL 模型下的模型奖励变化情况. 横坐标是扰动值大小,纵坐标是奖励值,不同颜色折线代表不同类型的攻击方法,横坐标值为 0 时模型没有受到攻击. 从图 2 中可以观察得到: 当 GAIL 模型受到攻击时奖励值有明显降低; 当扰动值大于 0.26 时, GAIL 模型将无法自愈,两足机器人很快摔倒,无法正常行走. 由此可以回答问题 1, 即 IL 模型容易被攻击者恶意干扰,存在安全漏洞. 模仿学习存在安全漏洞的原因值得关注: 首先模仿学习是建立在一致专家数据的基础上进行的学习, 已有专家数据是一个连续存储的状态动作对,但最终学习是要区分专家策略与非专家策略,同时应保证与专家策略无限接近. 另外,模仿学习模型是基于深度模型搭建的,类似于一种监督学习的过程,深度学习的安全问题[39] 已

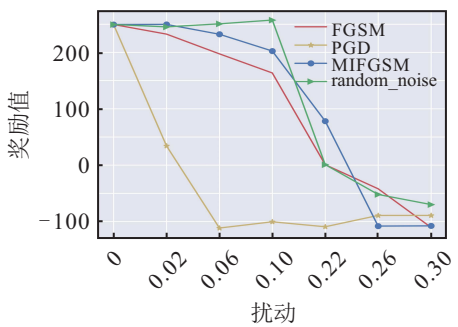

图 2 不同攻击方法下奖励值

有很多文献提及,故基于此模仿学习必然会存在安全问题. 图2的实验结果也进一步佐证了模仿学习模型的安全漏洞的存在.

3.5.2　模型鲁棒性增强训练(问题2)

本节使用两种防御方法来提高IL模型的受攻击能力,增强模型鲁棒性. 两种鲁棒性增强防御方法包括改进交叉熵和噪声网络. 其中:改进交叉熵是在模型损失方面进行改进;而噪声网络是应用现有针对强化学习模型的防御方法,将其迁移到IL模型,对模型结构层面进行改进. 下面验证本文提出的鲁棒性增强方法不影响原模型训练效果(问题2).

实验结果如图3所示,横坐标是游戏回合数,纵坐标是对应累积奖励值. 其中:PCE表示改进交叉熵方法,CE表示原交叉熵,CE_A&D表示对GAIL模型的actor动作网络和discriminator判别器网络加入高斯噪声. 从图3中可以看出:使用改进交叉熵训练模型后,模型的训练效果与原模型相当,而且改进交叉熵提高了模型的训练效果;而噪声网络降低了模型整体效果,但其抵御对抗攻击的能力还有待验证. 因此,本文提出的鲁棒性增强方法不仅不会影响原模型训练效果,还能提升原模型效果,训练结束后模型累积奖励值增大.

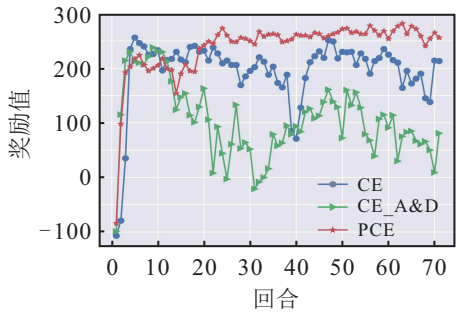

图3　改进交叉熵前后的测试奖励值

改进交叉熵之所以不影响模型整体效果,可能的原因是:改进交叉熵只是对专家与非专家策略的评判进行惩罚,拉大了专家策略与非专家策略的距离,能增强判别器的判别能力,不会破坏智能体的学习能力;而噪声网络可能会加重智能体学习负担,导致策略学习过程受到影响. 因此,图3中噪声网络模型学习效果低于正常模型,而改进交叉熵奖励值较高,学习效果较好.

3.5.3　面向白盒攻击的模型鲁棒性增强验证(问题3)

针对问题3,面向白盒攻击进行模型鲁棒性验证,本文在使用鲁棒性增强方法训练模型后进行对抗攻击,以验证本文提出的鲁棒性增强方法是否有效.

实验结果如图4所示,纵坐标表示游戏累积奖励值,不同颜色折线表示不同模型设置. 其中:PCE表示改进交叉熵的模型,CE表示原未改进交叉熵的模型,CE_A&D表示同时对GAIL模型的actor动作网络和discriminator判别器网络加入高斯噪声.

