

基于模仿学习的深度强化学习训练数据推断攻击

陈晋音^{1,2}, 瞿康赞¹, 郑海斌^{1,2}

(1. 浙江工业大学网络空间安全研究院, 浙江 杭州 310023; 2. 浙江工业大学信息工程学院, 浙江 杭州 310023)

摘要: 深度强化学习凭借其卓越的性能优势, 在PC游戏、机器人控制、自动驾驶等领域获得了广泛应用。该技术通常依赖大量的训练数据, 通过不断试错提升智能体的决策能力。已有研究指出, 深度强化学习存在训练数据泄露的隐私风险, 但相关攻击方法通常对过拟合的智能体更加有效。为此, 针对深度强化学习训练得到且具有泛化能力的智能体, 提出一种基于模仿学习的目标场景隐私攻击方法 (trajectory informed privacy attack strategy, TIPAS)。该方法通过分析智能体的行为轨迹, 采用基于行为克隆的模仿学习方法来构建影子智能体模型, 并利用影子智能体在多个场景中生成行为策略特征, 结合集成学习训练攻击模型, 实现在黑盒访问训练策略场景下精确推断出目标智能体的隐私信息。最后, 分别在4种网络拓扑场景和4种算法模型上进行了实验, 并以基于奖励均值与方差特征的支持向量机方法, 以及未使用集成学习的单一多层感知机模型作为对比基准, 通过推断准确率的指标对比验证了所提方法在攻击性能上的优越性。实验证明, 所提方法在Gym平台的4种经典游戏场景中利用差分隐私和正则化作为隐私保护方法的防御条件, 均具有优秀的攻击性能。

关键词: 深度强化学习; 模仿学习; 成员推断攻击; 隐私窃取攻击

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025060

Membership inference attack based on imitation learning for deep reinforcement learning

CHEN Jinyin^{1,2}, QU Kangyun¹, ZHENG Haibin^{1,2}

1. Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China
2. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract: Deep reinforcement learning (DRL), owing to its remarkable performance advantages, had been widely applied in fields such as game intelligence, robotic control, and autonomous driving. This technology typically relied on large amounts of training data, and the agent's decision-making ability was improved through continuous

收稿日期: 2025-05-13; 修回日期: 2025-08-16

基金项目: 国家自然科学基金 (62072406, 62406286); 浙江省自然科学基金 (LDQ23F020001); 浙江省重点研发计划 (2022C01018); 国家重点研发计划 (2018AAA0100801)

Foundation Items: The National Natural Science Foundation of China (62072406, 62406286), The Natural Science Foundation of Zhejiang Province (LDQ23F020001), Zhejiang Provincial Key R&D Program (2022C01018), The National Key R&D Program of China (2018AAA0100801)

引用格式: 陈晋音, 瞿康赞, 郑海斌. 基于模仿学习的深度强化学习训练数据推断攻击[J]. 网络与信息安全学报, 2025, 11(5): 50-67.

Citation Format: CHEN J Y, QU K Y, ZHENG H B. Membership inference attack based on imitation learning for deep reinforcement learning[J]. Chinese Journal of Network and Information Security, 2025, 11(5): 50-67.

trial and error. Previous studies had indicated that DRL faced privacy risks due to potential training data leakage; however, existing attack methods were generally found to be more effective against overfitted agents. To address this limitation, a trajectory-informed privacy attack strategy (TIPAS) based on imitation learning was proposed for DRL-trained agents with strong generalization ability. In this method, the agent's behavioral trajectories were analyzed, and a shadow agent model was constructed using behavior cloning. The behavioral policy features generated by the shadow agent in multiple environments were then integrated through ensemble learning to train an attack model, which enabled accurate inference of private information from the target agent under black-box access to the training policy. Experiments were conducted across four network topology scenarios and four algorithmic models. The proposed method was compared against two baselines: a support vector machine utilizing reward mean and variance features, and a single multilayer perceptron model without ensemble learning. The comparison of inference accuracy demonstrated the superiority of the proposed approach in attack performance. Experimental results show that excellent attack performance is achieved by the proposed method across four classic game environments on the Gym platform, even under defensive conditions where differential privacy and regularization techniques are employed.

Keywords: deep reinforcement learning, imitation learning, membership inference attack, privacy stealing attack

0 引言

近年来,深度强化学习在自主决策的各个领域取得了巨大成功,包括PC游戏^[1]、机器人控制任务^[2-3]和自动驾驶^[4-6]等。与传统监督学习直接从标注数据中学习静态映射关系不同,深度强化学习智能体通过与环境的动态交互和试错机制,持续优化其在状态空间与动作空间中的策略。智能体在不同环境中训练得到的策略差异,体现在其长期累积的状态—动作轨迹中,这为攻击者通过轨迹分析推断训练环境信息提供了可能。

随着深度强化学习在实际应用中的广泛推广,其训练过程可能涉及大量个人隐私数据^[7]。例如,在自动驾驶场景中,深度强化学习模型的训练往往依赖于包含GPS轨迹、驾驶行为等敏感数据。然而,目前针对深度强化学习隐私保护的研究尚不充分,特别是对具备良好泛化能力的深度强化学习模型的隐私风险探讨尤为欠缺,这给敏感数据的安全带来了潜在威胁。因此,深入探究深度强化学习模型在面对隐私窃取攻击时的脆弱性,具有重要的现实意义。

已有研究显示,深度强化学习策略的训练结果对超参数设置高度敏感,这些参数微调就可能导致不同结果^[8]。即便是当前较先进的深度强化学习算法,在任务间进行泛化仍存在困难。尽管训练后的智能体能够完成复杂任务,却难以将所

学经验有效迁移至新环境^[9],并常常出现对原始训练环境的过拟合。这些特性表明,深度强化学习模型可能隐式地记忆了训练环境的信息,因而容易遭受旨在窃取训练环境隐私信息的攻击。直观地看,具备良好跨环境泛化能力的深度强化学习模型,可能同时对隐私泄露攻击表现出更强的鲁棒性^[10]。

近年来,已有多种针对监督学习模型的隐私攻击方法被提出,如成员推断攻击与训练数据重构攻击。这类方法在图像识别、文本分类等典型监督学习任务中已被验证具有显著效果。然而,它们在深度强化学习任务中的适用性仍缺乏系统探索。深度强化学习依赖智能体与环境的持续交互,其策略更新与行为轨迹具有高度的动态性与不确定性,这一特点使监督学习中成熟的攻击范式难以直接迁移。以成员推断攻击为例,其有效性往往依赖于模型对训练数据的过拟合;而在深度强化学习中,为提高智能体在复杂环境中的泛化能力,研究者通常会主动采取措施抑制过拟合,从而减少此类攻击的可行性。类似地,训练数据重构方法在深度强化学习中也面临挑战。状态—动作空间的连续性、奖励信号的稀疏性等问题,常导致重构过程难以收敛,极易陷入不稳定状态,难以生成有效的结果。因此,如何设计一种既能适配深度强化学习特性,又能有效作用于高泛化能力智能体的隐私攻击方法,仍是该领域亟待解决的关键问题。

与针对过拟合深度强化学习模型场景不同,面向具有泛化能力的深度强化学习模型的隐私推断面临更加复杂的挑战与限制。首先,具有泛化能力的模型在多个不同的环境中均表现出较高的性能,这意味着隐私推断必须考虑其在各种环境中的行为差异。这些差异可能掩盖特定训练环境所遗留的特征,从而增加推断难度。其次,现有的隐私推断方法多依赖于环境奖励的统计特征(如均值与方差)来推测训练环境^[10],尽管这类方法在一定程度上能够揭示智能体训练时的环境信息,但在某些特定场景中,不同智能体的奖励水平可能相当接近,导致传统隐私推断方法难以有效区分,进而限制其推断能力。此外,当前隐私推断研究主要集中于过拟合场景下的深度强化学习模型,关注的是智能体在特定环境中的表现;而对于具备泛化能力的智能体,其隐私推断问题尚未深入研究。

针对上述问题,本文旨在探究具有泛化能力的深度强化学习模型所面临的训练数据泄露风险。以自动驾驶网络为例,攻击者在路由优化场景下获得一组候选的智能体配置及一种未知配置下的训练策略,其目标是推断出该策略实际是在哪一种配置下训练所得。在实际生活中通常存在多种网络配置方案,攻击者通过对比分析候选配置与目标策略的行为特征,有可能推测出实际部署中所使用的动态配置。一旦成功识别出训练配置,攻击者不仅能够揭示网络的运行机制,还可能进一步推断出网络架构、策略逻辑乃至数据流向等敏感信息。具体而言,若攻击者识别出某一网络配置所对应的优化策略,便可推断出该网络中的优先级规则、流量模式或路由路径,从而进一步揭示网络的管理策略及其潜在弱点。此类攻击不仅威胁网络运行安全,还可能泄露系统架构的关键细节,甚至暴露网络的物理部署位置或与特定供应商相关的技术信息,对系统整体安全构成严重威胁。

为主动识别具备泛化能力的深度强化学习模型所面临的训练数据泄露风险,本文提出一种基于模仿学习的目标场景隐私攻击方法(trajec-tory informed privacy attack strategy, TIPAS)。该方法通过分析智能体在执行任务过程中所产生的行

为轨迹,构建多个影子智能体,提取其在各候选场景下的策略行为特征,并借助集成学习技术训练攻击分类模型,实现在给定黑盒访问训练策略的条件下准确推断出训练环境转换动态等敏感特征。本文的主要贡献如下。

1) 首次针对具备泛化能力的深度强化学习模型的训练数据进行推断攻击。通过对训练策略与环境状态进行逆向分析,揭示了攻击者仅通过观测智能体行为即可推断出敏感信息的潜在威胁。

2) 针对传统基于奖励统计特征(如均值与方差)的隐私攻击方法在环境奖励水平相近时性能受限的问题,提出一种基于模仿学习的目标轨迹隐私推断攻击方法。通过构建影子智能体模拟目标策略并在不同环境中采集其行为轨迹,结合集成学习训练分类模型,显著提升了对训练场景的识别能力,并在具备隐私防御机制的环境中仍保持良好性能。

