面向网络空间防御的对抗机器学习研究综述

余正飞[1, 2] 闫 巧[3] 周 鋆[1, 2]

摘 要 机器学习以强大的自适应性和自学习能力成为网络空间防御的研究热点和重要方向. 然而机器学习模型在网络空间环境下存在受到对抗攻击的潜在风险, 可能成为防御体系中最为薄弱的环节, 从而危害整个系统的安全. 为此科学分析安全问题场景, 从运行机理上探索算法可行性和安全性, 对运用机器学习模型构建网络空间防御系统大有裨益. 全面综述对抗机器学习这一跨学科研究领域在网络空间防御中取得的成果及以后的发展方向. 首先, 介绍了网络空间防御和对抗机器学习等背景知识; 其次, 针对机器学习在网络空间防御中可能遭受的攻击, 引入机器学习敌手模型概念, 目的是科学评估其在特定威胁场景下的安全属性; 然后, 针对网络空间防御的机器学习算法, 分别论述了在测试阶段发动规避攻击、在训练阶段发动投毒攻击、在机器学习全阶段发动隐私窃取的方法, 进而研究如何在网络空间对抗环境下, 强化机器学习模型的防御方法; 最后, 展望了网络空间防御中对抗机器学习研究的未来方向和有关挑战.

关键词 网络空间防御, 对抗机器学习, 投毒攻击, 规避攻击, 对抗样本

引用格式 余正飞, 闫巧, 周鋆. 面向网络空间防御的对抗机器学习研究综述. 自动化学报, 2022, **48**(7): 1625−1649

DOI 10.16383/j.aas.c210089

A Survey on Adversarial Machine Learning for Cyberspace Defense

YU Zheng-Fei[1, 2] YAN Qiao[3] ZHOU Yun[1, 2]

Abstract Machine learning has the ability to learn in various conditions, and becomes a research hotspot and an important direction for cyberspace defense. Unfortunately, machine learning models have potential risks of suffering adversarial attacks in the cyberspace and may become the weakest part of the defense system. Therefore, it is of great benefit to discuss cyberspace defense scenarios and the fundamental issues about the possibility and security of using machine learning algorithms, which is the basis of building cyberspace defense system with machine learning models later on. Adversarial machine learning for cyberspace defense is an interdisciplinary research field. In this paper, we provide a comprehensive review of works related to this filed. Firstly, we present the background and related works of cyberspace defense and adversarial machine learning. Secondly, we provide a model to describe the adversarial model of attack against machine learning in cyberspace defense systems, and thoroughly assess its security attributes under specific threat scenarios. Specifically, we discuss the methods of launching evasion attacks in the test phase, launching poisoning attacks in the training phase, and launching privacy violation in the whole phase for cyberspace defense systems. On the basis of this, we study how to strengthen the machine learning models with different defense mechanisms in cyberspace. Finally, we discuss the future works and challenges of research on adversarial machine learning in cyberspace defense.

Key words Cyberspace defense, adversarial machine learning, poisoning attack, evasion attack, adversarial example

Citation Yu Zheng-Fei, Yan Qiao, Zhou Yun. A survey on adversarial machine learning for cyberspace defense. *Acta Automatica Sinica*, 2022, **48**(7): 1625−1649

近年来, 人工智能、云计算、大数据、物联网、区块链等新一代信息技术突飞猛进, 信息化和工业化深度融合, 各类安全事件和网络攻击频繁发生, 网络空间面临严重的安全威胁. 2016 年, 由恶意软件 Mirai 控制的僵尸网络发起分布式拒绝服务攻击, 造成美国东海岸大范围断网[1]. 2017 年, 勒索病毒软件 WannaCry 通过 “永恒之蓝” MS17-010 漏洞在全球范围大规模爆发, 至少 150 个国家、30 万名用户被感染, 造成经济损失高达 80 亿美元[2]. 2019 年, 中国网络空间研究院编写的《中国网络空间安全发展报告 (2019)》指出: 网络冲突和攻击成为国家间对抗主要形式[3]. 据不完全统计, 中国每年

收稿日期 2021-01-28 录用日期 2021-06-25
Manuscript received January 28, 2021; accepted June 25, 2021
国家自然科学基金 (61976142, 61703416), 湖湘青年英才支持项目 (2021RC3076), 长沙市杰出创新青年培养计划 (KQ2009009) 资助
Supported by National Natural Science Foundation of China (61976142, 61703416), Huxiang Youth Talent Support Program (2021RC3076), and Training Program for Excellent Young Innovators of Changsha (KQ2009009)
本文责任编委 赫然
Recommended by Associate Editor HE Ran
1. 国防科技大学系统工程学院 长沙 410073 2. 国防科技大学信息系统工程重点实验室 长沙 410073 3. 深圳大学计算机与软件学院 深圳 518061
1. College of Systems Engineering, National University of Defense Technology, Changsha 410073 2. Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073 3. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518061

因伪基站、恶意软件勒索等数字犯罪造成的损失达上百亿元[4].

上述事例表明, 网络空间安全不仅影响着国民经济的发展, 还关系着社会的稳定和国家安全. 事实证明, “被动防守永远无法确保安全”, 传统的以固定规则设定的网络空间防御体系变得效率低下, 在面对 “零日”漏洞以及各种高级可持续威胁时常常无能为力, 网络空间防御面临严峻挑战.

与此同时, 计算机算力的显著提升和数据量的日益递增, 带动了机器学习的快速发展, 人工智能迎来第三次发展浪潮. 以深度学习为代表的机器学习作为当前人工智能领域最热门的研究方向之一, 在计算机视觉、语音识别、自然语言处理等方面取得了一系列令人瞩目的研究成果, 成为引领未来的战略性技术.

机器学习技术在处理分类任务和决策问题时展现的突出能力, 成为网络空间防御中应用的新技术. 利用机器学习技术构建 “关口前移, 防患于未然”的积极防御战略, 提高网络空间安全监测和处理的效率, 得到学术界、工业界的广泛关注. 当前, 基于机器学习的网络空间安全研究在系统安全、网络安全及应用安全等层面已有不少解决方案和方法, 在包括垃圾邮件过滤、恶意软件检测、网络入侵检测、漏洞分析与挖掘等领域均取得了不错的效果[4]. 可以预见, 引入机器学习来解决网络空间安全问题是不容置疑的趋势, 机器学习在网络空间防御中的应用前景也会不断扩展.

值得注意的是, 将机器学习应用于网络空间防御问题并非新概念. 近年来, 各主流网络安全公司纷纷尝试利用机器学习改进或重制其安全产品. 然而, 利用机器学习解决网络空间防御问题仍处于初级阶段. 从模型的泛化能力、检测准确度以及实时性来看, 目前的技术解决方案均不能较好满足网络空间防御的应用需求. 导致这一现象的原因一般可概括为机器学习算法的安全性、功能及性能三个方面. 安全性是机器学习在网络空间防御应用首先应解决的基础性问题. 面向网络空间防御的机器学习需要在对抗多变的环境中处理大量数据, 对算法的安全性要求极为严苛. 在算法安全性得到保障的基础上, 机器学习的功能才可以得以实现. 在算法安全性和算法功能实现的基础上, 算法自适应性、可解释性等问题需要加以关注.

不幸的是, 机器学习本身存在易受对抗攻击的安全隐患. 网络空间更是高对抗环境, 无时无刻不在发生力量之间的相互对抗, 机器学习在这样的对抗环境下具有高度的脆弱性, 存在受到对抗攻击的潜在风险, 可能成为网络空间防御体系中最为薄弱的环节, 从而危害整个系统的安全. 例如, 机器学习模型训练过程中利用大量的网络流量、日志信息、系统信号等非结构化数据, 对这些输入数据进行投毒攻击, 会使得模型无法取得良好效果. 此外, 研究显示, 机器学习相关算法能够轻易地被对抗样本操控, 将对抗样本作为输入, 即使其中仅包含人类难以察觉的轻微扰动, 也会导致系统性能明显下降. 利用机器学习解决网络空间安全问题仍是极具挑战性的工作.

为应对以上挑战, 研究人员着手开展对抗机器学习相关研究, 提高机器学习算法在网络空间防御中的鲁棒性, 推动机器学习相关算法的应用[5−10]. 网络空间中广泛存在的对抗使得机器学习的应用面临严峻挑战. 以对抗样本生成和防御为中心的对抗深度学习, 无疑是对抗机器学习领域当前最受关注的研究热点. 然而, 网络空间是 “没有经过勘测的深海”, 科学分析安全问题场景, 从运行机理上展开研究, 探索机器学习算法应用的可行性, 对于机器学习在网络空间防御中的应用大有裨益.

1) 对抗机器学习相关综述. 随着对抗机器学习研究的深入, 相关研究成果不断涌现. 为此, 众多学者展开了对抗机器学习的综述工作, 对该领域进行了归纳与总结, 典型的综述文献及主要内容如表 1 所示. 在已有的综述文献中, 部分综述性工作从总体上对机器学习模型[7, 11−17] 或深度学习模型[18−21] 的对抗攻击和防御展开论述, 如文献 [7] 从对抗机器学习演进路线切入, 论述机器学习的攻防问题, 应用范围包括计算机视觉及网络安全, 文献 [19] 详尽论述深度学习中的对抗样本生成与防御技术; 部分综述性工作[17, 22−23] 围绕对抗机器学习中的隐私攻防进行了论述. 上述综述对通用机器学习模型和深度学习模型的对抗攻防问题进行了论述. 值得注意的是, 机器学习算法在与具体应用领域结合时, 往往具有鲜明的领域特点. 为此, 部分综述性工作围绕机器学习应用于各领域时的攻击与防御问题进行了研究, 主要包括计算机视觉[24−27]、自然语言处理[28]、生物医疗[29−31] 等, 较好地论述了各自领域所带来的新特点与新问题. 同时, 也有部分工作围绕网络空间防御中基于机器学习的入侵检测系统[32]、恶意软件检测[33] 对抗攻击与防御问题进行论述. 然而, 没有相关研究就机器学习在网络空间防御应用中存在的对抗攻击与防御进行综述. 本文在详尽调研与检索相关论文的基础上, 首次从网络空间防御这个角度切入, 开展对抗机器学习综述.