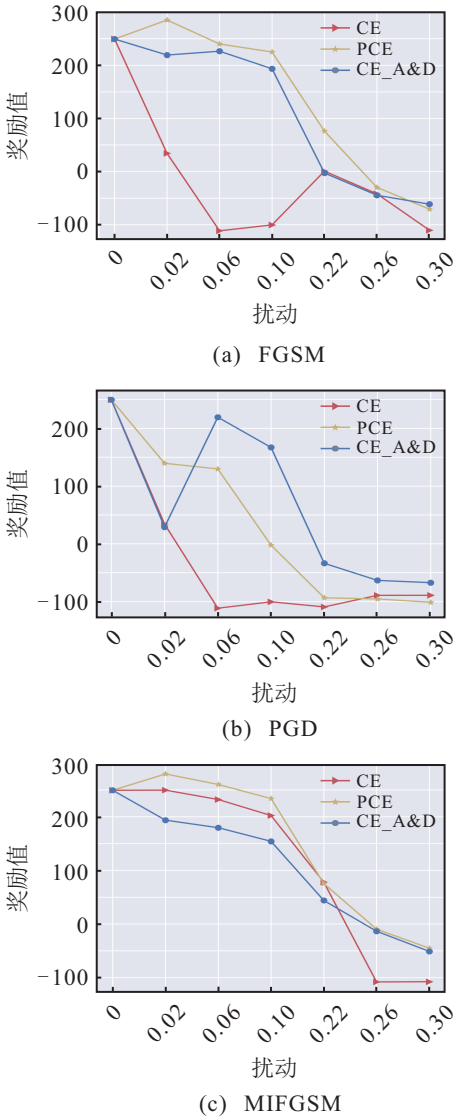

图4　不同攻击下的训练奖励值

从图4中可以看出:本文所提出的改进交叉熵在各种攻击方法下都具有一定的防御能力;在FGSM攻击方法下扰动小于0.26之前,改进交叉熵的防御效果都比噪声网络好,由于FGSM攻击方法下根据模型梯度生成的噪声针对性较强,模型没有在随机噪声下抵御能力强. 鲁棒模型抵御能力有一定上界,当扰动大于0.26时,虽然攻击后奖励值都比原模型高,但奖励已经大大降低,游戏进程受到很大影响. 在PGD以及MIFGSM攻击方法下也有类似规律. 而在PGD攻击方法下,噪声网络在一定范围扰动下防御能力更强,这是因为在扰动迭代过程中,噪声网络的参数噪声对扰动有一定的抵消恢复能力,但当扰动继续增大时将无法恢复. 总之,改进交叉熵对模型鲁棒性有一定提

升,从而回答了问题3,表明本文提出的鲁棒性增强方法是有效的.

3.5.4 面向黑盒攻击的模型鲁棒性增强验证(问题3)

同样针对问题3,面向黑盒攻击进行模型鲁棒性验证,本文在使用鲁棒性增强方法训练模型后进行对抗攻击,以验证本文提出的鲁棒性增强方法是否有效. 由于目前还没有面向模仿学习的黑盒攻击方法,本文选择易于实现且比较成熟的随机噪声(random_noise)攻击方法和策略诱导攻击方法(PIA).

实验结果如表1所示,表1中给出了不同鲁棒性增强方法以及原模型面向黑盒攻击的奖励值. 其中: PCE表示改进交叉熵的模型,CE表示原未改进交叉熵的模型,CE_A&D表示同时对GAIL模型的actor动作网络和discriminator判别器网络加入高斯噪声. 表1中显示,不同扰动值大小下改进交叉熵的模型奖励值都最高,鲁棒性较强,除了0.10以及0.12扰动值对应的原模型略高一点,但总体上改进交叉熵增强了模型的鲁棒性以及防御能力.

表 1 面向黑盒攻击的模仿学习模型鲁棒性验证结果

攻击方法	模型	扰动值大小									
		0.04	0.06	0.08	0.10	0.12	0.22	0.24	0.26	0.28	0.30
		奖励值									
random_noise	CE	253.42	250.53	251.21	**257.24**	**250.28**	0.86	−25.53	−52.01	−55.39	−69.91
	CE_A&D	24.64	29.80	27.19	28.64	32.03	−38.15	−34.98	−46.51	−32.68	−43.33
	PCE	**258.15**	**260.90**	**252.25**	241.46	248.39	**196.45**	**203.29**	**204.30**	**216.74**	**195.77**
PIA	CE	246.23	243.19	244.62	**253.81**	**248.58**	1.26	−22.21	−57.82	−59.92	−72.29
	CE_A&D	20.32	24.14	22.42	21.40	30.83	−39.34	−32.40	−44.63	−3.67	−53.28
	PCE	**247.55**	**253.91**	**249.98**	229.71	241.09	**187.25**	**210.82**	**193.27**	**210.03**	**202.41**

表1中作为对比方法的噪声网络性能最差,这与模型训练过程中添加的噪声有关,一定程度上影响了智能体的学习能力. 而交叉熵改进方法增强了判别器的判别能力,进而优化了动作actor网络,促进智能体模仿专家策略的效果,同时增强了自身的抗攻击能力.