3) 所提方法在4种不同的网络拓扑结构中对4种深度强化学习模型进行了验证,并进一步在Gym游戏环境中进行测试。实验结果表明,本文方法在推断准确率等关键指标上均优于基线算法,表现出更优的隐私推断效果与适用性。

1 相关工作

为深入理解本文研究的问题与后续提出的方法,本节将系统介绍深度强化学习、模仿学习及相关隐私窃取攻击的关键背景知识。

1.1 深度强化学习

深度强化学习是一种结合深度学习与强化学习的先进机器学习方法。它利用深度神经网络处理复杂的输入数据(如图像或视频),使得智能体能够在高维环境中学习决策。在这一过程中,智能体在不同状态下执行动作以获得奖励,最终目标是实现长期累积奖励的最大化。深度学习凭借其强大的特征提取能力,使智能体能够直接从原始感知数据中学习,而无须依赖人工设计特征。因此,深度强化学习特别适用于传统强化学习方法难以应对的复杂高维问题(如视频游戏、自动驾驶和机器人控制等)。在深度强化学习中,智能体的行为由策略函数 $\pi(a|s)$ 决定,即在状态 s 下选择动作 a 的概率分布。策略优化的目标

是最大化智能体在整个交互过程中所获得的预期累积奖励，其表达式为：

$$R = \sum_{t=0}^T \gamma^t r_t \quad (1)$$

其中， γ 为折扣因子，用于平衡短期与长期收益； r_t 为时间步 t 时刻智能体获得的即时奖励。

在自动驾驶领域，深度强化学习的应用尤为重要。Bojarski等^[11]提出了一种端到端的深度强化学习方法，通过深度Q网络（deep Q-network, DQN）直接从原始像素中训练自动驾驶决策系统，无须依赖传统计算机视觉中的中间处理步骤。Kendall等^[12]则提出一种仅需单日训练即可使自动驾驶系统掌握驾驶能力的方法。该方法结合深度强化学习与仿真环境，高效训练出能够应对复杂城市路况的智能体。此外，Kiran等^[13]在其综述中系统阐述了深度强化学习在自动驾驶中的多类应用，涵盖路径规划、车辆控制与传感器融合等方面，并对当前挑战和未来研究方向进行了探讨。这些研究成果不仅突显了深度强化学习在自动驾驶决策中的潜力，也预示着其将在未来技术演进中发挥关键作用。

1.2 模仿学习

模仿学习是一种通过观察专家行为来学习任务的机器学习方法，使智能体能够在缺乏明确奖励信号的情况下，通过模仿专家决策来掌握策略。在强化学习领域中，模仿学习常被用作预训练手段：智能体通过观察和复制专家行为，快速获得较优的初始策略，随后通过与环境的交互进一步优化该策略，从而提升整体性能。

Ng等^[14]提出了基于专家示范的学习方法，使智能体能够在没有明确奖励信号的情况下通过模仿完成复杂任务。Finn等^[15]开发了引导成本学习方法，将模仿学习与强化学习相结合，通过策略优化提升了学习效率。Codevilla等^[16]利用条件模仿学习实现了端到端的自动驾驶系统，展示了该方法在实际场景中的应用潜力。王晓芳等^[17]提出了一种基于深度强化学习与模仿学习的战斗机智能机动突防算法，解决了战斗机在攻击目标过程中遭遇防御弹拦截时需要同时满足突防与打击要求的复杂决策问题。Freeman等^[18]提出了一种无须目标值的学习框架，借助模仿学习

预测专家行为，特别适用于缺乏明确奖励信号的环境。这些研究成果体现了模仿学习在强化学习中的多样化应用与重要性。模仿学习为强化学习领域提供了一种新的学习范式：通过复制专家行为来引导智能体学习，摆脱了对传统奖励信号的依赖。该方法不仅加速了策略的初步形成，也为后续通过与环境的交互进一步优化策略提供了良好基础，从而提升整体学习效率与性能。

1.3 隐私窃取攻击

隐私窃取攻击^[19]是一种针对算法模型的攻击方式。在该类攻击中，攻击者通过精心构造的输入数据或查询，试图从目标模型中提取或推断其训练数据中所包含的敏感信息（如个人身份、医疗记录或其他私密内容）。这类攻击主要利用模型在训练过程中可能对部分训练数据产生“记忆”，或在输出中无意“泄露”关键信息的特性，若模型缺乏足够的隐私保护机制，此类信息泄露风险将进一步加剧。隐私窃取攻击一旦成功，可能导致敏感数据泄露，不仅严重威胁个人隐私与数据安全，甚至还会引发法律风险，从而削弱公众对人工智能技术的信任。

Gomrokchi等^[20]首次提出针对深度强化学习智能体的黑盒成员推断攻击。他们设计并引入了个体层面与集体层面的成员推断攻击模式，以评估如何利用时序与批次层面的相关性来揭示训练数据的存在程度。该项研究表明，深度强化学习模型在面对集体层面的成员推断攻击时表现出更高的脆弱性。在另一项研究中，Pan等^[10]开创性地探讨了深度强化学习场景中的隐私泄露问题，通过访问若干候选环境动态，提出一种能够推断给定策略所用训练环境的算法。然而，该方法仅对过拟合的深度强化学习模型具有较强推断能力，对于正常训练或具备良好泛化能力的深度强化学习模型则难以有效实施攻击。目前，针对具有泛化能力的深度强化学习模型的隐私安全问题，相关研究仍处于初步探索阶段，其隐私保护问题尚未得到充分重视。

2 方法

本文提出一种基于模仿学习的目标场景隐私攻击方法。该研究主要从3个层面展开：①对相关问题进行明确定义，随后构建整体系统框架；

②通过影子智能体构建轨迹数据集,并基于该数据集训练攻击模型;③利用训练好的模型对目标轨迹进行场景归属推断,判定其所属的候选训练场景,并通过实验验证所提方法的有效性。

2.1 问题定义

(1) 深度强化学习

深度强化学习^[21]的环境可建模为马尔可夫决策过程(Markov decision process, MDP),其由一个四元组(状态空间 S 、动作空间 A 、状态转移概率矩阵 T 和奖励函数 R)组成。在该框架中,折扣因子 γ 用于对未来奖励进行折现计算,取值范围为 $(0,1]$ 。深度强化学习模型通过学习策略函数 $\phi(a|s)$ 来指导智能体行动,该函数将环境状态映射为动作的概率分布,其优化目标是最大化智能体在整个交互过程中获得的期望累积折扣奖励 $E_{\pi}[\sum_{t=0}^{H-1} \gamma^t r_t]$,其中 H 为每一个训练回合的最大步长。状态转移概率矩阵 T 是MDP中的关键组成部分,刻画了智能体与环境交互过程中的动态特性。具体而言,状态转移矩阵 T 中的元素 $P(s'|s,a)$ 表示智能体在状态 s 下执行动作 a 后转移到状态 s' 的概率。在该矩阵中,每一行代表一个状态,每一列代表执行一个动作后的转移概率。本文聚焦于具有泛化能力的深度强化学习模型,其策略 π 通过典型深度强化学习算法(如DQN、PPO等)优化获得,能够在未见过的测试环境中仍保持较高性能,展现出良好的迁移与适应能力。

(2) 攻击者的目标

攻击者根据给定的一个已训练完成的强化学习策略 π ,结合马尔可夫决策过程中的已知组件(包括状态空间 S 、动作空间 A 及奖励函数 R),推断出状态转移概率矩阵 T 。攻击者通过隐私推断方法,能够获取目标模型与其所处环境的真实交互信息,从而推断出目标模型实际使用的状态转移概率矩阵和最符合目标模型的候选场景。

攻击者的目标可以归纳为以下3点。

1) 构建影子轨迹数据集。攻击者首先基于模仿学习方法,利用给定策略的轨迹数据训练影子模型,使其行为尽可能接近目标模型的行为。随后,将训练得到的影子模型部署到多个候选场景

中进行交互,生成并收集各场景下的影子轨迹,构建用于后续分析的数据集。

2) 构建分类模型。通过分类模型分析目标模型与环境交互生成的轨迹信息,并将不同候选场景作为分类目标,评估各候选场景与目标模型实际情况的匹配程度。

训练场景分类模型:攻击者构建一个分类模型,以目标模型与环境交互产生的轨迹作为输入、候选场景作为类别标签,通过监督学习方式训练该模型,从而评估每条轨迹属于不同候选场景的匹配概率。

3) 隐私推断效果。通过比较目标模型的轨迹信息与每个候选场景下的期望轨迹,攻击者试图识别出最有可能的状态转移概率矩阵。

(3) 攻击者的知识

由于可能存在多个不同的状态转移概率矩阵都能解释同一策略行为,本文假设攻击者具备一定的环境先验知识,即掌握一组具有不同状态分布特性的候选场景和给定策略的轨迹信息。在此条件下,攻击者的具体任务可形式化为一个多分类问题,即从候选场景集合中识别出目标策略最可能使用的训练环境。

2.2 系统框架

针对泛化能力强的深度强化学习模型,本文提出了一种基于模仿学习的目标场景隐私攻击方法。该方法的系统框架如图1所示。

(1) 攻击准备阶段

采用一种基于节点中心度优化的拓扑生成方法,为目标环境构建一系列具有多样化状态分布的候选场景。该方法首先计算图中各节点的中心度,进而通过调整节点间的连接方式,优先连接中心度较高的节点,以增强网络中关键节点与其他节点的交互能力。这一策略不仅有助于提升网络的连通性,还能够有效模拟更加复杂的节点间交互模式。

在此基础上,攻击者通过观察并获取目标模型的行为轨迹,采用基于行为克隆的模仿学习方法构建影子智能体模型。该模型通过模仿目标模型在不同状态下的决策行为,学习其在环境中的响应策略,从而能够在相似场景中复现目标模型的行为特征。通过这种直接的行为模仿,攻击者