2) 论文选取范围. 本文的综述范围主要是国内

表 1　对抗机器学习相关综述

Table 1　Related surveys about adversarial machine learning

类别	文献题目	主要内容	发表年份
机器学习模型	SoK: Security and privacy in machine learning[15]	分析机器学习模型的攻击面, 系统论述机器学习模型在训练和推断过程中可能遭受的攻击以及防御措施.	2018
	Wild patterns: Ten years after the rise of adversarial machine learning[7]	系统揭示对抗机器学习演进路线, 内容涵盖计算机视觉以及网络安全等领域	2018
	A survey on security threats and defensive techniques of machine learning: A data driven view[12]	从数据驱动视角论述机器学习的对抗攻击和防御问题.	2018
	The security of machine learning in an adversarial setting: A survey[13]	论述对抗环境下, 机器学习在训练和推断/测试阶段遭受的攻击, 提出相应的安全评估机制和对应的防御策略	2019
	A taxonomy and survey of attacks against machine learning[14]	论述机器学习应用于不同领域时的对抗攻击, 主要包括入侵检测、垃圾邮件过滤、视觉检测等领域.	2019
	机器学习模型安全与隐私研究综述[16]	从数据安全、模型安全以及模型隐私三个角度对现有的攻击和防御研究进行系统总结和归纳	2021
	机器学习安全攻击与防御机制研究进展和未来挑战[11]	基于攻击发生的位置和时序对机器学习安全和隐私攻击进行分类, 并对现有攻击方法和安全防御机制进行介绍	2021
深度学习模型	Survey of attacks and defenses on edge-deployed neural networks[18]	论述边缘神经网络的攻击与防御	2019
	Adversarial examples in modern machine learning: A review[19]	论述对抗样本生成与防御技术	2019
	A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and Interpretability[20]	论述深度神经网络(Deep neural network, DNN)的安全与可解释性	2020
	对抗样本生成技术综述[21]	围绕前传、起源和发展三个阶段对对抗样本进行综述	2020
机器学习隐私	机器学习的隐私保护研究综述[17]	着重论述机器学习的隐私保护技术	2020
	A survey of privacy attacks in machine learning[22]	论述机器学习中隐私攻击与保护技术	2020
	机器学习隐私保护研究综述[23]	着重论述机器学习的隐私保护技术	2020
计算机视觉	Threat of adversarial attacks on deep learning in computer vision: A survey[24]	论述计算机视觉中深度学习模型的攻击与防御	2018
	Adversarial machine learning in image classification: A survey towards the defender′s perspective[25]	从防御角度研究计算机视觉分类问题中的对抗机器学习	2020
	Adversarial examples on object recognition: A comprehensive survey[26]	论述神经网络在视觉领域应用时, 存在的对抗样本的攻防问题	2020
	Adversarial attacks on deep learning models of computer vision: A survey[27]	论述计算机视觉中深度学习模型的对抗攻击	2020
自然语言处理	Adversarial attacks on deep-learning models in natural language processing[28]	论述自然语言处理领域中深度学习模型的对抗攻击与防御问题	2020
生物医疗领域	Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective[29]	首次从对抗机器学习角度论述生物识别系统的安全问题	2015
	Toward an understanding of adversarial examples in clinical trials[30]	论述基于深度学习模型的临床实验中的对抗样本问题	2018
	Secure and robust machine learning for healthcare: A Survey[31]	从对抗机器学习的角度概述医疗保健领域中机器学习应用的现状、挑战及解决措施	2021
网络空间防御	Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues[32]	论述入侵检测系统中的对抗攻击问题以及应对措施	2013
	Towards adversarial malware detection: Lessons learned from PDF-based attacks[33]	论述基于机器学习的恶意便携式文档格式 (Portable document format, PDF)文件检测系统可能遭受的对抗攻击	2019

外人工智能和信息安全领域顶级会议和期刊, 包括安全与隐私专题研讨会 (IEEE Symposium on Security and Privacy, S&P)、计算机和通信安全会议 (ACM Conference on Computer and Communications Security, CCS)、安全专题研讨会 (Usenix Security Symposium)、网络与分布式系统安全研讨会 (ISOC Network and Distributed System Security Symposium, NDSS) 等国际信息安全四大会议, 另外还有学习表征国际会议 (International Conference on Learning Representations, ICLR)、知识发现与数据挖掘 (ACM Knowledge Discovery and Data mining, KDD)、神经信息处理

系统年会 (Annual Conference on Neural Information Processing Systems, NeurIPS)、国际人工智能联合会议 (International Joint Conference on Artificial Intelligence, IJCAI)、人工智能大会 (AAAI Conference on Artificial Intelligence, AAAI) 等人工智能领域顶级会议, 以及国内计算机领域部分重要刊物. 另外, 部分预印版论文对该方向研究有较大影响, 本文选取部分质量较好、方法较新、引用较高的预印版论文进行论述.

3) 本综述的主要贡献. 本综述以网络空间防御为问题背景, 全面综述机器学习在网络空间防御中可能遭受的对抗攻击、可采取的防御措施及以后的发展方向. 本综述的主要贡献有两个方面: 一是首次从网络空间防御层面论述机器学习的对抗攻击与防御; 二是探讨了该领域可能的发展方向与存在的挑战.

本文结构如下: 第 1 节介绍网络空间防御及对抗机器学习等背景知识; 第 2 节针对机器学习算法在网络空间防御中可能遭受的攻击进行威胁建模, 目的是科学评估其在特定攻击场景下的安全属性; 第 3 节针对应用于网络空间防御的机器学习算法, 讨论如何实现测试阶段的规避攻击, 训练阶段的投毒攻击以及机器学习全阶段的隐私窃取; 第 4 节论述当前主要采取的防御措施; 第 5 节讨论了当前研究的主要局限性以及下一步研究的可行方向.

1 背景知识

为了更好地阐明机器学习在网络空间防御中的应用, 对网络空间防御和对抗机器学习等背景知识给予简单解释和阐述.

1.1 网络空间防御

网络空间是信息环境中的一个全球域, 是依赖于空中、陆地、海洋和太空的物理域[34]. 1984 年, 美国科幻作家 William Gibson 的长篇小说《神经漫游者》中首次提出网络空间的概念, 意为 "看不到高山荒野, 也看不到城镇乡村, 只有庞大的三维信息库和各种信息在高速流动[35]" 的广袤空间. 2001 年 1 月, 美国首先提出网络空间是 "由独立的信息基础设施组成的全球域, 包括互联网、电信网、计算机系统和各类嵌入式处理器、控制器"[36]. 与之相对应, 国内将网络空间称之为网络电磁空间, 即 "融合于物理域、信息域、认知域和社会域, 以互联互通的信息技术基础设施网络为平台, 通过无线电、有线电信道、信号传递信息, 控制实体行为的信息活动空间". 2015 年, 方滨兴[37] 指出, 网络空间包括 "互联网、通信网、广电网、物联网、社交网络、计算系统、通信系统、控制系统等".

可以看出, 与仅包括局域网、域域网和广域网的传统网络相比, 网络空间作为一个全球域有其独有特性, 包括对环境的部分感知, 动态、离散、极度复杂以及对抗性强等特征. 因此, 网络空间防御能力是关系到国家战略安全的重要指标, 亟需学术界及工业界的关注.

网络空间防御以实现网络空间安全为根本目标, 既维护包括人、机、物等实体在内的基础设施安全, 也维护其中产生、处理、传输、存储的各种信息数据的安全. 广义上, 网络空间防御泛指为实现网络空间安全采取的一切措施. 从狭义上来讲, 网络空间防御专指对计算机、网络以及各种信息的一种保护措施, 通过预防、检测和响应攻击等行为, 实现网络空间安全, 以降低信息基础设施的脆弱性[38]. 当前, 网络空间防御手段主要包括防火墙、防病毒软件、网址过滤、终端检测和响应等, 涉及被动信息保障、主动诱骗、网络空间冲突规避和入侵检测技术等.

然而, 随着时代的发展, 网络空间节点数目呈指数级增长, 节点间的关系错综复杂, 导致网络中各种分布式拒绝服务攻击、撞库攻击、恶意软件、网络入侵等威胁越发复杂, 各种新程序、新技术、第三方/开源代码等应用使得攻击面越发宽广. 一切都在变化, 超越了传统安全防护手段的响应速度, 随之而来的是更大的风险、更高的复杂性, 网络空间防御愈发困难.

1.2 对抗机器学习

机器学习是多学科交叉专业, 属于计算机科学与人工智能的重要分支领域. 1950 年, 人工智能先驱 Turing[39] 在著名论文《计算机器与智能》中, 对图灵测试以及日后人工智能所包含的重要概念进行介绍. 同时, Turing 提出的 "通用计算机能否学习与创新" 问题, 引起各界广泛讨论. 机器学习概念由此引申而来, 并于 1959 年由计算机专家 Samuel[40] 首次提出.

机器学习能够利用经验数据改善系统自身性能, 并获得处理未知数据的泛化能力. Mohri 等[41] 从广义上将机器学习定义为使用经验来提高性能或做出准确预测的计算方法. 在经典的符号主义人工智能范式中, 系统输入的是规则 (即程序) 和需要根据这些规则进行处理的数据, 系统输出的是答案. 利用机器学习, 输入的是数据或从这些数据中预期得到的答案, 系统输出的是规则. 这些规则随后可应用于新的数据, 并使计算机自主生成答案.

机器学习在许多方面取得了巨大的成功, 然而算法本身的安全性问题一直没能得到很好解决. 对抗机器学习的核心思想就是分析机器学习算法在遭受特定攻击时的安全性, 并为设计更安全的学习算法制定适当对策. 如表 2 所示, 该领域已经成为机器学习与计算机安全交叉研究的热点, 并形成了初步的研究体系. Biggio 等[7] 认为该领域的开创性工作可以追溯到 2004 年, Dalvi 等[42] 和 Lowd 等[43−44] 先后针对垃圾邮件过滤问题展开研究. 2004 年文献 [42] 提出对抗分类概念, 指出早期基于机器学习的垃圾邮件检测系统中存在规避分类检测问题. 2005 年 Lowd 等[43] 进一步提出对抗学习概念. 2006 年 Barreno 等[8] 首次较为明确地提出机器学习安全问题的有关概念和知识, 包括机器学习系统的攻击分类和敌手建模, 并在 2010 年就机器学习安全相关概念进行了完善[45]. Dasgupta 等[46] 将对抗机器学习形式化为学习器与敌手围绕不同目标进行的博弈. 其中, 学习器的目标是对数据进行正确预测或分类, 敌手的目标则是诱使学习器对相关的数据做出错误的预测.

近年来, 深度学习成为机器学习领域的重要组成部分, 并有逐渐占据主导地位的趋势. 2014 年, 研究人员发现对抗样本在深度学习系统中普遍存在[47−48], 随后进一步指出深度学习本身存在明显的安全和隐私问题. 以上研究成果迅速激起了人们对深度学习安全性进行研究的兴趣, 越来越多的研究论文开始提出应对威胁的相关对策[49−53], 以使深度学习算法能够抵抗对抗攻击. 这些工作极大地促进了对抗机器学习研究的发展, 并推动对抗机器学习成为当前研究的热点. 围绕对抗机器学习中威胁建模、攻击方法和防御措施等 3 个方面已经做了大量工作, 主要包括: 1) 针对机器学习开展投毒攻击、规避攻击与隐私窃取; 2) 针对前述攻击提出安全评估的系统方法; 3) 设计适当的防御机制以应对这些威胁.

为进一步阐明上述攻防博弈的演变过程, 本文以跨站脚本 (Cross-site scripting, XSS) 攻击检测为典型案例进行分析. XSS 攻击是一种危害极其严重的网络攻击. 攻击者在网页中插入恶意代码, 当用户在引诱下点击包含 XSS 恶意代码的链接或浏览包含 XSS 代码的网页时, 就可能遭到 XSS 攻击.

XSS 攻击通常使用包含 “%” “&” “<” “#”等特殊字符或<*script*><*javascript*>等关键字的攻击

表 2　对抗机器学习时间线

Table 2　A timeline of adversarial machine learning history

年份	主要内容
2004	Dalvi 等[42] 和 Lowd 等[43-44] 研究了垃圾邮件检测中的对抗问题, 提出线性分类模型可能被精心设计的对抗样本所愚弄
2006	Barreno 等[8] 从更广泛的角度质疑机器学习模型在对抗环境中的适用性问题, 并提出一些可行措施来消除或降低这些威胁
2007	NeurIPS 举办 Machine Learning in Adversarial Environments for Computer Security研讨会. 2010年, *Machine Learning*期刊为该研讨会设立同名专题[54]
2008	CCS 举办首届人工智能与安全研讨会 AISec(Workshop on Artificial Intelligence and Security), 并且持续举办至2020年
2012	面向计算机安全的机器学习方法达堡展望研讨会 (Dagstuhl Perspectives Workshop on Machine Learning Methods for Computer Security), 探讨对抗学习和基于学习的安全技术面临的挑战和未来研究方向[55]
2014	KDD 举办安全与隐私特别论坛
2016	AAAI 举办面向网络空间安全的人工智能研讨会 AICS (Artificial Intelligence for Cyber Security), 此后至2019年每年举办一届
2017	为促进对抗样本的相关研究, 谷歌大脑 (Google Brain) 在 NeurIPS2017 上举办对抗攻击与防御挑战赛
2018	NeurIPS2018举办对抗视觉挑战赛, 目的是促进更加鲁棒的机器视觉模型和更为广泛可用的对抗攻击 Yevgeniy 等[6]撰写书籍 *Adversarial Machine Learning*, 并由 Morgan & Claypool 出版社发行
2019	Joseph 等[5]撰写书籍 *Adversarial Machine Learning*, 并由剑桥大学出版社发行 论文Adversarial attacks on medical machine learning[56] *Science* 期刊上发表, 指出医疗机器学习中出现新脆弱性问题, 需要新举措 论文 Why deep-learning AIs are so easy to fool[57] 在 *Nature*期刊上发表, 探讨深度学习遭受对抗攻击时的鲁棒性 KDD2019 举办首届面向机器学习和数据挖掘的对抗学习方法研讨会, 至今已连续举办两届 清华大学和阿里安全于天池竞赛平台联合举办安全 AI 挑战者计划, 至今已有 5 期. 同时, 每年底举办 AI 与安全研讨会, 至今已连续举办两届.
2020	KDD2020 举办首届面向安全防御的可部署机器学习国际研讨会 (Workshop on Deployable Machine Learning for Security Defense)
2021	AAAI2021 举办鲁棒、安全、高效的机器学习国际研讨会 (Towards Robust, Secure and Efficient Machine Learning)

注: 数据更新至2021年2月8日.