3.5.5 消融实验

本文还对交叉熵和噪声网络进行了消融实验. 在交叉熵的基础上再加入高斯噪声进行消融实验,也对actor动作网络和discriminator判别器网络分别加入高斯噪声进行消融实验,实验结果如图5所示. 其中: PCE_A表示在改进交叉熵的基础上对actor网络加入高斯噪声,PCE_D表示在改进交叉熵的基础上对discriminator网络加入高斯噪声; PCE_A&D表示在改进交叉熵的基础上对actor和discriminator网络同时加入高斯噪声,CE_A&D表示不改进交叉熵同时对actor和discriminator网络加入高斯噪声.

在图5中,本文给出了各种改进模型在不同攻击方法下扰动变化所对应的奖励值,每一种攻击方法下都对应5种模型. 横坐标是扰动值,纵坐标是对应的奖励值,不同风格的点线对应不同模型,横坐标扰动值为0时表示没有对模型进行攻击. 从图5中可以看出,每种组合下奖励值最大的都是改进后的模型,在FGSM攻击方法下,PCE和PCE_A的模型鲁棒性较强,说明改进交叉熵不仅本身具有防御效果,而且

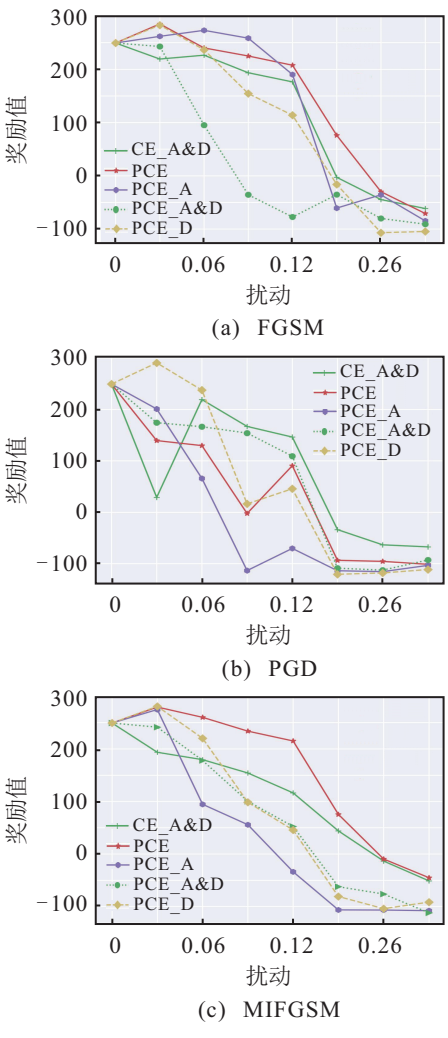

图 5 不同模型及攻击方法下的奖励值

对actor加入噪声后的模型也有一定的增强作用. 在PGD以及MIFGSM攻击方法下,只对actor模型加入噪声似乎对GAIL模型鲁棒性提升效果更好,这也说明在GAIL模型中actor网络对整体策略的学习起到关键作用. 另外,在PGD攻击方法下PCE的防御效果比不上CE_A&D,但在MIFGSM攻击方法下,二者效果相当. 总体上,本文提出的改进交叉熵具有一定的防御效果.

图5中PGD方法攻击下CE_A&D防御效果较好些,但随扰动的增大,本文方法达到与其相当的效果. 其他攻击方法下,本文提出的模型鲁棒性增强方法都有较好的效果,而且加入噪声网络之后使用改进交叉熵对噪声网络防御有一定的增强作用. 如图5(a)和图5(b)中PCE_A&D以及PCE_A对应模型的奖励值都较高,远远超过CE_A&D. 本文推测,只加入噪声网络在一定程度上会影响智能体的学习,但在噪声网络的基础上再改进交叉熵,可拉大专家策略与非专家策略的距离,提升模型的性能,有利于智能体对干扰环境的适应能力,尤其是PCE_A学习效果更好. 但也有一个问题,改进交叉熵同时加入噪声并不能保证面对所有攻击方法的防御效果都很好,而只有改进交叉熵整体上防御效果都很好,增强了模型鲁棒性.