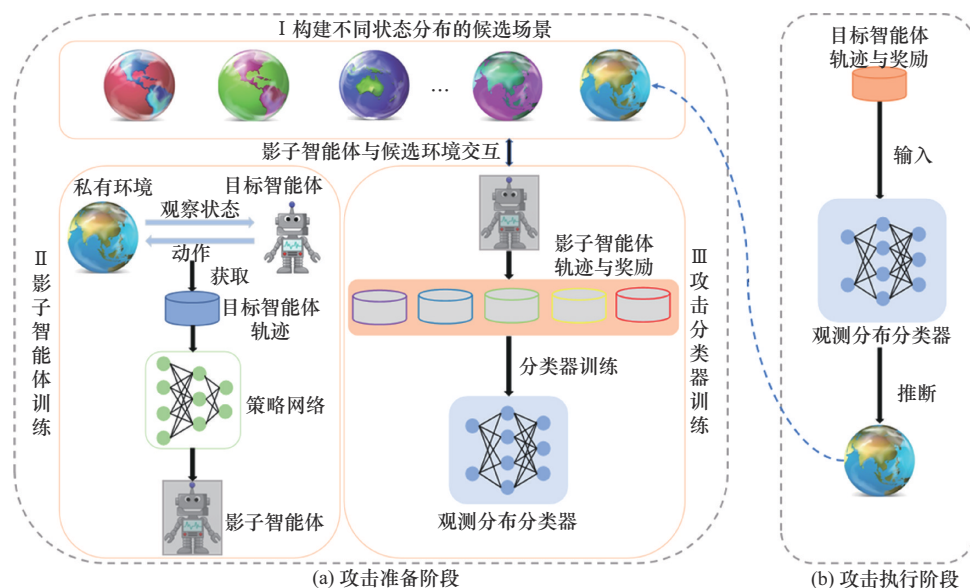


图1 TIPAS系统框架

Figure 1 Framework of the TIPAS

能够高效地构建一个与目标模型行为高度近似的影子智能体。

完成训练后，将影子智能体部署至实际环境中进行交互，收集其在各候选场景下的行为轨迹。这些轨迹数据包含智能体的行为模式与性能表现信息，用于后续构建攻击分类模型。在构建分类模型时，采用集成学习方法 BaggingClassifier，结合多层感知机（multilayer perceptron, MLP）模型进行训练，以提升分类性能并有效抑制过拟合现象。

(2) 攻击执行阶段

攻击者首先获取目标模型在特定场景下的行为轨迹数据，并将其输入至已训练完成的分类模型中。通过对目标轨迹进行特征提取，模型能够推断出目标智能体所对应的候选场景类型，从而实现目标智能体隐私信息的揭露。

2.3 轨迹数据构建

当使用攻击模型对目标模型进行隐私攻击时，需要构建影子智能体的行为轨迹数据。该过程如图1(a)所示，具体实现过程如下。

(1) 构建候选环境集合（见图1中“构建不同状态分布的候选场景”）

采用基于节点中心度优化的拓扑生成方法，为目标场景构建候选环境集合 $\varepsilon = \{E_1, E_2, \dots, E_n\}$ 。具体步骤如下：①定义一个包含 n 个节点的集合

$V = \{v_1, v_2, \dots, v_n\}$ ；②计算集合中每个节点的接近中心度 $C(v_i)$ ，并根据节点的中心度值对节点进行排序。在构建连接时，优先连接中心度较高的节点，以增强网络的连通性与节点间的紧密性。若在添加边的过程中产生孤立节点，则根据中心度优先连接这些孤立节点，确保整个图的连通性。此过程将持续迭代，直至所有节点均被连通且满足拓扑结构的需求。通过上述方法构建的一系列候选场景，其中的每个候选环境 E_i 具有不同的状态分布 S_i ，其多样性确保了影子智能体能够学习广泛的行为模式，从而提升攻击方法的泛化能力。不同的状态分布 S_i 表示不同的物理位置、任务场景或约束条件。例如，在自动驾驶网络场景中，不同的状态分布 S_i 代表着具有相同节点与连边数量，但结构不同的场景。

(2) 训练影子智能体模型（见图1中“影子智能体训练”）

影子智能体通过行为克隆模仿目标策略 π_{target} 的决策行为。目标策略 π_{target} 的决策能力源于其与环境的交互学习过程。因此，其在执行过程中产生的行为轨迹（即一系列状态-动作对）蕴含的训练环境的信息，是攻击分析的核心依据。首先，搭建影子智能体模型结构，包括输入层、隐藏层和输出层3个部分。输入层负责接收环境的当前状态信息，并将其传递给隐藏层，以便模

型能够基于当前环境进行相应的决策。模型包含两个隐藏层，每个隐藏层由 64 个神经元组成，采用 ReLU 激活函数。模型通过非线性变换提取复杂特征，以适应目标模型的行为模式。输出层采用 Softmax 激活函数，其输出维度与动作空间维度相匹配，表示在给定状态下选择各动作的概率分布。通过 Softmax 函数，模型能够将最后一层的线性输出转换为概率值，确保所有可能动作的选择概率之和为 1。该层预测行为 $P(a|s)$ 的定义如下：

$$P(a|s) = \frac{e^{z_a}}{\sum_j e^{z_j}} \quad (2)$$

其中， z_a 是对应动作 a 的线性输出， z_j 是所有动作的线性输出。

在搭建完模型后，攻击者着手收集给定策略的行为轨迹数据。这些数据包括状态 s 及其对应的动作 a ，形成数据集 $D = \{(s_i, a_i)\}_{i=1}^N$ ，其中 N 是样本数量。接下来，定义影子智能体为一个策略函数 $\pi_\theta(s)$ ，该函数的输入为状态 s ，输出为动作 a ， θ 表示影子智能体模型的参数。影子智能体的训练目标是 최소화 预测动作 $\pi_\theta(s)$ 与目标模型真实动作 a 之间的差异。这一过程可以视为监督学习任务，通过最小化损失函数 $L(\theta)$ 来调整模型参数 θ ，其中损失函数定义为目标动作 a 与预测动作 $\pi_\theta(s)$ 之间的均方误差。

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N \|a_i - \pi_\theta(s_i)\|^2 \quad (3)$$

为优化该损失函数，采用梯度下降法更新影子智能体模型的参数 θ ，其更新规则如下：

$$\theta = \theta - \alpha \nabla_\theta L(\theta) \quad (4)$$

其中， α 表示学习率， $\nabla_\theta L(\theta)$ 表示损失函数相对于模型参数的梯度。通过上述步骤，攻击者能够训练出一个高度逼近目标模型行为特征影子智能体模型。

(3) 构建轨迹数据集（见图 1 中“影子智能体与候选环境交互”）

将训练完成的影子智能体模型 M 置于候选环境 E_i 中进行交互，生成轨迹 τ 。每条轨迹 τ 由一系列状态—动作对组成，表示为：

$$\tau = \{(s_0, a_0), (s_1, a_1), \dots, (s_t, a_t)\} \quad (5)$$

其中， s_t 表示时间步 t 的状态， a_t 表示在时间步 t

根据当前状态与策略所选的动作。影子智能体每执行完一个动作，环境状态将发生转移。针对每个候选环境 E_i ，需要生成一组轨迹数据，构成该场景的轨迹集：

$$D_i = \{\tau_{i1}, \tau_{i2}, \dots, \tau_{im}\} \quad (6)$$

随后，在不同环境中重复上述交互过程，使用相同的影子智能体模型 M 与策略 π ，分别获取各环境 E_j 对应的轨迹数据 D_j 。最终，整合所有环境的轨迹数据，构建出完整的轨迹数据集：

$$D = \{D_1, D_2, \dots, D_n\} \quad (7)$$

2.4 攻击模型训练

在得到完整轨迹数据集后，攻击模型的目标是推断目标模型曾在哪一种状态分布的环境中进行训练。该过程如图 1 (a) 中“攻击分类器训练”所示。下面详细介绍其实现原理。

首先，为每个候选环境 E_i 的轨迹数据赋予相应的环境标签。对于每一条轨迹 τ_{ij} ，其对应的环境标签为 i ，据此构建带标记的数据集：

$$D = \{(\tau_{ij}, i) | \forall i, j\} \quad (8)$$

MLP (multilayer perception) 具有出色的表示学习能力，能够有效处理复杂的非线性问题，并从高维数据中提取有用的特征^[22]，因此被广泛用来处理图像识别^[23-24]、语音识别^[25]等分类任务。为提升模型的鲁棒性，本文采用 Bagging-Classifer^[26]与 MLP 结合的集成方法。Bagging-Classifier 是一种用于分类任务的集成学习方法，能够通过组合多个 MLP 子模型，有效减少模型方差，从而提高分类性能。每个 MLP 子分类器的网络结构统一定义为：

$$\tilde{y} = \text{Softmax}(W_2 \cdot \text{ReLU}(W_1 \cdot x_i + b_1) + b_2) \quad (9)$$

其中， $\tilde{y}$ 表示第 i 个 MLP 子分类器的预测结果， W_1 和 W_2 分别表示隐藏层和输出层的权重矩阵， x_i 表示第 i 个分类器输入的轨迹特征向量， b_1 和 b_2 均为偏置向量，ReLU() 为激活函数，Softmax 用于输出类别概率。组合多个分类器模型后对每个子模型进行不同数据的采样并独立训练，每个分类器独立优化其参数，采用交叉熵损失函数评估预测标签与真实标签之间的差异，其表达式为：

$$L = - \sum_{i=1}^N y_i \log(\tilde{y}_i) \quad (10)$$

其中， N 表示样本数量， y_i 表示真实类别标签， $\tilde{y}_i$ 表示预测标签。最后对所有子分类器的预测结

果进行投票，确定最终的预测类别：

$$\tilde{y} = \operatorname{argmax} \sum_{i=1}^M \tilde{y}_i \quad (11)$$

其中， M 表示分类器的个数， $\tilde{y}_i$ 表示第 i 个分类器的预测结果，通过 argmax 函数返回最大概率的标签所对应的类别，即最终的预测类别 $\tilde{y}$ 为获得最多投票的类别。