载荷. 由于正常用户不大可能使用这些关键字或者字符, 从而可以通过基于规则的过滤器和文本分类器对其进行检测, 并在输出数据之前对潜在的威胁字符进行编码及转义操作, 比如将<*script*>转成空字符串< >. XSS 攻击发动者试图通过恶意混淆以规避检测, 从而规避这些防御措施, 比如将<*script*>替换成<*SCRscriptIPT*>来绕过防御. 针对上述攻击, 研究人员对网站防御系统进行相应地升级改进, 利用设定的规则将<*SCRscriptIPT*>替换成<*SCRIPT*>, 进而成功检测到 XSS 攻击.

之后, 攻击者又利用恶意混淆技术对 XSS 攻击进行改进, 通过编码技术规避检测. 随后, 研究人员迅速提出对数据进行解码操作, 进而实现 XSS 代码反混淆的防御技术. 图 1 显示了混淆代码经过解码还原为原始代码的一个实例[58]. 此后, 基于编码技术的 XSS 攻击有所减少, 但攻击者一直在不断尝试新的伎俩, 以使 XSS 攻击规避系统检测.

```
ca%3D%26id_secteur_activite%3D%26
date_operation%3D%26mots_cles%3D
%22%3E%3Cscript%3Ealert%28%27Xss
%20By%20Atm0n3r%27%29%3C/script
%3E%26x%3D40%26y%3D12
```

(a) 编码后的混淆代码
(a) Encoded obfuscation codes

```
ca=&id_secteur_activite=
&date_operation=&mots_cles=
"><script>alert('Xss
By Atm0n3r')</script>
&x=40&y=12
```

(b) 解码得到的原始代码
(b) Decoded original codes

图 1 混淆代码经过解码被还原为原始代码[58]
Fig. 1 Original codes decoded and restored by obfuscated codes[58]

综上所述, 为解决网络空间中日益复杂的安全问题, 机器学习在网络空间防御中得到广泛应用. 然而, 机器学习技术并没有成为解决此类威胁的最终答案. 网络空间是未来战争的首战场, 针对机器学习的对抗攻击广泛存在. 利用机器学习解决网络空间防御问题, 会引入特定漏洞, 熟练的攻击者可以利用这些漏洞危害整个系统. 在实际问题中, 攻击者必须遵循某些约束. 例如垃圾邮件必须仍然具备传递可用信息的功能, 部署在主机的恶意软件必须能够正确运行并利用存在的漏洞. 在某些情况下, 相关约束使得最优攻击的计算比较困难.

2 机器学习威胁建模

机器学习面临容易遭受对抗攻击的问题, 从机器学习性能角度考虑, 可以看作模型鲁棒性或泛化能力不足; 但是从安全角度考虑, 其实所谓 “安全” 的概念是模型行为超出意料, 让模型原本的设计者手足无措, 因此可以认为是存在 “潜在威胁”.

机器学习系统行为要符合预期目标, 就要对模型可能面临的威胁有全面了解. 对系统进行威胁建模是对攻击者具有的敌手目标、知识和能力进行合理的假设, 是从不同维度刻画潜在敌手的有效保证. 然而, 目前为止, 大多数机器学习系统在设计时只对应了一个很弱的敌手模型. 在这样的前提条件下设计出来的机器学习系统, 只能运行于不存在攻击的封闭环境之中. 在开放的网络空间环境中, 攻击者既可以干扰机器学习训练过程, 训练出不符合预期要求的模型; 也可以利用对抗样本让训练良好的机器学习模型预测错误.

在网络攻防博弈中, 无论是攻击方还是防御方的行为和策略都在动态变化. 本质上, 两者之间是此消彼长的博弈过程, 仅仅根据防御方的防御能力和防御配置对其安全效能进行评估是非常片面的. 因此, 对应于较弱威胁场景设计的机器学习系统很难应用于网络空间防御领域.

在统一的研究框架之下, 分析对抗机器学习中蕴含的假设, 对当前所面临的攻击和可能出现的攻击进行明确分类, 并提前预设防御措施, 是威胁建模的意义所在. 2006 年, 文献 [8] 首次提出对机器学习攻击进行分类, 主要区分攻击影响、安全违规、攻击指向性 3 个维度. 之后, Yevgeniy 等[6]提出从攻击时机、攻击信息、攻击目标 3 个维度对敌手模型进行刻画. 如表 3 所示, Biggio等[7, 10, 59]多次对敌手模型进行改进, 提出从敌手能力、敌手目标、敌手知识维度进行刻画, 能够适用于具有不同攻击策略的应用场景, 更为科学合理. 本文在论述时主要沿用该模型.

2.1 敌手目标

如表 3 所示, 敌手目标包括完整性违规、可用性违规以及隐私窃取. 完整性违规是指在不影响合法用户对系统功能正常使用的同时, 达到规避检测的目的. 比如针对 PDF 文件, 在良性 PDF 文件中注入恶意软件检测器无法检测到的负载, 虽然用户可以正常阅读 PDF 文件内容, 但后台已经在执行各种恶意操作. 可用性违规是指危害合法用户对系统功能正常使用, 甚至造成系统对用户的拒绝服务. 比如, 在恶意 PDF 检测中, 攻击者注入多个良性的脚本指令以触发假警报, 从而造成系统宕机停转, 不能正常运行. 隐私窃取是指对机器学习算法进行

表 3　基于威胁建模的机器学习攻击分类

Table 3　Classfication of attacks against machine learning based on threat model

敌手能力	敌手目标			敌手知识
	模型完整性	模型可用性	隐私窃取	
测试数据	规避攻击	—	模型提取 模型反演 成员推断	白盒攻击 黑盒攻击
训练数据	投毒攻击(后门攻击)	投毒攻击(油蛙攻击)	模型反演 成员推断	白盒攻击 黑盒攻击

反向工程, 获取系统的模型参数、用户的信息或训练数据等隐私. 例如, 攻击者可以通过添加文本、添加脚本代码等方式, 增量更改 PDF文件的某些特征, 观察目标分类器对各种更改产生的反应, 从而推断目标相关的隐私信息.

2.2　敌手知识

敌手知识是指攻击者掌握的机器学习模型的相关知识, 可能包括以下要素: 1) 训练数据 $\boldsymbol{\mathcal{D}}$, 包含 n 个训练样本; 2) 特征集 $\boldsymbol{\mathcal{X}} \subseteq \mathbf{R}^d$, 主要指预处理和特征提取算法以及提取的特征类型; 3) 模型结构和参数 $f : \boldsymbol{\mathcal{X}} \mapsto \mathbf{R}$, 包括决策函数以及在训练过程中的最小目标函数、超参数, 甚至是训练得到的模型参数. 从而, 可以用由基本元素 $\theta = (\boldsymbol{\mathcal{D}}, \mathcal{X}, f)$ 组成的抽象空间 $\boldsymbol{\Theta}$ 描述攻击者对目标系统的知识. 如表 3 所示, 本文根据敌手知识将攻击划分为白盒攻击和黑盒攻击, 主要区别在于两者对目标模型具有不同的访问权限.

1) 白盒攻击. 假设攻击者可以完全获取目标模型的结构和参数等信息, 白盒攻击可以表示为 $\theta = (\boldsymbol{\mathcal{D}}, \mathcal{X}, f)$, 这种情况在实际中很少发生. 然而, 柯克霍夫原则认为 “密码系统应该就算被所有人知道系统的运作步骤, 仍然是安全的[60]”. 依赖于模型结构和参数的机密性而建立的安全性是危险的策略, 因为如果这些机密被暴露, 系统的安全性将立即受到损害. 与现代加密方法类似, 安全的机器学习系统需要尽量减少对机密性的依赖. 研究白盒攻击可以观察系统遭受攻击时性能下降的上限, 如果机器学习系统在白盒攻击中表现良好, 自然也能抵御其他任何攻击. 利用白盒攻击对机器学习系统进行考量, 有利于理解攻击对机器学习系统所能产生的最坏影响, 进而从根本上进行综合评估并加以改进.

2) 黑盒攻击. 假设攻击者无法得知目标模型采用的训练数据, 也无法获取模型结构和具体的参数, 只能获取模型的最终决策结果, 黑盒攻击可以表示为 $\theta = (\widehat{\boldsymbol{\mathcal{D}}}, \widehat{\mathcal{X}}, \widehat{f})$. 在这种情况下, 攻击者拥有的敌手知识最弱, 因此如果能对模型成功实施黑盒攻击, 则该模型的防御能力很弱. 实际上, 黑盒并非完全 “黑”, 攻击者可以利用对目标系统的少量知识来分析模型的脆弱性. 例如, 对于分类函数 f, 攻击者必须知道是用于垃圾邮件过滤还是恶意软件检测. 同时, 也应该知道为了成功实施攻击, 需要对样本进行哪种类型的转换. 因此, 攻击者应该知道包括使用的特征种类、执行静态还是动态分析等基本信息. 同样, 对于训练数据 $\boldsymbol{\mathcal{D}}$, 如果目标系统用于恶意软件检测, 虽然攻击者不知道确切的权重等训练参数, 但可以推断出训练样本是良性软件或是恶意软件. 也就是说, 可以利用目标系统的反馈对决策和标签进行推断.

需要说明的是, 文献 [7, 33] 在白盒攻击和黑盒攻击之间引入了灰盒攻击, 用以描述攻击者掌握目标模型部分结构和参数的状态. 本文在综述过程中, 未在网络空间防御应用中发现有关灰盒攻击的代表性论文, 因此在后续内容中不对灰盒攻击进行论述.

2.3　敌手能力

敌手能力指攻击者对机器学习模型的操作权限. 如表 3 所示, 根据敌手操纵数据时的约束, 可以分为对训练阶段进行 “毒化”的投毒攻击, 包括对具有特定后门触发器的输入样本做出特定错误预测的后门攻击, 测试阶段试图规避模型检测的规避攻击, 以及可以在机器学习全过程进行的隐私窃取. 以恶意软件检测问题为例, 规避攻击要求攻击者仅可访问用于测试的网络流量传输控制协议转储, 而投毒攻击则赋予攻击者访问用于区分攻击行为与正常行为的模型的权限. 针对上述情况, 敌手能力可以形式化为优化问题中的约束条件.

3　攻击方法

机器学习安全性的一个重要来源在于设计和训练机器学习模型时所做的错误假设, 即: 模型的训练和运行都在理想化条件下, 训练集和测试集服从独立同分布特性[61], 且机器学习不存在遭受对抗攻击的威胁. 当前, 基于 “正常人”在 “正常行为”中产

生 “正常数据”得到的机器学习模型能够很好地投入应用. 然而, 网络空间攻防博弈异常激烈, 安全威胁时刻存在. 耗费巨大代价构建的机器学习模型在攻防过程中可能漏洞频出、不堪一击. 例如大部分宣称实际效果很好的商业化产品, 往往都对前提条件进行了严格限制, 但在实际之中又难以达到这些限制条件. 如图 2 所示, 通过研究对抗机器学习问题, 从根本上建立科学有效的防御机制, 提高机器学习算法的安全性, 是对抗机器学习的意义所在. 攻击者可能通过不同途径获得模型的算法、参数以及内部结构等相关信息, 从而对机器学习展开攻击, 进而降低模型性能甚至破坏模型可用性. 需要注意的是, 攻击者包括 “人为的有意攻击”以及 “受系统或环境所限, 数据中可能产生的潜在的无意攻击”.