4 结 论

本文研究了模仿学习的安全漏洞问题. 通过各种攻击方法对模仿学习进行安全漏洞挖掘,验证了模仿学习的脆弱性,同时本文也提出了面向模仿学习的防御方法,增强了模仿学习模型的鲁棒性. 本文在模型损失和模型结构两个方面进行鲁棒性增强,针对模型损失提出了改进交叉熵的模仿学习模型鲁棒性增强防御方法,同时也针对模型结构将噪声网络应用于模仿学习模型以实现鲁棒性增强,将噪声网络视为对比方法,本文通过大量实验表明这两种防御方法都是有效的,但本文方法的防御效果更好. 实验中模仿学习模型采用GAIL模型来验证本文方法的有效性,不仅验证了面向白盒攻击的有效性,也验证了面向黑盒攻击的有效性. 针对噪声网络和改进交叉熵,本文也进行了消融实验,本文对改进交叉熵的网络模型进行了actor和判别器网络添加噪声的消融实验,验证了模型鲁棒性的提升效果.

虽然大量实验验证了本文所提出的方法是有效的,但本文工作仍存在一些缺陷. 其中实验场景较少,同时也没有验证是否适用于其他模仿学习模型,虽然实验表明,改进交叉熵确实提高了模型鲁棒性,但性能提升并不明显. 未来工作将进一步研究模仿学习模型的安全问题,进一步验证多种场景以及多种算法模型的适用性,而且本文防御效果还有待进一步提升.

参考文献(References)

[1] Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[2] Ross S, Gordon G, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning[C]. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, 2011: 627-635.

[3] Kuefler A, Morton J, Wheeler T, et al. Imitating driver behavior with generative adversarial networks[C]. 2017 IEEE Intelligent Vehicles Symposium. Los Angeles, 2017: 204-211.

[4] Giusti A, Guzzi J, Cirean D C, et al. A machine learning approach to visual perception of forest trails for mobile robots[J]. IEEE Robotics and Automation Letters, 2016, 1(2): 661-667.

[5] Eysenbach B, Gupta A, Ibarz J, et al. Diversity is all you need: Learning skills without a reward function[J/OL]. 2018, arXiv: 1802.06070.

[6] Ho J, Ermon S. Generative adversarial imitation learning[C]. Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016. Barcelona, 2016, 29: 4565-4573.

[7] Fortunato M, Azar M G, Piot B, et al. Noisy networks for exploration[J/OL]. 2017, arXiv: 1706.10295.

[8] Brockman G, Cheung V, Pettersson L, et al. OpenAI gym[J/OL]. 2016, arXiv: 1606.01540.

[9] Pomerleau D A. Alvinn: An autonomous land vehicle in a neural network[J]. Advances in Neural Information Processing Systems, 1988, 1: 305-313.

[10] Torabi F, Warnell G, Stone P. Behavioral cloning from observation[J/OL]. 2018, arXiv: 1805.01954.

[11] Ng A Y, Russell S J. Algorithms for inverse reinforcement learning[C]. Proceedings of the 17th International Conference on Machine Learning (ICML). Stanford, 2000, 1: 663-670.

[12] Arora S, Doshi P. A survey of inverse reinforcement learning: Challenges, methods and progress[J/OL]. 2018, arXiv: 1806.06877.

[13] Abbeel P, Ng A Y. Apprenticeship learning via inverse reinforcement learning[C]. Proceedings of the 21st International Conference on Machine Learning. Banff, 2004: 1-8.

[14] Wang R H, Ciliberto C, Amadori P, et al. Random expert distillation: Imitation learning via expert policy support estimation[J/OL]. 2019, arXiv: 1905.06750.

[15] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari

with deep reinforcement learning[J/OL]. 2013, arXiv: 1312.5602.

[16] Levine S, Popovic Z, Koltun V. Nonlinear inverse reinforcement learning with gaussian processes[J]. Advances in neural information processing systems, 2011, 24: 19-27.

[17] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J/OL]. 2014, arXiv: 1406.2661.

[18] Fei C, Wang B, Zhuang Y Z, et al. Triple-GAIL: A multi-modal imitation learning framework with generative adversarial nets[J/OL]. 2020, arXiv: 2005.10622.