2.5 目标轨迹推断

在训练完攻击模型后，将目标模型的行为轨迹数据 x 输入该攻击模型中。每个独立的MLP子分类器首先对输入数据进行特征提取，并输出对应的分类预测结果。

$\tilde{y}_i = \operatorname{Softmax}(W_2^i \cdot \operatorname{ReLU}(W_1^i \cdot x + b_1^i) + b_2^i)$ (12)
其中， $\tilde{y}_i$ 表示第 i 个MLP子分类器对输入的目标模型的行为轨迹数据 x 的预测结果。通过 $\operatorname{Softmax}$ 函数对MLP子分类器的输出进行归一化处理，得到其在各候选场景上的概率分布。 $\operatorname{BaggingClassifier}$ 汇总所有子分类器的预测结果，采用投票机制确定最终的预测标签，其中最大概率的标签对应的环境类别即为推测的目标候选环境。

$$\operatorname{env} = \operatorname{argmax} \sum_{i=1}^M \tilde{y}_i \quad (13)$$

其中， env 表示最终预测的目标候选环境标签， $\operatorname{argmax} \sum_{i=1}^M \tilde{y}_i$ 表示从所有MLP子分类器的预测结果中选择概率最高的类别作为最终预测的目标候选环境。综上，基于模仿学习的目标场景隐私攻击算法的流程如算法1所示。

算法1 基于模仿学习的目标场景隐私攻击算法

输入 目标智能体轨迹数据 T_{tar}

输出 根据轨迹数据推断出的场景 S_{inf}

1) 初始化：构建候选场景集合 $S = \{S_1, S_2, \dots, S_n\}$

2) $\pi_\theta(s) \leftarrow$ 使用目标智能体轨迹数据 T_{tar} 和式(3)定义的损失函数训练影子智能体模型

3) for env in S :

将影子策略置于环境 env 中进行测试，收集轨迹数据 $\tau = f(\pi_\theta(s), \operatorname{env})$

4) 收集每个 env 场景下的轨迹数据

$$D_{\operatorname{env}} = \{\tau_{\operatorname{env}1}, \tau_{\operatorname{env}2}, \dots, \tau_{\operatorname{env}m}\}$$

5) 构建完整的数据集 $D = \{D_1, D_2, \dots, D_n\}$

6) $f_i(x) \leftarrow$ 根据 $\operatorname{BaggingClassifier}$ 集成学习构建多个分类器

7) $f'_i(x) \leftarrow$ 基于数据集 D 与式(10)所示的损失函数训练攻击模型

8) $y_i' \leftarrow$ 根据式(12)获取每个分类器的输出预测结果

9) $S_{\text{inf}} \leftarrow$ 通过式(13)投票机制确定最终预测的目标候选环境标签

10) return: 推断出的场景 S_{inf}

2.6 方法有效性分析

基于模仿学习的目标轨迹隐私推断方法，通过影子智能体模型获取候选场景的轨迹特征信息，并将其用于训练攻击模型。影子智能体模型采用行为克隆方式进行训练，其核心目标是直接模仿目标智能体的决策行为。该方法的关键在于通过逐步逼近目标模型的行为，确保影子智能体在相同环境下能够复制目标智能体的策略，进而在攻击阶段利用这一特征进行场景识别和隐私推断。影子智能体模型的训练旨在通过最小化预测行为与目标行为之间的差异，实现对目标策略的复现。具体而言，在给定状态 s_t 下，影子智能体选择的目标动作作为 a_t ，并通过监督学习得到其策略表示：

$$a_t = f(s_t, \theta) \quad (14)$$

其中， $f(\cdot)$ 为影子智能体的策略函数， θ 表示神经网络中的权重参数。通过迭代训练与参数调整，影子智能体能够逐步学习目标智能体在不同状态下的决策规则和行为模式。

影子智能体通过学习目标模型的策略，在相同环境下生成大量的轨迹数据。这些轨迹由连续的状态-动作对构成，反映了智能体在不同场景中的决策逻辑，可作为区分场景的有效特征。具体而言，第 j 个候选场景的轨迹可表示为：

$$T_j = \{(s_t, a_t) | s_t \in S, a_t \in A\} \quad (15)$$

其中， s_t 表示状态， a_t 表示在状态 s_t 下执行的动作。轨迹数据所包含的状态-动作关联信息，为分类器提供了具有判别能力的特征输入。基于这些轨迹特征，分类器的学习目标可表述为：

$$f: T_j \rightarrow C_j \quad (16)$$

其中， C_j 表示场景特征的标签。通过学习不同候选场景的轨迹特征，分类器能够辨识不同候选场景下的轨迹模式，从而实现对目标智能体所在环

境的有效推断。为进一步提升分类性能,本文采用集成学习方法组合多个分类器共同决策,不仅增强了对轨迹特征的学习能力,还显著提高了分类的准确率与鲁棒性。

综上所述,该方法的核心在于利用目标深度强化学习策略在不同候选环境中表现出的行为轨迹差异进行隐私推断。借助影子智能体对目标策略的高度还原,攻击者通过集成分类器能够精确识别目标智能体所处的训练场景,从而有效揭示其隐私信息。

3 实验

本节对基于模仿学习的目标场景隐私攻击方法进行实验验证与结果分析。首先介绍实验环境设置与所采用的评价指标,并说明实验场景及用于对比的基线方法;随后,在不同场景下对所提隐私攻击方法的性能进行评估,分析其时间消耗;接着,通过参数敏感性分析探究数据集规模与轨迹长度对攻击效果的影响,测试该方法在一般防御机制下的攻击性能;最后,在 Gym 游戏环境^[27]中进一步验证所提方法的有效性。

3.1 实验环境设置

实验采用的硬件与软件配置如下:CPU 为 Intel i7-7700K @ 4.20GHz,内存为 16 GB×4,操作系统为 Ubuntu 16.04,编程语言为 Python 3.7.16,深度学习框架为 TensorFlow-2.7.0。

实验平台基于 OpenAI Gym^[27]框架实现光传输网络(optical transport network, OTN)模拟器。该模拟器是一种专用于光传输网络研究的开源强化学习环境。为满足用户多样化的研究需求,模拟器支持设置不同拓扑结构的网络(从简单的点对点链路扩展至复杂网状网络)。

Almasan 等^[28]提出将图神经网络(graph neural network, GNN)集成到深度强化学习智能体(DRL-GNN)中,以解决自动驾驶网络中的路由优化问题。该方法利用 GNN 捕捉网络拓扑中链路流量之间的关系,提取有意义的结构特征,从而使智能体在未知拓扑上仍能保持良好的性能。研究表明,采用消息传递神经网络(message passing neural network, MPNN)的深度强化学习模型在不同网络拓扑中表现出较强的泛化能力。为此,本文选取 DQN^[29]、DDQN^[30]、Duel-

ing DQN^[31]和 PPO^[32]算法分别与 MPNN 相结合作为基准策略,探究具备泛化能力的模型所面临的隐私风险。

实验采用以下评价指标。

1) 平均回合奖励(average round reward, ARR):用于评估深度强化学习智能体的综合表现。奖励值越大,意味着智能体性能越好。

2) 推断准确率(inference accuracy, IA):作为隐私推断攻击效果的衡量标准,定义为攻击模型正确分类的样本数占总样本数的比例,具体计算公式如下:

$$IA = \frac{N_{\text{correct}}}{N_{\text{total}}} \quad (17)$$

3.2 实验网络场景介绍

本文在以下 4 种不同结构的 OTN 环境中进行实验,各网络具有相同的节点数与连边数,但在结构上存在差异。

1) NSFNET(国家科学基金会网络):该网络包含 14 个节点和 21 条光路,用于模拟高效传输环境。

2) GEANT 网络:该网络是一个泛欧学术骨干网,包含 24 个节点和 37 条光路,用于研究不同结构下的网络性能和隐私攻击的鲁棒性。

3) GBN:该网络是基于 NetworkX Python 库^[33]生成的随机网络,包含 17 个节点和 25 条光路。其平均节点度与 NSFNET 相近,用于分析在相似节点度下不同结构对攻击效果的影响。

4) Topology 网络:该网络是选自 Topology Zoo^[34]数据集中的真实网络,包含 50 个节点和 70 条光路,用于研究大规模网络中隐私与性能之间的关系。

3.3 对比算法

本文选取以下两种方法作为对比算法,用于验证所提方法的有效性。

1) 基于奖励均值与方差特征的支持向量机(support vector machine, SVM)隐私推断方法^[10]。该方法通过分析策略在不同候选环境下的性能表现来推断其训练环境。为了建立分类模型,首先在 N 种环境动态下训练 $N \times m$ 个策略,用于构建训练数据集,每个策略采用 m 个随机种子进行初始化。相应候选迁移动态 T_i 上训练的第 j 个策略表示为 π_j^i 。然后在 N 种不同环境动态下

收集各策略的情景奖励，每种策略进行 k 次实验。计算各次实验中情景奖励的均值和方差特征，用于构建每个策略的特征向量 R^{2N} 。然后根据特征向量与每个环境动态候选对应的标签拟合一个分类器。在测试期间，给定一个目标策略 π ，目标以类似的方式构建特征向量：通过 k 次不同的实验，计算 N 种动态中每种动态的情景奖励，从而预测其原始训练环境。在本节图表中，该方法简称为 SVM。

2) 单一 MLP 模型方法：为验证集成学习的有效性，本文还引入了不采用集成学习的单一 MLP 模型作为对比算法。

3.4 基于模仿学习的隐私攻击性能验证

针对基于奖励均值和方差特征的 SVM 隐私推断方法，本节在不同网络结构的候选环境中针对

每个环境构建了 m 个随机种子，并基于不同的基准算法模型进行训练。以第一个随机种子在 NSFNET 和 Topology 两种拓扑下的训练过程为例，其训练过程的奖励收敛曲线分别如图 2 和图 3 所示。