与一般领域不同, 用于网络空间防御的机器学习的显著特点是对安全性要求十分严格. 比如, 利用机器学习构建新闻推荐系统, 效果不好最多也就是给用户推荐了一些不感兴趣的内容, 并不会造成太大损失. 而在网络空间防御中则完全不同, 利用机器学习技术进行入侵检测, 无论是误报或漏报, 对用户来说都会造成实际的或潜在的严重损失. 在网络空间部署基于机器学习的防御系统时, 必须设计符合现实情况的攻击并将其作为测试基准, 开展科学合理的安全评估. 如表 4 所示, 本节围绕攻击发生的阶段不同, 区分规避攻击、投毒攻击和隐私窃取 3 个部分对攻击方法进行论述, 可以涵盖现有攻击方法. 其中, 规避攻击发生于机器学习测试/推断过程, 投毒攻击发生于训练过程, 隐私窃取则发生于机器学习全阶段, 可以为顺利实施规避攻击或投毒攻击提供丰富的先验知识, 达到破坏机器学习安全性的目的.

3.1 规避攻击

规避攻击是指攻击者对恶意样本进行特定修改, 从而改变分类器决策、规避系统检测, 进而对机器学习系统的安全性带来严重威胁. 在网络空间防御中, 规避攻击最早被用于垃圾邮件检测[42], 随后扩大至恶意 PDF 文件检测[62] 和网络入侵检测[63] 等领域. 用于执行规避攻击的样本通常被称为对抗样本, 与离群值或噪音相比, 对抗样本主要有两方面特点: 一是大多数对抗样本由攻击者特意创建用来攻击机器学习模型, 通常具有人造的属性; 二是通常要求对抗样本 “不易察觉”, 根据应用领域的不同, 对 “不易察觉”的具体要求也不尽相同. 在计算机视觉中, 可以通过限制修改像素的数量或者图片改变的幅度, 从而使得生成的对抗样本不易被人类视觉所分辨. 在网络空间防御中则有其他要求, 比如, 针对恶意软件检测要求生成的对抗样本能够保留恶意软件的特定功能, 电子邮件仍能正常发送及阅读等等.

针对网络空间防御问题, 实现规避攻击的方式各不相同, 本文将规避攻击区分为基于模仿、基于梯度、基于迁移等 3 大类. 其中, 基于模仿的规避攻击采用启发式算法, 直接修改文件结构或内容. 基于梯度的规避攻击利用梯度下降求解优化问题, 对输入样本执行细粒度的修改, 以最小化 (最大化) 样本被归类为恶意 (良性) 的概率. 基于迁移的规避攻击主要利用了对抗样本的跨模型迁移性, 可以应用于无法获取模型梯度的各种攻击场景.

3.1.1 基于模仿的规避攻击

模仿是社会学习的重要形式之一. 在学术研究中, 模仿也是一种很好的启发和借鉴方式. 如图 3

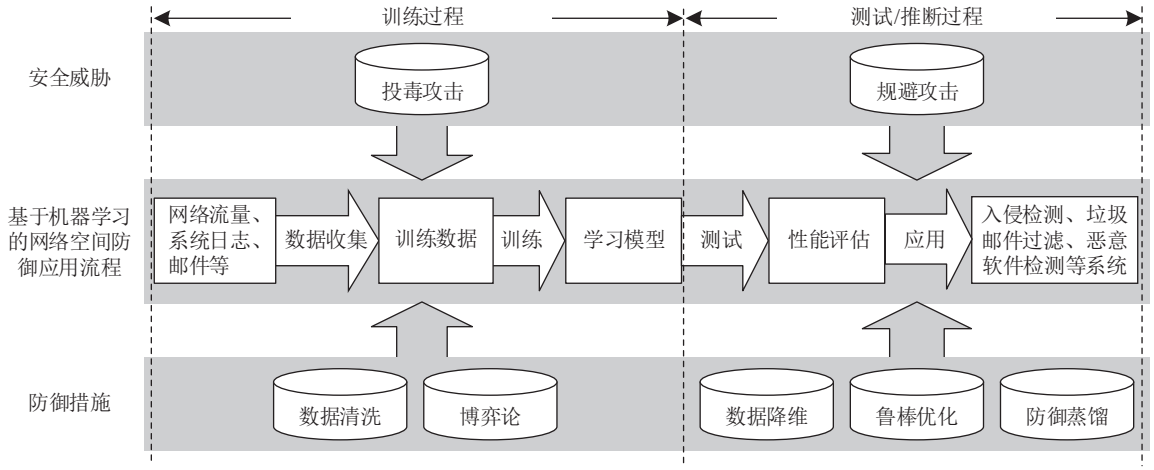


图 2 网络空间防御中的对抗攻击与防御措施

Fig. 2 Adversarial attack and defense methods for cyberspace defense

表 4　网络空间防御中的典型对抗攻击

Table 4　Typical adversarial attacks for cyberspace defense

攻击方法		相关文献	应用领域	特点
规避攻击	基于模仿的规避攻击	[42, 44, 64–66]	垃圾邮件检测	模仿攻击采用启发式算法, 尝试向恶意文件中添加良性特征或者向良性文件中注入恶意特征, 从而实现规避
		[67]	流量分析	
		[68]	恶意软件检测	
		[62, 69–75]	恶意 PDF 文件分类	
	基于梯度的规避攻击	[75–77]	恶意 PDF 文件分类	基于梯度的规避攻击利用梯度下降求解优化问题, 对输入样本执行细粒度的修改, 以最小化 (最大化) 样本被归类为恶意 (良性) 的概率
		[9, 78–79]	恶意软件检测	
		[63, 80]	入侵检测	
	基于迁移的规避攻击	[70, 81]	恶意 PDF 文件分类	基于迁移的规避攻击主要利用了对抗样本的跨模型迁移性, 可以应用于无法获取模型梯度的各种攻击场景
		[82–84]	入侵检测	
		[85]	XSS 检测	
		[86]	域名生成	
		[87–89]	恶意软件检测	
投毒攻击	可用性攻击	[8, 44, 90–92]	垃圾邮件检测	可用性攻击的目的是增加测试阶段的分类误差, 从而造成拒绝服务
		[93–94]	入侵检测	
	完整性攻击	[95–96]	异常检测	完整性攻击的目的是使得恶意软件特定子集被模型误分类
		[97–98]	恶意软件检测	
隐私窃取	模型提取攻击	[99]	—	隐私窃取主要目的是窃取机器学习模型或训练数据的信息
	模型反演攻击	[100–101]		
	成员推断攻击	[102–103]		

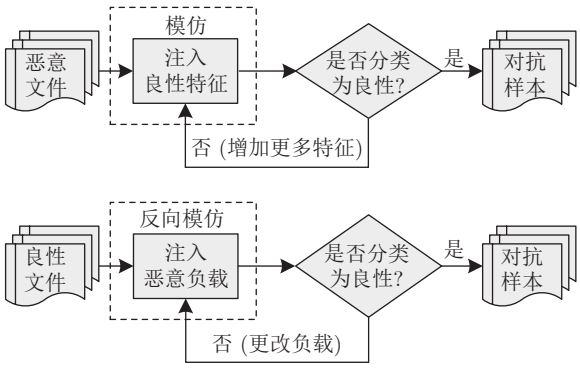


图 3　模仿攻击 (上图) 和反向模仿攻击 (下图)[33]

Fig. 3　Mimicry attacks (top) and reverse mimicry attacks (bottom)[33]

所示, 具体到规避攻击, 还可以分为模仿攻击和反向模仿攻击. 前者向恶意样本添加良性样本的特征, 后者向良性样本中添加恶意样本的特征, 使得修改后的样本在实现恶意功能的同时, 将结构差异降至最低, 从而规避分类器的检测[33]. 模仿攻击在实现方法上比较简单、对模型没有特定依赖性, 广泛适用于各种场景.

早期的模仿攻击主要针对垃圾邮件检测问题, 分类器也大多是相对简单的朴素贝叶斯和最大熵. 2004 年, John Graham-Cumming 在麻省理工学院举办的垃圾邮件大会上率先提出针对贝叶斯分类器的 “好词”攻击思想. 随后, 文献 [44, 64] 初步尝试使用临时策略规避基于统计的垃圾邮件过滤器. 文献 [42–44] 以垃圾邮件过滤器为例, 研究用于线性分类器的规避攻击, 在保持垃圾邮件可读性的同时, 对垃圾邮件内容进行尽量少的修改. 主要手段包括好词插入攻击[44, 64], 核心思想是在垃圾邮件中插入可能出现在合法电子邮件中的 “好词”; 以及 “坏词”混淆攻击[65–66], 核心思想是对经常出现在垃圾邮件中的 “坏词”进行混淆处理. 同一时期, 类似模仿攻击的还有针对网页识别的流量变型[67], 主要思想是模仿检测系统关注的流量特征. 结果表明, 利用攻击样本对贝叶斯分类器进行攻击, 可以使流量检测分类器对网页识别的准确率从 98% 降低到 4%.

随着技术的不断发展, 模仿攻击也逐渐应用于更为复杂的恶意软件检测领域. Smutz 等[62] 针对恶意 PDF 文件检测首次提出模仿攻击, 主要思想是在白盒场景下, 将对于目标分类器最具有判别性的特征注入测试样本, 使其特征空间中的条件概率分布或距离尽可能接近良性样本. 结果表明, 只要在恶意软件样本中注入前 6 个判别特征, 就可以使基于随机森林算法的恶意 PDF 文件检测系统 PDFRate 的检测准确率降低 20% 左右. 随后, Šrndić 等[69] 对其进行改进, 通过注入对分类器决策影响最大的特

征来测试基于径向基函数支持向量机 (Support vector machine, SVM) 的恶意 PDF 文件检测系统 Hidost的鲁棒性, 并成功误导分类算法.

然而, 上述文献的分析仅限于特征空间的操作, 并没有创建出真实可用的恶意 PDF 文件样本. 为克服上述局限性, Šrndić 等[70] 在真实场景下实现对 PDFRate 的规避攻击. 如图 4 所示, PDF 文件主要包括文件头、文件体、交叉引用表和文件尾 4 个部分. 文献 [70] 的主要思想是在交叉引用表和文件尾之间插入特定内容, 从而影响 PDFRate 的分类输出; 同时, 该方法可以使普通 PDF 阅读器跳过插入的内容对文件进行解析. 值得注意的是, 文献 [70] 指出, 尽管这种注入策略能够有效规避目标系统, 但通过改进解析过程可以很容易使其失效. Suciu 等[71] 也采用类似的思想, 将字节添加到恶意二进制文件末尾或空白区域, 从而创建对抗样本.

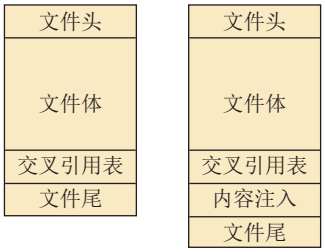


图 4 原始 PDF 文件 (左图) 和修改后的 PDF 文件 (右图)
Fig. 4 The original (left) and modified (right) PDF file

在黑盒场景下, 最简单的模仿攻击是将良性 PDF 文件的全部内容复制到恶意样本. Corona 等[72] 将多个良性样本的 *JavaScript* 应用程序接口 (Application programming interface, API) 特征添加到恶意 PDF 文件中, 并观察恶意样本检测率随注入特征增加产生的变化. 结果表明, 由于 *JavaScript* API 在本质上是动态的特征, 利用这种方法产生的样本仍然能在真实环境中被 *Lux0R* 检测. Šrndić 等[69] 也采用了类似的攻击方法, 结果表明, Hidost 对这种模仿攻击具有较高的鲁棒性. 类似的工作还有, Rosenberg 等[68] 在软件中注入不会更改其恶意功能的虚假 API 调用, 成功规避动态恶意软件检测模型. 反向模仿也被用于具体的攻击, Maiorca 等[73] 将恶意内容注入良性 PDF 文件, 成功规避恶意 PDF 文件检测系统 PDFRate. 这种方法的缺点是只能作用于具有简单载荷的恶意 PDF 文件.