[19] Lin J H, Zhang Z Z. ACGAIL: Imitation learning about multiple intentions with auxiliary classifier GANs[C]. Pacific Rim International Conference on Artificial Intelligence. Cham: Springer, 2018: 321-334.

[20] Wang Z Y, Merel J, Reed S, et al. Robust imitation of diverse behaviors[J/OL]. 2017, arXiv: 1707.02747.

[21] Reddy S, Dragan A D, Levine S. Sqil: Imitation learning via reinforcement learning with sparse rewards[C]. The 8th International Conference on Learning Representation (ICLR). Addis Ababa, 2020: 1-14.

[22] Baram N, Anschel O, Mannor S. Model-based adversarial imitation learning[J/OL]. 2016, arXiv: 1612.02179.

[23] Jeon W, Seo S, Kim K E. A Bayesian approach to generative adversarial imitation learning[C]. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018. Montreal, 2018: 7440-7450.

[24] Blondé L, Kalousis A. Sample-efficient imitation learning via generative adversarial nets[J/OL]. 2018, arXiv: 1809.02064.

[25] Song J M, Ren H Y, Sadigh D, et al. Multi-agent generative adversarial imitation learning[J/OL]. 2018, arXiv: 1807.09936.

[26] Gu Z Y, Jia Z Z, Choset H. Adversary A3C for robust reinforcement learning[J/OL]. 2019, arXiv: 1912.00330.

[27] Lin Y C, Liu M Y, Sun M, et al. Detecting adversarial attacks on neural network policies with visual foresight[J/OL]. 2017, arXiv: 1710.00814.

[28] Fischer M, Mirman M, Stalder S, et al. Online robustness training for deep reinforcement learning[J/OL]. 2019, arXiv: 1911.00887.

[29] Lütjens B, Everett M, How J P. Certified adversarial robustness for deep reinforcement learning[J/OL]. 2019, arXiv: 1910.12908.

[30] Pinto L, Davidson J, Sukthankar R, et al. Robust adversarial reinforcement learning[C]. Proceedings of the 34th International Conference on Machine Learning-Volume 70. Sydney, 2017: 2817-2826.

[31] Behzadan V, Munir A. Mitigation of policy manipulation attacks on deep Q-networks with parameter-space noise[C]. International Conference on Computer Safety, Reliability, and Security. Cham: Springer, 2018: 406-417.

[32] Havens A J, Jiang Z H, Sarkar S. Online robust policy learning in the presence of unknown adversaries[J/OL]. 2018, arXiv: 1807.06064.

[33] Lee X Y, Havens A, Chowdhary G, et al. Learning to cope with adversarial attacks[J/OL]. 2019, arXiv: 1906.12061.

[34] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J/OL]. 2014, arXiv: 1412.6572.

[35] Dong Y P, Liao F Z, Pang T Y, et al. Boosting adversarial attacks with momentum[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 9185-9193.

[36] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J/OL]. 2017, arXiv: 1706.06083.

[37] Fawzi A, Moosavi-Dezfooli S M, Frossard P. Robustness of classifiers: From adversarial to random noise[J/OL]. 2016, arXiv: 1608.08967.

[38] Behzadan V, Munir A. Vulnerability of deep reinforcement learning to policy induction attacks[C]. International Conference on Machine Learning and Data Mining in Pattern Recognition. Cham: Springer, 2017: 262-275.

[39] Huang Y J, Hu H, Chen C Y. Robustness of on-device models: Adversarial attack to deep learning models on android apps[C]. 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice. Madrid, 2021: 101-110.

作者简介

李晓豪 (1999−), 男, 硕士生, 从事人工智能、数据挖掘等研究, E-mail: 2112103051@zjut.edu.cn;

郑海斌 (1995−), 男, 讲师, 博士, 从事人工智能、机器学习等研究, E-mail: haibinzheng320@gmail.com;

王雪柯 (1995−), 女, 初级工程师, 硕士, 从事人工智能、强化学习等研究, E-mail: 3049128970@qq.com;

张京京 (1988−), 男, 高级工程师, 博士, 从事信息系统安全的研究, E-mail: lg02103@163.com;

陈晋音 (1982−), 女, 教授, 博士, 从事人工智能的研究, E-mail: chenjinyin@zjut.edu.cn;

王巍 (1980−), 男, 高级工程师, 博士, 从事网络安全的研究, E-mail: wwlofty@gmail.com;

赵文红 (1981−), 女, 讲师, 硕士, 从事优化计算的研究, E-mail: wwzwh@sohu.com.