从训练曲线可以看出，各模型在不同网络拓扑环境下均表现出良好的收敛特性。具体而言，在 NSFNET 与 Topology 两种拓扑中，各基准模型在经过充分训练后，其平均回合奖励均逐渐趋于稳定，表明模型已成功学习到有效的决策策略并实现性能优化。值得注意的是，尽管网络拓扑结构不同，但是各模型的训练过程表现出较高的一致性，收敛轨迹相近。相比之下，在同一拓扑下的不同候选场景中，模型的训练过程呈现较大的差异。这一现象主要源于各候选场景在网络资源分配与链路状态配置上的不同。在资源充裕的场

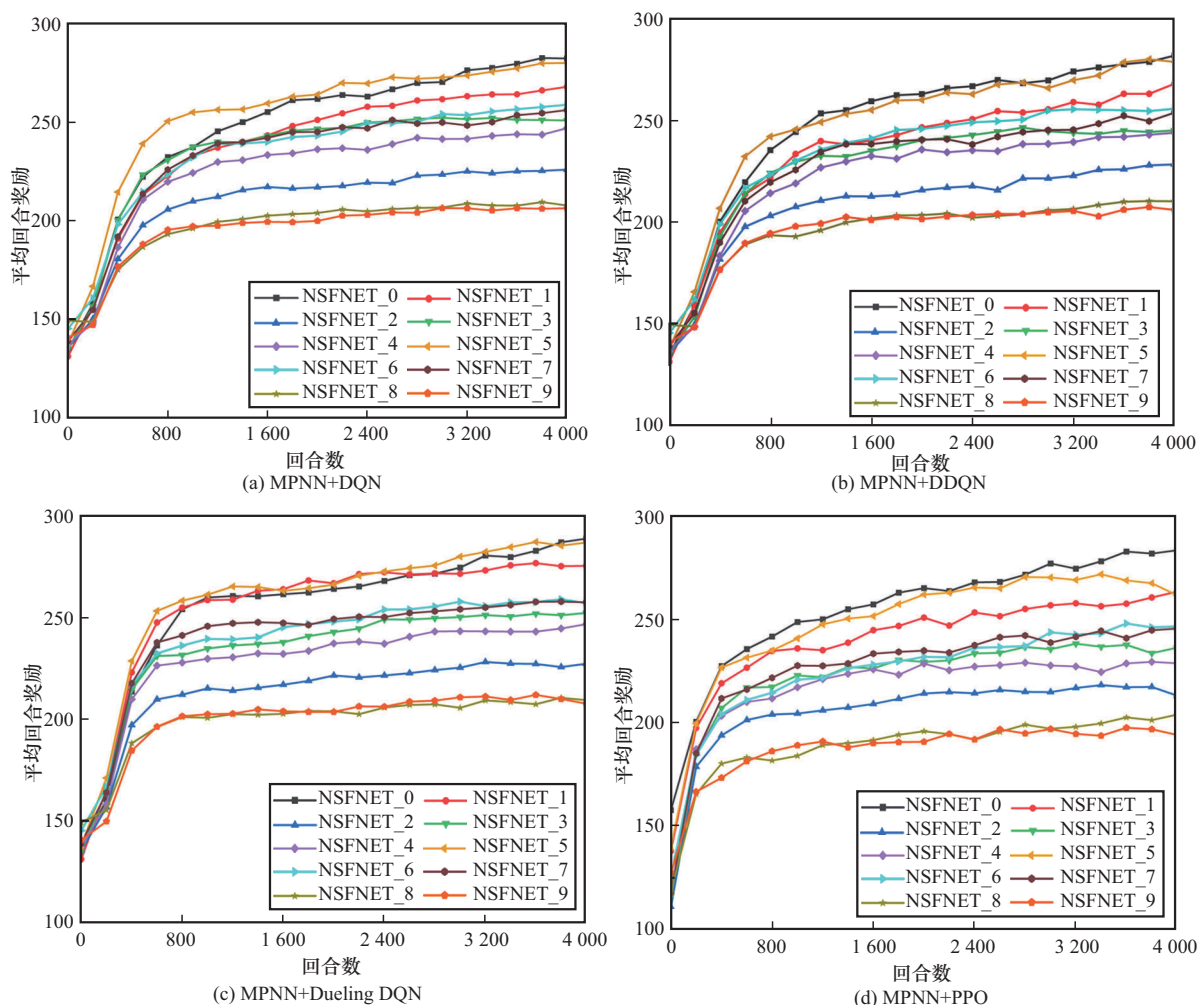


图2 NSFNET不同候选场景下替身模型训练曲线

Figure 2 NSFNET surrogate model training curve under different candidate scenario

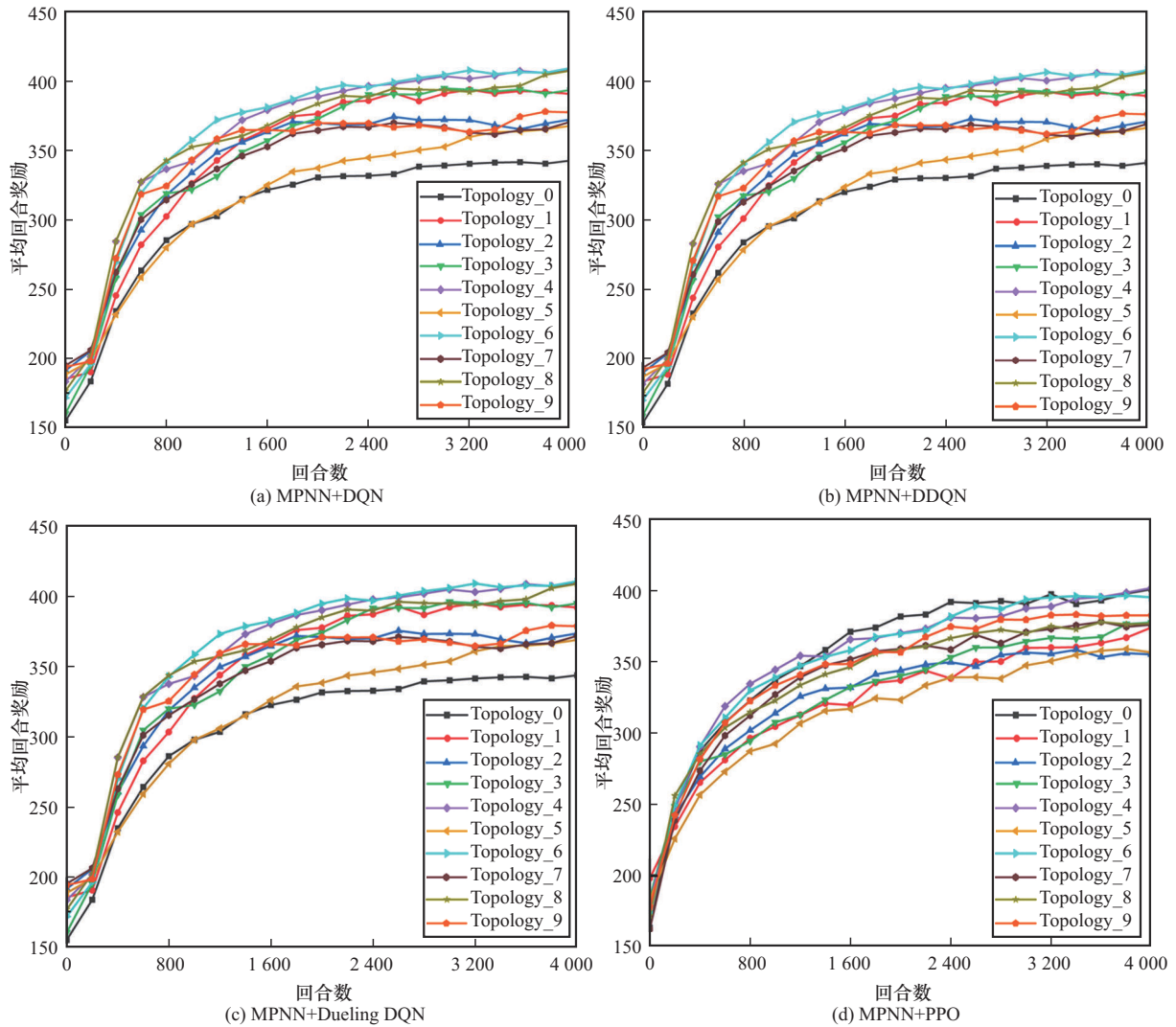


图3 Topology不同候选场景下替身模型训练曲线

Figure 3 Topology surrogate model training curve under different candidate scenario

景中, 模型更容易寻找到最优路径与资源分配策略, 从而获得较高奖励; 而在资源受限的场景中, 由于资源分配无法满足用户需求, 模型获得的奖励显著减少, 进而影响其训练过程与最终性能。

本文对提出的TIPAS在4种网络拓扑下的候选场景的推断准确率进行了评估, 并与基于奖励值和均方差特征的SVM方法及单一MLP模型方法进行了对比。如表1所示, TIPAS在各种网络拓扑下均取得了最高的推断准确率, 显著优于两种对比方法。这一优势主要源于TIPAS对深度强化学习行为轨迹中时空模式等深层特征的捕捉能力。与传统SVM方法仅依赖奖励统计量相比, TIPAS能够提取更具判别性的隐私特征; 而与单一MLP模型相比, 其采用的BaggingClassifier集

成策略有效提升了模型的鲁棒性与泛化能力, 从而在不同网络环境中均表现出更稳定的推断性能。

这一结果表明, TIPAS方法具有卓越的泛化能力和适应性, 能够在不同网络环境中对目标场景进行精准识别。其主要优势在于, 该方法不仅利用了丰富的场景轨迹数据, 还通过特征提取与表示更全面地捕捉了轨迹中的信息。同时该方法采用了BaggingClassifier集成学习策略, 通过结合多个基分类器的预测结果来降低模型的方差, 从而在处理复杂网络结构时进一步提升了模型的稳定性与预测精度。这种设计使得TIPAS方法更适合应对实际环境中变动性较强的网络结构。