大部分模仿攻击主要是向样本中添加特征, 然而, 也有部分研究成果考虑对样本进行特征删除从而实现模仿攻击. 2016 年, Xu 等[74] 提出基于遗传编程算法的黑盒攻击, 从注入和删除特征的角度对恶意 PDF 文件进行修改. 实验中, 利用布谷鸟沙盒将样本与原始种子的动态行为进行比较, 验证文件可用性. 结果表明, 从 Contagio 恶意 PDF 文件数据集中选出的 500 个恶意样本种子, 可以产生近 17K 个规避样本, 对 PDFrate 和 Hidost 等恶意 PDF 文件分类器的规避率高达 100%. 该攻击首次考虑从 PDF 文件中删除对象, 但它对计算性能的需求很大, 利用以上方法实现规避攻击的场景非常有限. 主要有两方面原因: 一是遗传算法不利用梯度等任何结构化信息来驱动优化空间中的搜索, 因此优化过程中需要执行大量查询; 二是攻击需要对操纵的样本进行动态分析. 如果攻击者的目标仅仅是成功实现攻击, 那么这并不是什么问题, 但基于遗传算法的规避攻击可能不适合在时间有限、计算资源有限、对目标系统查询次数有限的场景生成攻击.

3.1.2 基于梯度的规避攻击

模仿攻击主要适用于结构和参数较为简单的机器学习模型, 往往也只能生成次优的样本. 同时, 模仿攻击需要准确地知道应该修改哪些特征, 在适用性上表现出明显不足. 随着计算机算力的提升以及机器学习的发展, 研究人员寻求使用不同的攻击方法修改测试样本, 使其更好规避机器学习模型检测. 其中, 梯度下降是一种有效且使用相对较多的优化算法, 适用于大多数分类器. 该方法应用梯度下降法, 在损失函数最可能增加的方向上扰动测试样本, 使恶意样本向着合法样本的区域移动, 从而被分类器错分为正常样本. 由于正常样本和恶意样本的特征差别较大, 直接修改恶意样本有时不能有效规避分类器的检测. 因此, 通常会先计算所有正常样本特征属性的平均值, 然后使用梯度下降的攻击方法生成对抗样本规避分类器的检测, 降低分类器的可用性.

2013 年, Biggio 等[75, 77] 在白盒场景下针对恶意 PDF 文件检测器提出基于梯度的攻击方法, 重点讨论了如何规避恶意 PDF 文件检测系统 Slayer[104], 虽然仅针对线性支持向量机、径向基函数支持向量机和神经网络等三种不同的学习算法进行测试, 但理论上该方法适用于具有可微判别函数的任何分类器. 实验结果表明, 不管是线性恶意软件检测器还是非线性恶意软件检测器, 同样容易受到规避攻击. 之后, Smutz 等[76] 针对 PDFRate 进行了相同的攻击测试, 结果表明, 基于梯度的攻击可以完全避开系统的检测.

2014 年, Szegedy 等[47] 寻求对图像添加视觉“不易察觉”的轻微扰动, 同时引起机器学习的错误

分类. 在该思想的指导下, 利用限制内存 BFGS (Broyden-Fletcher-Goldfarb-Shanno) 方法成功创建对抗样本. 随后, 研究人员陆续提出快速梯度符号法 (Fast gradient sign method, FGSM)[48]、基于雅可比矩阵的显著图攻击算法 (Jacobian-base saliency map attack, JSMA)[105]、C&W[106] (该算法以 Carlini 和 Wagner 两位作者名字首字母命名) 等对抗样本生成方法, 甚至对实际部署的人脸识别系统实现了成功攻击[101, 107]. 鉴于对抗样本在计算机视觉领域取得的丰富成果, 研究人员开始考虑将计算机视觉中表现良好的攻击方法迁移到网络空间防御问题.

在将扰动攻击方法从计算机视觉迁移到网络空间防御的具体过程中, 存在新问题, 面临新挑战: 一方面, 计算机视觉中连续的、可微的输入, 被网络空间中离散的、通常是二进制数值的输入代替, 从而使得特征只能被完全添加或删除, 而不能部分更改; 另一方面, 网络空间防御对生成的样本提出了新的约束, 规避攻击的有效实施不应更改或破坏样本的恶意功能. 在计算机视觉中, 可以利用技术手段限制对抗样本与原始图像之间的距离, 从而使得产生的样本对人类视觉来说保持不变. 在网络空间中则必须保证修改后保留其语义和恶意功能, 这与以往计算机视觉中对抗样本的应用形成了对比.

Grosse 等[9] 在这方面首先取得突破. 2016 年, 他们通过改进 JSMA 算法, 生成 Android 恶意软件的对抗样本, 对基于神经网络的恶意软件检测模型进行白盒攻击, 相关实验表明, 对于各种规模的神经网络, 对抗样本可以使模型错误分类率高达 40% ~ 84%. 该研究收录于次年的欧洲计算机安全研究研讨会. 理论上, 该方法可以推广应用于任何可微的机器学习分类器. 然而, 这项工作仅在特征空间中进行攻击, 并且假设可以生成与任意特征向量匹配的样本, 在此过程中不会生成真实的恶意文件, 可用性较低.

之后, 陆续又有较多成果涌现: Kolosnjaji 等[78] 和 Kreuk 等[79] 对 FGSM 进行改进, 通过在恶意软件样本末尾追加字节的方式创建对抗样本. Huang 等[80] 利用 FGSM 和 JSMA 算法对基于多种深度学习方法的入侵检测系统实现攻击. Clements 等[63] 借鉴 FGSM、JSMA、C&W 和弹性网络算法[108] 等 4 种不同攻击算法攻击基于深度学习的网络入侵检测系统, 结果表明, 利用 4 种方法都能成功生成对抗样本.

3.1.3　基于迁移的规避攻击

随着机器学习云服务的发展, 机器学习算法通常被集成到防病毒软件或入侵检测应用等网络空间防御系统, 并以托管的方式在云端运行. 攻击者无法直接获取模型参数, 但可以通过查询输入的方式与模型交互, 并获得表征分类置信度的分数或简单的二元标签等输出. 在这样的情况下, 对于攻击者来说, 目标模型是一个黑盒模型, 对机器学习的攻击是黑盒攻击. 此时, 创建对抗样本的难度有所增加.

当前, 对黑盒模型的规避攻击主要基于对抗样本跨模型迁移性. 如图 5 所示[109], 比例越高表明有效的对抗样本越多. 由图 5 可以看出, 对于 DNN、SVM、逻辑回归 (Logistic regression, LR)、决策树 (Decision tree, DT)、邻近算法 (k-Nearest neighbor, kNN) 以及集成学习 (Ensemble learning, Ens.) 等具有不同结构以及在不同数据集上训练的模型, 对抗样本仍然能够保持有效性. 这意味着攻击者可以通过构建代理模型的方式, 利用已知机器学习模型构造对抗样本, 然后攻击相关的未知模型. 具体来说, 攻击者首先给黑盒模型输入各种特征的样本, 然后观察模型对不同样本在输出上产生的差异, 进而根据差异对模型的决策边界进行估计, 利用估计的边界生成代理模型, 最后针对该代理模型生成对抗样本. 由于对抗样本具有可迁移性, 仍然会导致原始模型错误分类.

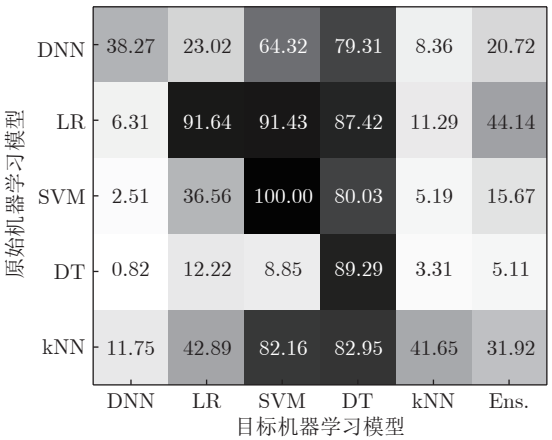


图 5　跨模型迁移矩阵[109]

Fig. 5　The cross-model transferability matrix[109]

Šrndić 等[70] 利用梯度下降核密度估计在基于 SVM 的代理分类器上生成对抗样本, 进而对 PDFRate 发动黑盒攻击. 实验结果表明, 梯度下降核密度估计攻击能够将 PDFRate 的输出分数降低 29% ~ 35%, 验证了黑盒攻击对 SVM 和随机森林分类器的有效性. 在网络空间防御应用中, 将样本输入基于机器学习的恶意软件检测系统, 可以得到的输出信息往往只有样本类别, 即样本为恶意或良性的分类标签. 在此情形之下, 恶意软件检测系统属于二

元黑盒, 对攻击者来说是最具挑战性的情况. Dang 等[81] 提出基于爬山算法, 针对恶意 PDF 文件分类器发动黑盒攻击, 通过对改变检测器和测试器决策所需变形步骤的数量进行量化, 将二元黑盒问题转换成产生分数的黑盒问题, 进而实现规避攻击.

2014 年, Goodfellow 等[110] 提出生成对抗网络 (Generative adversarial networks, GAN), 在机器学习领域得到广泛应用[111]. 生成对抗网络由生成器和判别器构成, 生成器通过学习真实的数据分布生成逼真的数据, 判别器用于判别数据是否真实. 生成对抗网络是另一种基于迁移的规避攻击方式. 与前述通过添加微小扰动生成对抗样本的方法不同, 生成对抗网络采用博弈论中的纳什均衡思想, 通过多回合的对抗训练, 使得生成器学到目标样本分布, 从而用于规避攻击. 研究人员发现, 利用 GAN 生成的对抗样本效果拔群, 随后各种基于 GAN 的对抗样本生成方法在网络空间防御中被广泛应用, 比较有代表性的包括恶意流量生成网络[82]、自适应常规流量生成网络[83]、拒绝服务攻击迹线生成网络[84]、基于深度学习的域名生成算法 (Deep learning based domain generation algorithms, DeepDGA)[86]、恶意软件样本生成网络 MalGAN[87]、安卓恶意软件样本生成网络[89]. 其中, DeepDGA 利用 GAN和长短期记忆网络构建模型, 生成的随机域名难以被随机森林等机器学习算法检测. MalGAN 对静态 API 特征调用进行分析, 并且构造了稀疏的二进制特征以指示程序调用 API, 通过对目标模型进行探测, 从而创建一个完全可微的代理模型. 然后, 在改进的 GAN 中使用代理模型进行梯度计算, 生成 API 调用序列的恶意软件对抗样本, 用来规避恶意软件检测系统. 与传统的基于梯度的对抗样本生成算法相比, MalGAN 算法能大大降低系统检测率.

近年来, 强化学习也被用来实现基于迁移的规避攻击. Anderson 等[88] 提出基于强化学习的通用框架, 用于攻击静态可移植可执行 (Portable executable, PE) 恶意软件检测系统. 该框架不要求模型具有可微性, 也不需要恶意软件检测系统在检测样本时能够对样本进行评分. 如图 6 所示, 智能体与恶意软件检测系统进行多轮博弈, 学习一组可能使恶意软件样本实现规避检测的操作集合, 进而实现对静态 PE 恶意软件检测系统的黑盒攻击, 并直接产生可规避的恶意软件样本. 将基于强化学习得到的恶意软件样本上传到在线查毒网站 VirusTotal, 发现与原始样本相比, 均值检测率呈现明显下降. 文献 [88] 利用强化学习生成恶意软件对抗样本的独特之处包括: 1) 生成的对抗样本能够在真实环境正常运行并具有恶意功能; 2) 成功实现对黑盒检测

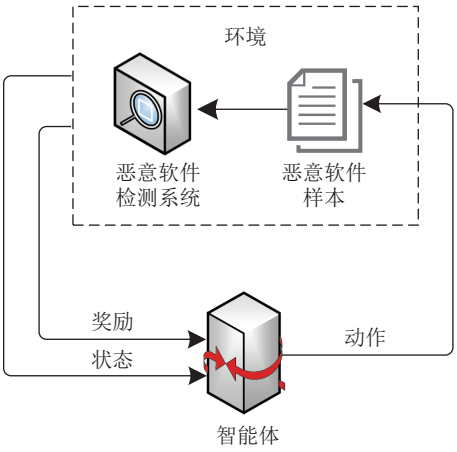


图 6 基于强化学习的恶意软件规避框架

Fig. 6 Framework of malware evasion based on reinforcement learning

系统的规避. 类似的研究还有, Fang 等[85] 提出基于强化学习的 XSS 攻击对抗样本生成算法, 在规避黑盒和白盒检测的同时, 能够保留样本的攻击特征.