表1 不同算法的推断准确率对比

Table1 Comparison of the accuracy of different algorithms

网络拓扑	候选场景 (10)	算法模型	TIPAS	SVM	单一 MLP
NSFNET	14 节点, 21 连边	MPNN+DQN	92.46%	42.50%	83.67%
		MPNN+DDQN	93.78%	38.75%	71.82%
		MPNN+Dueling_DQN	92.85%	27.50%	78.29%
		MPNN+PPO	91.40%	21.25%	73.75%
GEANT	24 节点, 37 连边	MPNN+DQN	90.94%	28.75%	71.47%
		MPNN+DDQN	92.57%	25.00%	76.04%
		MPNN+Dueling_DQN	93.77%	48.75%	81.87%
		MPNN+PPO	90.27%	56.25%	77.32%
GBN	17 节点, 25 连边	MPNN+DQN	92.63%	21.25%	75.66%
		MPNN+DDQN	90.18%	3.75%	85.52%
		MPNN+Dueling_DQN	81.50%	27.5%	74.48%
		MPNN+PPO	88.40%	26.25%	77.83%
Topology	50 节点, 70 连边	MPNN+DQN	88.97%	36.25%	75.50%
		MPNN+DDQN	87.87%	22.5%	78.61%
		MPNN+Dueling_DQN	91.40%	20.00%	72.38%
		MPNN+PPO	78.46%	15.00%	73.50%

相比之下, 基于奖励均方值和方差特征的 SVM 方法在推断过程中仅依赖于奖励值信息, 未能充分挖掘轨迹数据中蕴含的复杂特征表达, 因此在捕捉场景的潜在复杂关系时表现出明显不足。尽管 SVM 在某些线性任务上表现良好, 但在处理非线性特征时能力有限, 难以适应实际场景中的多样性和复杂性。

单一 MLP 模型虽然具备对非线性关系的拟合能力, 但是其仅依赖单一的网络结构, 容易受到数据噪声和样本复杂性的干扰, 使得其在推断任务中的稳定性和泛化能力相对较差。因此, 单一 MLP 模型在不同网络拓扑下的表现不够理想。

此外, 本文还测试了 4 种拓扑下 3 种隐私推断方法构建攻击模型时间开销, 结果如表 2 所示。从表中数据可以看出, SVM 方法的时间开销最小 (仅需 0.01 s)。这得益于其线性核函数所带来的计算简便性 (训练过程相对简单, 时间复杂度较低)。而 TIPAS 的时间开销最高达到 282.03 s。这主要是因为 TIPAS 采用了 Bagging-Classifer 集成学习策略。该策略需要构建多个基分类器, 并对每个基分类器进行训练, 从而在提

升模型泛化能力和稳定性的同时增加了训练时间。尽管时间成本较高, 但其在准确率与泛化能力上的显著提升, 因此, 时间成本的增加是可以接受的。

表2 不同算法的时间开销对比

Table2 Comparison of time costs of different algorithms

网络拓扑	TIPAS/s	SVM/s	单一 MLP/s
NSFNET	245.66	0.01	57.07
GEANT	231.00	0.01	50.11
GBN	191.98	0.01	42.19
Topology	282.03	0.01	61.88

3.5 参数敏感性分析

本文开展了参数敏感性分析, 探究数据集规模和轨迹长度对隐私推断攻击效果的影响。本文在 4 种网络拓扑下构建了包含不同轨迹长度的数据集, 评估了在多种轨迹长度条件下 4 种模型在隐私推断攻击中的表现。实验中, 轨迹长度被定义为智能体完成分配任务所经历的总步数。本实验分别构建了轨迹数量为 20 条、40 条、60 条、80 条和 100 条的数据集, 并基于这些数据集展开实验分析。图 4 展示了在 4 种网络拓扑下不同数

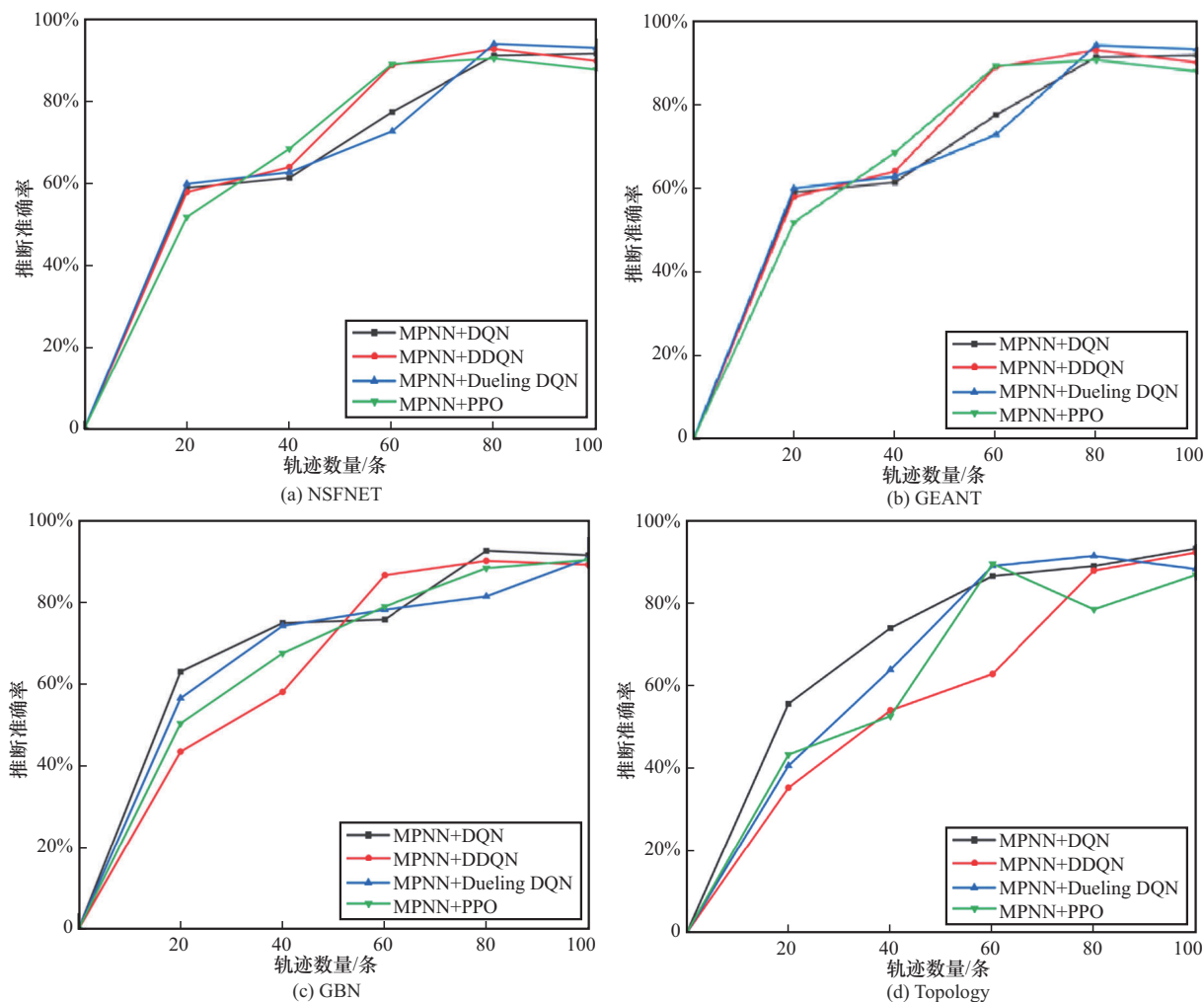


图4 不同网络拓扑下不同数据集规模的隐私推断准确率

Figure 4 Privacy inference accuracy of different data set sizes under different network topologies

数据集规模的隐私推断准确率。实验结果显示,随着轨迹数据集规模的增加,隐私推断攻击的准确率在 NSFNET、GEANT、GBN 和 Topology 拓扑中均呈现上升趋势。具体而言,当轨迹数量达到 80 条和 100 条时,隐私推断准确率分别达到最高水平。这主要源于更大规模的数据集提供了更加丰富的行为特征,从而增强了攻击者对智能体行为模式的捕捉与推断能力,最终提升了攻击准确率。此外,在 Topology 拓扑中,针对 MPNN+PPO 模型的隐私推断攻击在轨迹数量为 60 条时达到峰值。这一现象可能与 Topology 拓扑的结构特性及该模型的训练机制有关。在该场景下,60 条轨迹已足以充分描述 MPNN+PPO 智能体的行为特征,进一步增加数据规模可能引入冗余信息,导致攻击效果趋于饱和。此外,本文进一步

分析了轨迹长度对攻击模型推断准确率的影响。实验固定轨迹数据集规模为 80 条,并分别构建了轨迹长度为 5 步、10 步、15 步、20 步和 25 步的数据进行测试。图 5 展示了在数据集规模相等的条件下,4 种网络拓扑下不同轨迹长度的隐私推断准确率。从实验结果来看,在 4 种网络拓扑下,随着轨迹长度的增加,攻击模型的推断准确率均呈逐步上升的趋势,并在轨迹长度达到 20 步或 25 步时达到峰值。这主要源于较长的轨迹数据能够反映目标模型在训练过程中更完整的行为特征,从而帮助攻击模型更准确地地区分其所训练的具体候选场景。同时,更长的轨迹也涵盖了更丰富的时序信息与环境交互细节,增强了攻击模型对目标模型训练环境的推断能力,进而提升了攻击的准确率。