3.2 投毒攻击

投毒攻击主要是对训练数据进行修改或者以添加少量数据的方式对训练数据进行污染, 其本质是针对训练数据寻求全局或局部分布的扰动, 进而破坏模型可用性和完整性, 降低机器学习系统性能. 当前, 大多数工作集中于利用对抗样本实现规避攻击, 对投毒攻击的研究却相对少得多, 主要原因是投毒攻击需要解决双层优化问题, 也就是比较具有挑战性的非凸优化问题. 在凸优化问题中, 局部最优解同时也是全局最优解, 在一定意义上, 这个特性使凸优化问题更易于解决; 相比之下, 一般的非凸优化问题更难以解决.

投毒攻击是对主流机器学习算法的一种重要威胁, 包括 SVM、DT、贝叶斯分类器和层次聚类等常规算法都受到了投毒攻击的危害. 投毒攻击属于因果攻击, 具有两方面特点: 一是, 投毒攻击发生在机器学习训练之前或训练过程中, 其造成的危害普遍难以解决; 二是, 机器学习的性能在很大程度上取决于训练数据的质量, 投毒攻击可以从根本上制约机器学习的性能. Kearns 等[112] 研究表明, 当模型的预测误差小于 ε 时, 修改训练集的最大容忍度概率 b 应满足 $b \leq \varepsilon/(\varepsilon+1)$. 特别是, 随着大数据技术的突飞猛进, 网络空间中各种开放的、共享的数据被用来训练机器学习算法, 以机器学习即服务兴起为代表, 各种带有错误标签或有偏差的数据在有意或无意之中被添加到训练数据中, 因而降低了训练数据整体质量. 在数据赋能的网络空间防御服务中,

攻击者以影响机器学习服务性能为目标, 针对数据库发起投毒攻击, 可以实现 “立竿见影”的效果, 同时得到极为丰厚的回报, 使得投毒攻击频繁发生. 比如, 全球领先的杀毒公司卡巴斯基实验室, 被指控向 VirusTotal 注入假阳性样本, 具体来说就是将良性的、系统关键文件标记为恶意文件[113]. 当卡巴斯基的竞争对手在新样本上训练病毒扫描器时, 这些扫描器会将合法的系统文件标记为恶意, 从而毒害竞争对手的杀毒产品.

从敌手目标角度来看, 投毒攻击可分为可用性攻击和完整性攻击. 可用性攻击目的是降低分类器分类精度并产生高误报率, 从而无法为正常用户提供服务, 甚至使系统拒绝服务. 完整性攻击可以使机器学习产生较高漏报率, 降低成功检测到恶意活动的几率.

3.2.1　可用性攻击

红鲱鱼攻击是一类常用的可用性攻击, 主要思路是在训练样本中引入真正恶意样本不具备的伪特征, 降低真正恶意样本被模型检测为恶意样本的概率, 从而产生高误报率. 红鲱鱼攻击最早被用于攻击垃圾邮件过滤器. 由于早期垃圾邮件过滤器模型较为简单, 对攻击者的敌手知识要求也较低, 有时只需具备简单的令牌 (单词) 字典. 文献 [8, 44, 90] 研究了针对朴素贝叶斯垃圾邮件过滤器的字典攻击. 基本思想是, 攻击者通过发送电子邮件, 让垃圾邮件检测模型关注特定的 “坏词”. 随后在垃圾邮件中不出现这些单词, 这会导致垃圾邮件过滤器将它们误分类为正常邮件. 类似的研究还有文献 [91−92].

Kim 等[93] 研究了针对蠕虫特征码自动生成的入侵检测系统 Autograph 的投毒攻击, 主要思想是使 Autograph 确信网络节点受到感染并将该类流量识别为恶意后, 不断发送该类流量, 进而导致机器学习系统拒绝服务. Rubinstein 等[94] 采取了类似的思想, 研究利用主成分分析的敏感性发动拒绝服务攻击. 具体来说, 攻击者在训练数据中注入大量虚假流量, 通过增加目标流链路上方差的形式, 误导异常检测系统并产生高昂的计算开销, 从而使其可用性显著降低.

一个有趣的现象是, 与 SVM 和随机森林等浅层学习相比, 针对深度学习的可用性投毒攻击的研究较少. 有两方面原因: 一是计算复杂性; 二是深度学习需要的数据量比浅层学习高很多, 对数据投毒的固有脆弱性较小, 为使准确性降到足够低, 可能需要更大的攻击强度. 在训练样本动辄以十万、百万为量级的深度学习中, 即便对 10% 的训练样本进行投毒攻击也拥有很高的难度.

3.2.2　完整性攻击

完整性攻击是使机器学习产生较高漏报率的投毒攻击. 其中, 油蛙攻击通过多次微小攻击达到投毒攻击的目标, 是一类常用的攻击方法. 在机器学习系统部署过程中, 原始训练数据大多是保密的, 攻击者很难对其进行修改. 而恶意软件分类、垃圾邮件检测等系统为了增强适应能力需要定期更新决策模型, 从而为攻击者发动油蛙攻击提供了可趁之机. 部分机器学习模型在更新过程中利用在线质心异常检测方法[114] 对常规攻击或异常进行检测. 原理比较简单, 如式 (1) 所示, 首先计算训练数据的均值 $c=\sum_{i=1}^{n}x_i/n$, 进而计算新样本 x 与质心 c 的距离并将其与设定的阈值进行比较, 如果距离大于阈值则模型存在异常, 反之则否:

$$f(x)=\begin{cases}1, & \parallel x-c \parallel > \theta \\ 0, & \text{否则}\end{cases} \tag{1}$$

如图 7 所示, 攻击者利用在线学习模型周期性更新的特点, 将伪装后的样本注入用于再训练决策模型的训练数据中, 使用于识别用户特征的模型训练数据中心值 $\boldsymbol{X}_c$ 在投毒样本的作用下逐步向攻击特征的中心值 $\boldsymbol{X}_a$ 移动, 从而导致用于异常检测的质心发生增量改变. 由于每个增量变化都足够小, 系统难以检测到这种变化; 然而, 随着时间的推移, 模型的质心会移动到攻击者预期的位置. Nelson 等[95] 针对无限窗口的定期再训练异常检测方法进行研究, 表明攻击者需要指数级的数据量才能成功实施投毒攻击. 然而, 无限窗口的假设也阻碍了学习算法适应数据分布中合法变化的能力. 随后, Kloft 等[96] 在有限窗口的情况下, 对在线质心异常检测方法进行安全评估. 研究显示, 攻击者仅限于特定数量的样本时, 需要控制 5% ~ 15% 的流量.

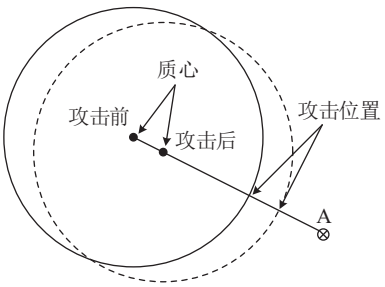


图 7　针对质心异常检测的投毒攻击

Fig. 7　The illustration of poisoning attack for centroid anomaly detection

油蛙攻击也被用于聚类算法, 如图 8 所示, Biggio 等[97] 提出桥接攻击, 并在后续研究中对其进行完善[98]. 其主要思想是以迭代的方式将攻击点引入簇之间的

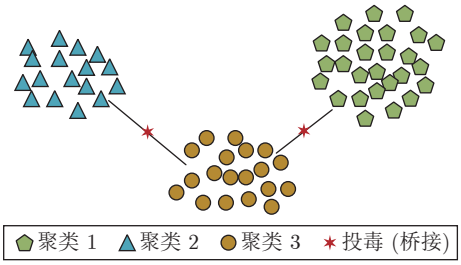


图 8 单连接分层聚类的桥接攻击

Fig. 8 Bridge-based attacks against single-linkage clustering

空间, 通过对簇进行合并操作, 从而使系统无法正常使用. 研究表明, 在基于分层聚类的恶意软件聚类开源工具 Malheur 上, 针对两个恶意软件数据集进行测试, 仅需要注入 2% 的样本就能利用桥接攻击将簇的数量从 40 个减少到 5 个. 研究认为, 对于使用簇内距离的任意聚类算法, 桥接攻击都比较有效.

后门攻击也是一类经常使用的完整性攻击. 后门模式启用时, 机器学习模型会按攻击者的意图将样本区分为指定类别. 后门模式未启用时, 机器学习模型不会改变决策边界, 但准确性有可能降低. 后门攻击的显著特点是, 这种攻击方式成功与否不明显依赖于攻击者的敌手知识. 攻击者只需要将“后门”嵌入相关样本, 同时保持样本标签不变并将嵌入 “后门”的样本加入训练集. 值得注意的是, 攻击者如果拥有模型、训练集以及学习算法等敌手知识, 那么可以利用这些知识优化后门攻击, 从而确保模型精度不受影响的同时, 以最小的攻击强度实现后门攻击. 文献 [115] 针对目前最先进的人脸识别系统提出后门攻击, 实施方法很简单, 基本思路就是将包含不易察觉的水印或者具有特定内容的图片以后门的方式插入, 包括脸上戴有眼镜、图片背景中有树木、天空中有飞翔的小鸟等特定特征.

随着机器学习训练成本越来越高, 外包成为机器学习训练常用的一种形式. 在机器学习外包服务中, 攻击者可以通过控制模型参数从而发动后门攻击, 甚至提供模型训练服务的第三方本身就是投毒攻击的发动方. 在这种高度脆弱的场景中, 攻击者可以很容易地将后门嵌入到所学习的模型中.

3.3 隐私窃取

对抗机器学习以规避攻击或投毒攻击为手段, 最终目的是破坏模型本身的安全性. 然而, 安全与隐私总是如影随形. 隐私窃取利用模型提取、模型反演[100] 以及成员推断等反向工程, 窃取机器学习模型或训练数据的信息, 强化攻击者拥有的敌手知识, 为顺利实施规避攻击或投毒攻击提供丰富的先验知识, 达到破坏机器学习安全性的目的.

模型提取的目的是提取基于私有数据训练的目标模型 f 的参数, 从而构建一个代理模型 $\widehat{f}$, 使得代理模型 $\widehat{f}$ 与目标模型 f 对输入样本产生相似的预测结果. 如图 9 所示, Tramèr 等[99] 对 BigML 和亚马逊等在线机器学习服务提供商进行了模型提取攻击, 主要是从观察到的输入输出对 $(x, h_\theta(x))$ 中恢复参数 θ. 研究表明, 针对 LR、DNN 和 DT 等常用模型, 使用相对较少的查询便可以学习与黑盒机器学习服务的决策非常相似的分类器. 一旦黑盒模型被实施反向工程, 攻击者就不再需要向在线机器学习服务提供商订阅机器学习服务. 然而, 文献 [99] 对机器学习模型使用随机查询方式, 其显著弱点是据此发起的攻击容易被检测. Papernot等[101] 提出了一种更符合实际情况的多阶段查询方式. 首先收集具有代表性的样本作为初始训练集并训练初始代理分类器, 而后进行数个阶段的数据收集和再训练. 对于每个阶段, 攻击者将新生成的样本扩充到当前训练集, 再重新训练代理分类器. 利用这种方法, 每个后续阶段都将得到更接近真实决策边界的分类器.