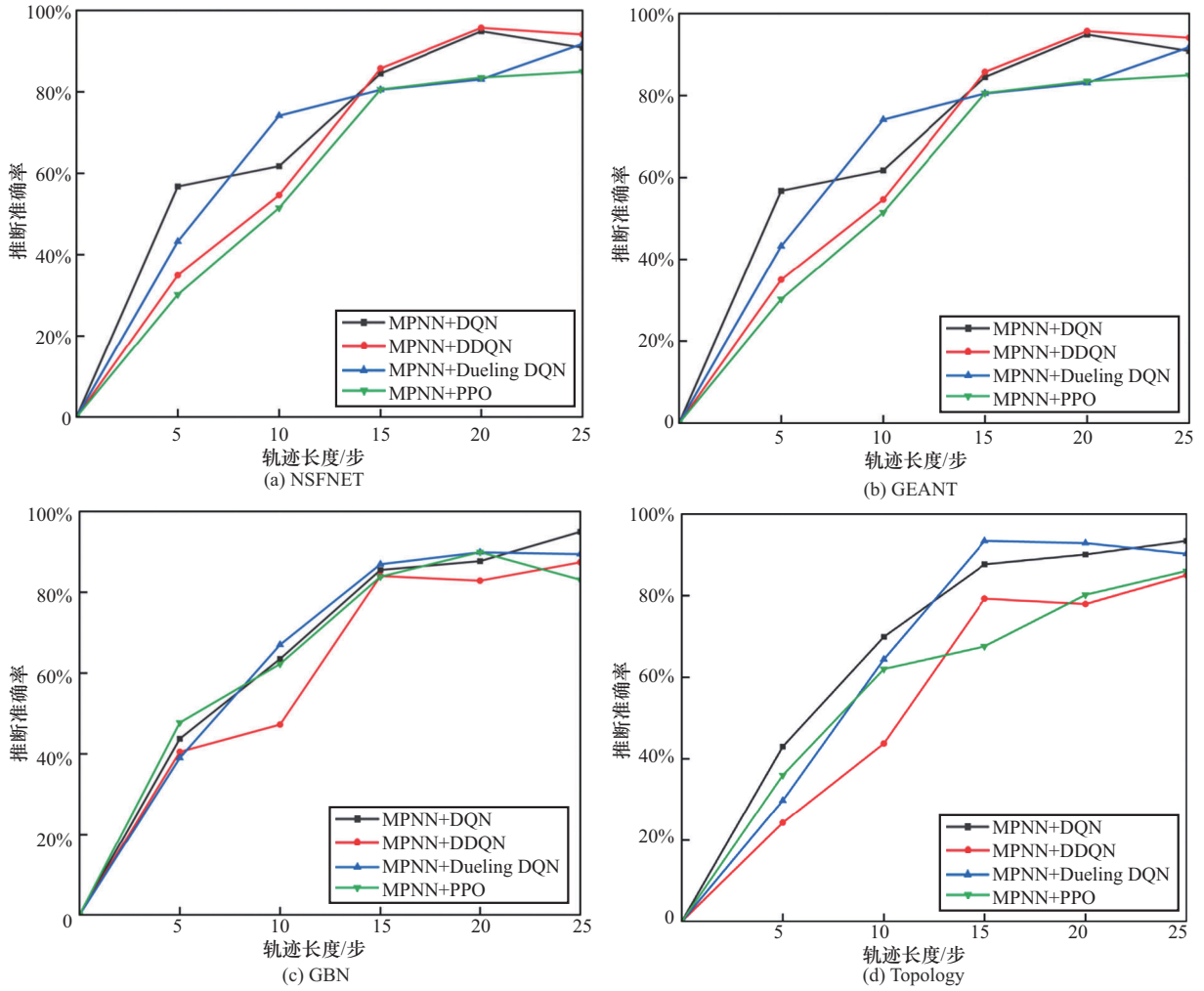


图5 不同网络拓扑下不同轨迹长度的隐私推断准确率

Figure 5 Privacy inference accuracy of different trajectory lengths under different network topologies

3.6 防御条件下的攻击性能

本节探究 TIPAS 和单一 MLP 模型在隐私保护条件下的性能表现。实验针对不同的网络拓扑与智能体模型，分别采用了差分隐私 (differential privacy, DP) [35] 和 Drop_out [36] 两种隐私保护方法进行验证。

为在深度强化学习中实现差分隐私的效果，本文采用了一种简化的策略扰动机制，该机制的核心在于降低模型对训练数据（尤其是状态-动作映射关系）的敏感性，从而降低单个样本在策略行为中的可识别性。具体而言，在策略网络的 Softmax 输出层之前引入温度缩放操作，即将每个动作的 logit 输出除以一个较大的温度系数 ($T=3$)，如式(18)所示。

$$\pi(a|s) = \frac{\exp\left(\frac{z_a}{T}\right)}{\sum_i \exp\left(\frac{z_i}{T}\right)} \quad (18)$$

该方法不会显著改变最终选择的动作（即 $\arg \max$ 结果基本不变），但能够显著平滑动作概率分布，使模型在不同状态下的决策倾向更加模糊。这种概率分布的平缓化弱化了环境特征在行为中的显性表达，从攻击者视角来看，相当于抑制了可用于环境推断的隐私信号，从而达到保护策略隐私的目的。

Drop_out 作为另一种防御策略，通过在训练过程中随机屏蔽部分神经元的激活来防止模型过拟合。在本实验中，统一在策略网络的隐藏层中设置 Drop_out 概率为 0.5。尽管 Drop_out 并非直

接为隐私保护设计，但其正则化作用能够间接降低模型对输入数据的依赖性，因此也具备一定的防御效果。

实验结果如表 3 所示，从表中数据可以看出，在相同的隐私保护条件下，本文提出的 TIPAS 在隐私推断攻击中的表现优于单一 MLP 模型。以 NSFNET 拓扑下 MPNN+PPO 模型的行为轨迹隐私推断为例，在施加 DP 和 Drop_out 防御后，TIPAS 的攻击成功率分别下降 3.76% 和 5.15%，而单一 MLP 模型的攻击成功率则分别下降 8.8% 和 14.09%。这表明 TIPAS 在面对隐私防御机制时性能波动更小，具备更强的抗干扰能力。

该优势主要源于 TIPAS 采用的 BaggingClassifier 集成学习策略。该策略将多个弱分类器组合成一个强分类器，从而减少模型方差，增强模型的泛化能力。因此，即使在隐私保护措施引入噪声的情况下，TIPAS 仍能较好地捕捉数据中的关键特征，从而在攻击任务中保持较高的鲁棒性与稳定性。相比之下，单一 MLP 模型缺乏此类集成机制，在防御条件下更容易受到干扰，导致攻击性能显著下降。

3.7 Gym 游戏场景下隐私推断性能分析

本节通过在 Gym 平台的 4 个典型环境（CartPole、MountainCar、LunarLander 和 BipedalWalker）中开展实验，进一步验证 TIPAS 的有效性。实验评估了基于不同算法（包括 DQN、PPO、A2C 和 DDQN）的智能体模型在上述环境中的表现，并通过调整关键环境参数构建多个候选场景，以模拟不同的训练条件。各环境的核心任务分别为实现平衡杆稳定（CartPole）、完成小车爬山控制（MountainCar）、操控月球着陆器平稳着陆（LunarLander）及实现双足机器人行走（BipedalWalker）。本实验旨在验证 TIPAS 在典型深度强化学习环境中的隐私推断准确率，进而说明其具有良好的通用性与适用性。实验结果如表 4 所示。

基于奖励均值和方差的 SVM 方法依赖于智能体在非训练环境中表现出的性能差异。对于泛化能力较强的模型（如 PPO），其在多个候选场景中均表现良好，导致基于奖励均值差的 SVM 方法难以准确识别其原始训练环境，推断准确率较低。对于泛化能力较弱的模型（如 A2C），基于奖励均值与方差特征的 SVM 方法的推断准确

表 3 不同隐私保护方法下的隐私推断准确率对比
Table3 Comparison of privacy inference accuracy under different privacy protection methods

网络拓扑	算法模型	DP		Drop_out	
		TIPAS	单一 MLP	TIPAS	单一 MLP
NSFNET	MPNN+DQN	89.84%	75.55%	87.54%	70.24%
	MPNN+DDQN	88.35%	61.97%	86.72%	56.44%
	MPNN+Dueling_DQN	90.12%	69.09%	89.02%	63.66%
	MPNN+PPO	87.64%	64.95%	86.25%	59.66%
GEANT	MPNN+DQN	89.73%	64.01%	87.45%	58.39%
	MPNN+DDQN	90.36%	68.57%	88.50%	63.43%
	MPNN+Dueling_DQN	91.45%	74.70%	88.45%	69.40%
	MPNN+PPO	90.21%	67.72%	88.43%	62.36%
GBN	MPNN+DQN	89.68%	66.86%	87.82%	61.40%
	MPNN+DDQN	88.65%	76.40%	85.59%	70.61%
	MPNN+Dueling_DQN	80.74%	67.42%	79.45%	62.22%
	MPNN+PPO	86.81%	67.92%	84.84%	62.41%
Topology	MPNN+DQN	86.39%	66.12%	84.53%	60.41%
	MPNN+DDQN	83.25%	70.97%	81.41%	65.93%
	MPNN+Dueling_DQN	86.53%	64.83%	83.91%	59.23%
	MPNN+PPO	78.23%	65.95%	76.17%	60.78%

表4 TIPAS 与对比算法推断准确率对比
Table4 Comparison of the accuracy of TIPAS and comparison method

环境	环境参数	算法模型	TIPAS	SVM	单一 MLP
CartPole	Lnegth	DQN	90.88%	46.25%	84.01%
	Force	PPO	92.80%	22.67%	80.62%
	Masspole	A2C	95.37%	71.25%	87.38%
	Masscart	DDQN	89.47%	32.50%	76.69%
MountainCar	Position	DQN	97.00%	22.35%	82.28%
	Speed	PPO	97.50%	25.50%	91.78%
	Force	A2C	86.25%	27.25%	78.42%
	Gravity	DDQN	81.00%	28.75%	71.52%
LunarLander	MainPower	DQN	92.85%	26.25%	84.79%
	SidePower	PPO	93.94%	15.00%	81.58%
	Gravity	A2C	95.34%	62.50%	79.23%
	WindPower	DDQN	93.70%	26.25%	83.57%
BipedalWalker	Friction	DQN	89.90%	35.00%	88.49%
	Force	PPO	81.76%	25.00%	80.29%
	Height	A2C	97.65%	61.25%	79.82%
		DDQN	94.40%	48.75%	87.35%

率相对较高，特别是在 CartPole、LunarLander 和 BipedalWalker 场景中。

此外，实验结果显示，TIPAS 的推断性能整体高于单一 MLP 模型方法。这主要得益于 TIPAS 所采用的集成学习框架，该框架具备更强的泛化能力与环境适应性，能够更好地应对复杂环境及多样化算法带来的识别挑战，从而进一步验证了集成学习在提升隐私推断任务性能方面的优势。

4 结束语

本文针对具有泛化能力的深度强化学习模型所面临的隐私泄露风险，提出了一种基于模仿学习的目标轨迹的隐私推断方法。该方法首先利用给定的目标轨迹，通过模仿学习构建影子智能体；随后在多种候选场景中对该智能体进行测试，收集其行为轨迹数据；接着采用集成学习策略构建多分类攻击模型，并使用所收集的数据对其进行训练，最终实现对目标轨迹所属训练场景的有效推断。实验结果表明，与传统基于奖励均值与方差的 SVM 及单一 MLP 模型的方法相比，本文方法在推断准确率方面具

有明显优势。然而，该方法仍存在不足，包括对高质量数据与先验知识的依赖性较强，以及计算时间开销较大。未来研究将聚焦于提升算法效率、减少对先验知识的依赖，并进一步增强方法的通用性与可解释性。

本文进一步揭示，深度强化学习模型不仅易受隐私窃取攻击，还能高精度还原其训练环境的特定信息。实验结果表明，即便泛化能力优异的深度强化学习模型，其行为轨迹仍蕴含可被辨识的环境特征，这反映了深度强化学习所特有的“轨迹-环境”耦合特性。该发现为设计具备隐私保护功能的深度强化学习算法提供了重要见解，凸显在模型训练与部署中嵌入隐私保障机制的必要性。

参考文献：

- [1] TAN R J, ZHOU J, DU H B, et al. An modeling processing method for video games based on deep reinforcement learning[C]//Proceedings of the 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Piscataway: IEEE Press, 2019: 939-942.
- [2] LIU R R, NAGEOTTE F, ZANNE P, et al. Deep reinforcement learning for the control of robotic manipulation: a focussed mini-review[J]. Robotics, 2021, 10(1): 22.

- [3] IBARZ J, TAN J, FINN C, et al. How to train your robot with deep reinforcement learning: lessons we have learned[J]. *The International Journal of Robotics Research*, 2021, 40(4/5): 698-721.
- [4] PALANISAMY P. Multi-agent connected autonomous driving using deep reinforcement learning[C]//*Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*. Piscataway: IEEE Press, 2020: 1-7.
- [5] ASHRAF N M, MOSTAFA R R, SAKR R H, et al. Optimizing hyperparameters of deep reinforcement learning for autonomous driving based on whale optimization algorithm[J]. *PLoS One*, 2021, 16(6): e0252754.
- [6] FENG S, SUN H W, YAN X T, et al. Dense reinforcement learning for safety validation of autonomous vehicles[J]. *Nature*, 2023, 615 (7953): 620-627.
- [7] 刘圣龙, 黄秀丽, 江伊雯, 等. 面向多方数据融合分析的隐私计算技术综述[J]. *网络与信息安全学报*, 2024, 10(6): 24-36.
- LIU S L, HUANG X L, JIANG Y W, et al. Review of privacy computing techniques for multi-party data fusion analysis[J]. *Chinese Journal of Network and Information Security*, 2024, 10(6): 24-36.
- [8] HENDERSON P, ISLAM R, BACHMAN P, et al. Deep reinforcement learning that matters[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*. Hong Kong: AAAI Press, 2018: 3207-3214.
- [9] COBBE K, KLIMOV O, HESSE C, et al. Quantifying generalization in reinforcement learning[C]//*Proceedings of the International Conference on Machine Learning*. New York: PMLR, 2019: 1282-1289.
- [10] PAN X, WANG W, ZHANG X, et al. How you act tells a lot: privacy-leaking attack on deep reinforcement learning[C]//*Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems*. Berlin: Springer, 2019: 368-376.
- [11] BOJARSKI M, DEL TESTA D, DWORAKOWSKI D, et al. End to end learning for self-driving cars[J]. *arXiv Preprint*, arXiv: 1604.07316, 2016.
- [12] KENDALL A, HAWKE J, JANZ D, et al. Learning to drive in a day[J]. *arXiv Preprint*, arXiv: 1807.00412, 2018.
- [13] KIRAN B R, SOBH I, TALPAERT V, et al. Deep reinforcement learning for autonomous driving: a survey[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(6): 4909-4926.
- [14] NG A Y, RUSSELL S. Algorithms for inverse reinforcement learning[C]//*Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann: San Francisco, 2000: 663-670.
- [15] FINN C, LEVINE S, ABBEEL P. Guided cost learning: deep inverse optimal control via policy optimization[C]//*Proceedings of the 33rd International Conference on Machine Learning*. Cambridge: MIT Press, 2016: 49-58.
- [16] CODEVILLA F, MÜLLER M, LÓPEZ A, et al. End-to-end driving via conditional imitation learning[C]//*Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*. Piscataway: IEEE Press, 2018: 4693-4700.
- [17] 王晓芳, 顾焜仁. 一种深度强化学习与模仿学习结合的突防策略[J]. *宇航学报*, 2023, 44(6): 914-925.
- WANG X F, GU K R. A penetration strategy combining deep reinforcement learning and imitation learning[J]. *Journal of Astronautics*, 2023, 44(6): 914-925.
- [18] FREEMAN C D, METZ L, HA D. Learning to predict without looking ahead: world models without forward prediction[C]//*Proceedings of the 32nd Conference on Neural Information Processing Systems (2019)*. Vancouver: NeurIPS, 2019: 5380-5391.
- [19] 任奎, 孟泉润, 闫守琨, 等. 人工智能模型数据泄露的攻击与防御研究综述[J]. *网络与信息安全学报*, 2021, 7(1): 1-10.
- REN K, MENG Q R, YAN S K, et al. Survey of artificial intelligence data security and privacy protection[J]. *Chinese Journal of Network and Information Security*, 2021, 7(1): 1-10.
- [20] GOMROKCHI M, AMIN S, ABOUTALEBI H, et al. Membership inference attacks against temporally correlated data in deep reinforcement learning[J]. *IEEE Access*, 2023, 11: 42796-42808.
- [21] FRANÇOIS-LAVET V, HENDERSON P, ISLAM R, et al. An introduction to deep reinforcement learning[J]. *Foundations and Trends in Machine Learning*, 2018, 11(3/4): 219-354.
- [22] POPESCU M, BALAS V E, PERESCU-POPESCU L, et al. Multi-layer perceptron and neural networks[J]. *WSEAS Transactions on Circuits and Systems Archive*, 2015, 8(7): 579-588.
- [23] GIRSANG N D. Classification of batik images using multilayer perceptron with histogram of oriented gradient feature extraction[C]//*Proceeding of International Conference on Science and Engineering*. 2021: 197-204.
- [24] 王文尚, 高凡, 侯冰莹, 等. 基于多层感知机的 ATM 英文凭单识别[J]. *信息技术*, 2024, 48(5): 144-148, 154.
- WANG W S, GAO F, HOU B Y, et al. English ATM voucher identification based on multilayer perceptron[J]. *Information Technology*, 2024, 48(5): 144-148, 154.
- [25] ALI-ALNUAIM A, ZAKARIAH M, SHUKLA P K, et al. Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier[J]. *Journal of Healthcare Engineering*, 2022.
- [26] NOOR F, IQBAL M J, YAQOOB S, et al. Depression detection in social media using bagging classifier[J]. *Depression*, 2023.
- [27] BROCKMAN G, CHEUNG V, PETERSSON L, et al. Openai gym[J]. *arXiv preprint*, arXiv: 1606.01540, 2016.
- [28] ALMASAN P, SUÁREZ-VARELA J, RUSEK K, et al. Deep reinforcement learning meets graph neural networks: Exploring a routing optimization use case[J]. *Computer Communications*, 2022, 196: 184-194.
- [29] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[EB]. *arXiv Preprint*, arXiv: 1312.5602, 2013.
- [30] SALH A, AUDAH L, ALHARTOMI M A, et al. Smart packet transmission scheduling in cognitive IoT systems: DDQN based approach[J]. *IEEE Access*, 2022, 10: 50023-50036.
- [31] BAN T W. An autonomous transmission scheme using dueling DQN for D2D communication networks[J]. *IEEE Transactions on Vehicular Technology*, 2020, 69(12): 16348-16352.
- [32] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. *arxiv preprint*, arxiv: 1707.06347, 2017.
- [33] HAGBERG A, SWART P S, CHULT D. Exploring network structure, dynamics, and function using NetworkX[R]. Los Alamos National Lab, Los Alamos, 2008.
- [34] KNIGHT S, NGUYEN H X, FALKNER N, et al. The Internet to-

pology zoo[J]. IEEE Journal on Selected Areas in Communications, 2011, 29(9): 1765-1775.

[35] MIRONOV I .Renyi differential privacy[J]. arXiv preprint, arXiv: 1702.07476. 2017.

[36] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[作者简介]



陈晋音（1982-），女，宁波象山人，浙江工业大学教授，主要研究方向为人工智能安全、图数据挖掘和进化计算。



瞿康贇（2001-），男，安徽天长人，浙江工业大学硕士生，主要研究方向为人工智能安全、深度学习和强化学习。



郑海斌（1995-），男，浙江台州人，浙江工业大学讲师，主要研究方向为深度学习、人工智能安全和图像识别。