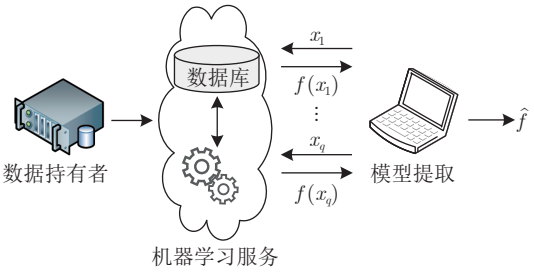


图 9 模型提取攻击

Fig. 9 Model extraction attacks

模型反演利用机器学习算法具有容易过拟合的缺陷, 即模型会隐性记忆部分过拟合的训练数据的缺陷, 通过查询、探测机器学习模型等方式, 为攻击者重建部分或全部训练数据[100, 102]. 需要注意的是, 模型提取与模型反演存在明显区别, 前者的目标是获得代理模型, 后者的目标是获取训练数据.

成员推断的目的是了解目标系统是否将特定样本用作机器学习训练数据的一部分, 其直接危害是导致提供训练数据的用户信息泄露, 包括用户地理位置数据、网页使用记录等敏感信息. 如图 10 所示, Shokri 等[103] 研究通过训练影子模型进行成员推断攻击. 训练影子模型使用与训练目标模型相同的机器学习平台. 目标模型和影子模型的训练数据集在格式上相同但不相交, 同时, 影子模型的训练数据

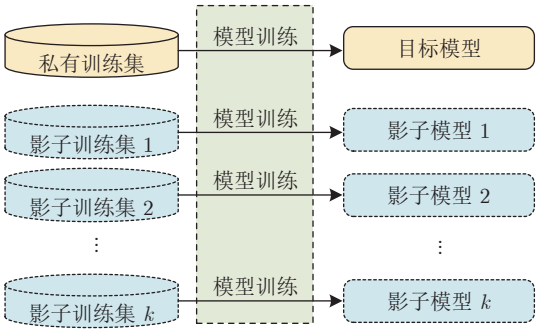


图 10　用于成员推断攻击的影子模型

Fig. 10　Shadow models for membership inference

集可能存在重叠. 所有模型的内部参数均经过独立训练. 这种方法主要利用了模型训练集包含特定样本时, 模型对这些样本的输出可能表现出置信度方面的差异. Hayes 等[116] 在文献 [103] 的基础上, 利用生成对抗网络, 将成员推断攻击扩展到生成模型的无监督学习.

4　防御措施

攻击与防御犹如军备竞赛, 此消彼长. 尽管各种对抗攻击使机器学习模型的性能显著下降, 但相应的防御技术也在不断发展之中, 随后又研究出新的攻击方法将防御措施击破. 防御措施的研究旨在为遭受对抗攻击的机器学习系统提供安全保证. 众多防御措施的提出, 推动了对抗机器学习的不断发展[7].

如表 5 所示, 与攻击方法相对应, 本文将防御措施区分为规避防御、投毒防御以及隐私保护. 需要注意的是, 规避防御和投毒防御之间也存在相互交叉的空间. 用于规避防御的措施也很可能用于投毒防御.

4.1　规避防御

4.1.1　数据降维

研究表明, 高维数据拥有较大的攻击面, 有利于对抗样本的生成. 在对模型进行测试之前, 利用特征约简、特征压缩等数据降维方法降低数据表示的复杂性, 将数据投影到低维空间以减少攻击面, 可以在一定程度上增加生成对抗样本的难度. Bhagoji 等[117] 提出了基于主成分分析降维的防御方法. Zhang 等[118] 研究利用特征选择防御规避攻击, 基于流行的前向特征选择和后向特征消除算法, 提出基于封套的对抗特征选择方法. 特征选择利用部分特征集进行训练, 可减少算法的时间和计算复杂度, 并在较小的训练集上提供更好的学习效果. 研究表明, 特征选择对基于线性 SVM 分类器的垃圾邮件检测以及高斯核 SVM 分类器的恶意 PDF 文件检测的规避攻击能取得良好效果. Wang 等[119] 提出随机消除样本特征, 使 DNN 具有不确定性, 从而阻止对抗样本的生成. 在具有 14 679 个恶意软件变体和 17 399 个良性软件的真实数据集上进行的实验表明, 随机特征消除可以显著提高模型对对抗样本的鲁棒性, 同时保持分类精度.

然而, 数据降维虽然能在一定程度上处理对抗攻击, 但降低了模型对真实样本进行分类的准确性. 同时, 数据降维之后可攻击的特征更少, 如果攻击者知道选择的特征, 可能会导致规避攻击的进一步发生.

4.1.2　鲁棒优化

鲁棒优化又称作最小最大优化, 基本思想是把对抗学习作为极小极大问题进行求解, 通过操纵训练数据以保持模型的泛化能力. 内层问题是通过尽可能干扰训练集使训练损失最大化, 而外层问题是

表 5　网络空间防御中用于对抗攻击的典型防御措施

Table 5　Typical defense against adversarial attacks for cyberspace defense

防御措施		相关文献	应用场景	简述
规避防御	数据降维	[117−118]	垃圾邮件检测	可以有效防御对抗攻击, 但模型对正常样本的精度可能降低
		[118−119]	恶意软件检测	
	鲁棒优化	[120−124]	恶意软件检测	基本思想是模型在训练时存在 “盲点”, 将构造的对抗样本注入训练集, 以提高模型的泛化能力
	防御蒸馏	[125−129]	恶意软件检测	难以防御 C&W 攻击方法
投毒防御	数据清洗	[130]	异常检测	该方法将投毒攻击视为离群值进行处理
		[131−136]	—	
	博弈论	[137−141]	垃圾邮件检测	该方法将博弈论的思想用于处理垃圾邮件的投毒攻击
隐私保护	差分隐私	[142−149]	—	该方法的难点在于如何平衡模型可用性与隐私保护效果
	模型压缩	[109]		该方法可用于缓解成员推断攻击
	模型集成	[150]		该方法的主要思想是将模型中低于特定阈值的损失梯度设为零, 可以用于防御模型提取攻击

最小化引起相应训练损失的情况.

Al-Dujaili 等[120] 将鲁棒优化的思想应用于恶意软件检测, 从而可以将机器学习目标形式化为:

$$\theta^* = \arg\min_{\theta} \mathrm{E}_{(x,\,y)\sim D}\left[\max_{x'\in \boldsymbol{S}(x)} L\left(\theta,\, x',\, y\right)\right] \tag{2}$$

式中, $\boldsymbol{S}(x)$ 是保留恶意软件 x 相关功能的二进制指示向量集, L 是原始分类模型的损失函数, y 是基准标签, θ 是可学习参数, D 表示数据样本 x 的分布. Demontis 等[121] 对支持向量机的训练目标进行修改, 保证学习到的权重向量不稀疏, 主要缺陷是在未遭受攻击的情况下, 可能降低对正常数据的分类准确性.

Goodfellow 等[48] 从鲁棒优化的角度分析对抗样本的存在, 将其归因于训练算法的 "盲点". 进而提出利用对抗训练的思想, 将构造的对抗样本注入训练集, 以提高模型的泛化能力. Yang 等[122] 针对静态恶意软件检测系统, 提出将对抗训练作为检测恶意软件对抗样本的解决方案. 在 2018 年 ICLR 会议上, Tramèr 等[123] 进一步提出集成对抗训练, 将从多个模型产生的对抗样本注入训练集, 来学习更为鲁棒的模型. Li 等[124] 将集成对抗训练用于强化恶意软件检测模型, 从而防御众多规避攻击.

对抗训练在有效提升模型鲁棒性的同时, 也存在一定局限性. 一是基于对抗训练的防御属于启发式, 缺少收敛性和鲁棒性的形式化保证, 类似于非对抗环境下的主动学习; 二是在对抗训练过程中, 需要尽可能把所有对抗样本涵盖在训练集中, 在实际应用中构建大量用于对抗训练的对抗样本代价昂贵且不现实. 虽然对抗训练存在上述局限性, 但它仍然是目前最有效的防御方法之一.

4.1.3 防御蒸馏

2015 年, Hinton 等[125] 提出知识蒸馏算法, 基本构想是将大型网络中的知识压缩到一个更容易部署的单一模型中. 在 2016 年 S&P 会议上, Papernot 等[53] 将知识蒸馏算法改进为防御蒸馏算法. 防御蒸馏为被保护的原始模型训练一个蒸馏之后的模型, 该模型具有平滑的输出面, 对干扰的敏感度也更低, 可有效提高模型鲁棒性.

然而, Carlini 等[106] 在 2017 年 S&P 会议上提出 3 种对抗攻击, 通过使用模型的部分参数还原防御蒸馏的效果, 使其无效. Hosseini 等[126] 研究了黑盒模型中的防御蒸馏方法, 并证实它不会提高分类器鲁棒性, 原因是当输入数据的一些特征被修改之后, 防御蒸馏便无效. 随后, Papernot 等[127] 提出可扩展的防御蒸馏技术以保证其有效性, 主要思想是在训练蒸馏的模型时, 训练数据的标签不仅包括原始模型输出的样本分类置信度, 还有样本的分类标签. 该研究认为防御蒸馏破坏了攻击所需的损失函数梯度, 并没有从实质上解决问题, 因此无法抵御 Carlini 等[106] 提出的攻击, 而可扩展的防御蒸馏则解决了这一问题.

在基于静态分析的恶意软件分类问题中, Grosse 等[128] 提出通过降低神经网络对对抗样本的敏感性, 提高恶意软件分类器的分类精度, 并对防御蒸馏以及前述的特征约简、对抗训练等 3 种不同防御机制进行比较. 实验结果表明: 1) 特征约简作为降低输入复杂度和简化分类器训练的常用方法, 可能无法抵御对抗攻击, 甚至由于对抗样本生成过程的简化, 可能会产生不利的影响; 2) 防御蒸馏确实降低了误分类率, 但改进幅度远小于计算机视觉中的性能; 3) 对抗训练可以提高模型的抵抗力, 但是易受再训练时模型参数的影响. 与文献 [128] 类似, Stokes 等[129] 提出将防御蒸馏应用于基于动态分析的恶意软件分类问题.

4.2 投毒防御

4.2.1 数据清洗

投毒攻击的主要原理是, 通过污染数据集对机器学习模型的性能造成影响. 因此, 从一定角度上讲, 可以采用针对离群值的方法进行处理. 其中, 数据清洗就是一种比较好的方法. 数据清洗先对训练集进行净化处理, 然后进行标准的模型学习. 运用净化处理消除数据投毒, 主要包括两方面内容: 一是检测和移除投毒样本; 二是通过特征选择或特征压缩移除可能主要由攻击者使用的次要信号成分或特征. 例如, 利用主成分分析移除低能量系数, 基本思想是对特征空间中距训练数据足够远的样本进行检测并移除. 由于这些样本通常很少出现在由训练数据填充的特征空间区域, 因此这些样本也被称为盲区规避点.

Cretu 等[130] 在网络异常检测中, 对机器学习训练阶段进行扩展, 主要是引入一个净化阶段以过滤污染的数据. 净化阶段分别利用部分训练数据生成多个 "微模型", 使用这些 "微模型"为每个输入样本生成临时标签, 然后对 "微模型"进行组合投票, 从而确定训练数据中可能代表攻击属性的特征. 实验结果表明, 扩展得到的净化阶段可以使网络异常检测系统尽可能 "无攻击", 显著提高未标记训练数据的质量. Laishram 等[131] 通过识别投毒数据样本点并对其进行过滤, 防御通过标签翻转产生的投毒攻击. Steinhardt 等[132] 对数据依赖和数据独立净化的最坏情况损失进行研究. Metzen 等[133] 提出训练

一个从分类网络的中间层获取输入, 并对投毒样本进行过滤的检测器网络. Feinman 等[134] 使用最终隐藏层特征空间中的核密度估计和贝叶斯神经网络不确定性估计检测针对 DNN 模型的投毒样本.

机器遗忘学习[135] 是类似数据清洗的另一种防御方法, 基本思想是从训练集中删除投毒数据, 无需重新训练分类. 上述方法主要问题是通用性不够, 很难扩展到更为复杂的模型. 之后, Bourtoule 等[136] 提出分片隔离切割聚合 (Sharded isolated sliced aggregated, SISA) 训练框架, 用于实现通用机器学习模型的遗忘学习. 如图 11 所示, 首先将数据分片, 分片的数据又被切割成更小的碎片. 利用新碎片扩充训练集之前, 将其以众多分片的形式呈现并保存其参数, 从而在分片上训练多个相互隔离的组元模型. 在测试阶段, 将测试数据馈送至所有组元模型, 采用类似于集成学习的方法对其进行聚合并输出测试结果. 当需要遗忘学习数据时, 只需对包含遗忘学习数据点的组元模型, 利用保存的参数进行重新训练. 其主要限制是防御者必须知道哪些是不需要学习的内容, 在实际操作中有一定难度.

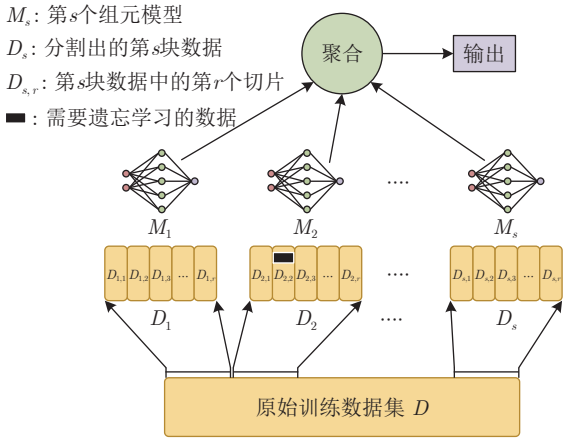


图 11　SISA 训练示意图[136]

Fig. 11　The illustration of SISA training[136]

4.2.2　博弈论

利用博弈论进行防御, 具有独特优势. 广义上讲, 博弈论的目标是要找到表征最优策略的均衡. Brückner 等[137] 针对垃圾邮件检测问题, 提出在静态博弈中, 寻找凸损失函数唯一纳什均衡的方法. 随后, Brückner 等[138] 将斯塔克尔伯格预测博弈应用于垃圾邮件检测领域. 在此基础上, Brückner 等[139] 将纳什均衡与 LR 和 SVM 模型相结合, 构造出相应的纳什变体模型, 在垃圾邮件检测问题中表现出优于非纳什变体模型的性能. 其他类似方法包括 Sengupta 等[140] 利用贝叶斯斯塔克尔伯格博弈构建移动目标防御, 主要思想是将用于软件安全范式的移动目标防御引入对抗机器学习. 移动目标防御的关键是实现多个学习器之间的最优切换, 为此引入贝叶斯斯塔克尔伯格博弈, 在保证合法用户高精度的前提下, 降低攻击成功率.

利用博弈论进行防御似乎比较有效. 然而在对抗环境下, 预设的攻击策略能否反应现实情况, 以及能反应哪种程度的现实情况, 实际上仍属未解决的开放问题. 主要原因是对抗机器学习不具备完善的规则, 不能用适当的回报函数来设定攻击者的真实目标, 导致攻击者在实际行动中的行为可能偏离预期. 针对这一问题, Biggio 等[141] 提出利用概率模型对攻击行为进行预测, 之后又在文献 [10] 提出使用学习算法的反馈对后续攻击行为进行预设.

4.3　隐私保护

针对机器学习的隐私窃取, 常用的防御技术主要包括差分隐私[142−143]、正则化、模型压缩、模型集成以及添加噪声等. 本文对差分隐私进行详细介绍, 并对其他几种防御方法进行概述.

1) 差分隐私. 差分隐私是 Dwork 等[143] 提出的一种具有严格数学理论支撑的隐私定义, 旨在保证攻击者无法根据输出差异推测个体的敏感信息. 差分隐私主要是在模型训练过程中引入随机性, 即添加一定的随机噪声, 使输出结果与真实结果具有一定程度的偏差, 以防止攻击者恶意推理. 满足差分隐私的算法, 其输出结果的概率分布不会因数据集中一条记录的增加、删除或修改, 而产生明显差异, 在一定程度上避免了攻击者通过捕捉输出差异进而推测个体记录的敏感属性值. 形式上, 差分隐私的定义如下:

定义 1. ε-差分隐私. 对任意的邻接数据集 D 和 $D' \in \boldsymbol{\mathcal{D}}$. 给定随机算法 f: $\boldsymbol{\mathcal{D}} \mapsto \mathbf{R}$ 和任意输出结果 $\boldsymbol{S} \subseteq \mathbf{R}$, 若不等式 (3) 成立, 则称算法 f 满足 ε-差分隐私. 即:

$$\max_{\boldsymbol{S}} \left[\ln \frac{\Pr\left[f\left(D\right) \in \boldsymbol{S}\right]}{\Pr\left[f\left(D'\right) \in \boldsymbol{S}\right]} \right] \leq \varepsilon \tag{3}$$

式中, 邻接数据集指有且仅有 1 条记录不同的 2 个数据集. 不等式左边可视为算法访问数据集后造成的隐私损失, ε 用于控制算法的隐私保护程度, 称为隐私预算. 差分隐私机制将算法的隐私损失控制在有限范围内, ε 越小则算法隐私保护效果越好. 常用的有拉普拉斯机制[143]、指数机制[144] 和高斯机制[145]. 这些机制中, 噪声大小取决于算法的敏感度.

差分隐私机制是目前机器学习隐私保护研究中最常采用的方法之一[146]. 由于模型训练过程需要多

次访问敏感数据集, 如数据预处理、计算损失函数、梯度下降求解最优参数等, 必须将整个训练过程的全局隐私损失控制在尽可能小的范围内. 对于结构简单的模型, 此要求较容易实现. 然而, 对结构复杂、参数量大的深度学习模型而言, 模型可用性与隐私保护效果比较难以平衡, 这是该技术面临的最大问题与挑战. 文献 [147−148] 表明, 针对成员推断攻击, 只有牺牲模型性能, 差分隐私才能在一定程度上保护隐私. Jayaraman 等[147] 对 LR 和 DNN 模型的成员推断攻击评估了多种用于差分隐私的松弛, 结果表明, 这些松弛对性能与隐私之间的平衡产生了影响. 直观地说, 尽管松弛可以减少所需的附加噪声, 但也增加了隐私泄漏. McMahan 等[149] 将差分隐私用于针对联邦学习等多方参与场景的隐私防御, 该方法对参与者数量有具体要求, 当参与者数量少于 30 时, 实验效果不明显.

2) 正则化[103, 116, 150−153]. 由于成员推断攻击与过拟合有一定关系, 利用正则化可以缓解过拟合. 因此正则化也被用于隐私防御.

3) 模型压缩. 文献 [109] 提出, 将深度学习模型中低于特定阈值的损失梯度设置为零, 可以用于防御模型提取攻击. 相应的实验结果表明, 只有 20% 的梯度设置为零, 对模型性能的影响可以忽略不计.

4) 模型集成. 文献 [150] 提出利用模型集成的方法, 可以在一定程度上缓解成员推断攻击的影响.

5) 添加噪声. 文献 [154] 提出通过在数据中添加噪声的方式, 随机改变部分数据的标签, 可以在一定程度上防御成员推断攻击.

5 研究展望与挑战

随着网络空间的不断发展和演化, 基于机器学习的网络空间防御系统也面临新的攻击机制与更为严苛多样的需求, 对机器学习模型安全提出了新的挑战. 亟需解决的各种问题随之涌现而来, 同时也催生出更多新的研究方向.

5.1 图数据对抗攻防成为新的研究分支

网络空间广泛存在的图数据为网络空间防御算法的研究提供了全新的研究思路, 图卷积网络[155]、变分图自编码器[156] 和 GraphSAGE[157] 等图神经网络模型的提出促进了图结构数据上机器学习的研究进展, 为网络空间防御提供了更为丰富的手段. 比如, 针对 Android 恶意软件检测提出的 HinDroid[158]、AiDroid[159] 和 Scorpion[160] 等系统, 考虑了应用程序和 API、移动设备识别码、签名等类型实体之间的高级语义关系, 并引入结构化的异构图来建模这种复杂的关系, 已成功地应用于反恶意软件行业, 初步表明图神经网络模型应用于网络空间防御的可行性.

伴随图神经网络模型研究的发展, 关注图神经网络模型安全性和脆弱性的图数据对抗攻防问题也逐渐成为研究热点. Zügner 等[161] 首次研究了图数据的对抗攻击, 提出了图对抗攻击的一般框架, 如图 12 所示, 在节点特征和图结构几乎没有被干扰的情况下, 对指定节点进行攻击并使节点被错分. Zhu 等[162] 提出利用对抗防御框架改进图卷积网络模型, 以提高其鲁棒性. 区别于文献 [161] 中同构图的对抗攻防问题, 文献 [163] 在恶意软件检测问题中首次研究了异构图对抗攻防问题. 此外, Sun 等[164] 就近年来出现的 100 多篇图数据对抗攻防论文进行综述, 在介绍图数据对抗攻防研究现状、不足和未来方向的同时, 还对图数据对抗攻防的评估指标和常用数据集进行了综述, 指出图数据对抗攻防是未来该领域的主要方向.

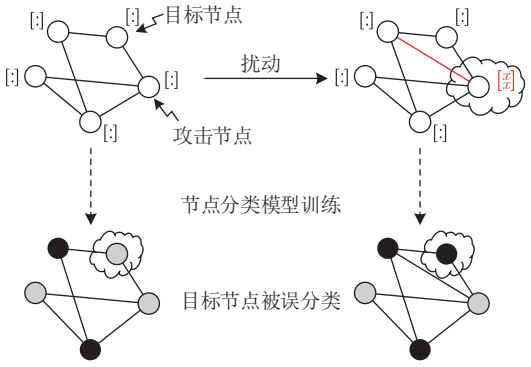


图 12 图神经网络对抗攻击[161]

Fig. 12 Adversarial attacks on graph neural networks[161]

与传统对抗机器学习相比, 图数据的对抗攻防研究面临特有的挑战: 1) 图数据的攻防对象包括图的节点、连边、子图甚至是整个图结构, 敌手模型更为复杂; 2) 图数据各组成部分之间互为关联、互相影响, 如何对图数据进行攻击并确保相关扰动 “不可察觉”还没有统一的定义; 3) 图的结构和特征是离散数据, 在对抗样本生成上存在较大挑战.

需要注意的是, 上述部分挑战也是网络空间防御所具有的特性. 在利用传统非图数据进行网络空间防御研究的同时, 利用图数据也许会为相关研究的发展提供更为友好的支持. 针对图数据的对抗攻防问题, 未来的工作可以从以下方面开展: 1) 深入探究面向图数据对抗学习的基础理论, 构建适合图结构数据的对抗度量模型和形式化描述方法; 2) 改进已有的攻击与防御策略, 本文提及的相关策略都是当前对抗机器学习较为成熟的策略, 因此可以迁

移至图数据的对抗攻防研究中, 并针对图数据的特殊性进行改进和优化.

5.2 实施更为合理的攻击

机器学习的敌手模型主要包括敌手目标、知识和能力 3 个方面, 在网络空间具体的攻击过程中, 可能需要对这 3 个方面的合理性进行考虑.

1) 对敌手知识进行合理假设. 在众多研究中, 对抗攻击往往看起来非常有效, 但提出的防御措施通常在后续工作中被绕过[165]. 然而, 要了解这些攻击在实际中产生的安全威胁, 必须对现实中攻击者的能力和局限性有客观的认识. 在高估攻击者能力的假设下, 对投毒和规避攻击进行评估, 不能准确描述对机器学习模型构成的安全威胁. 文献 [92, 105, 166] 假设攻击者知道分类器的结构和学习参数, 文献 [167] 进一步假设攻击者具有包括特征向量以及样本类别等在内的真实数据分布. 然而, 关于这些分布的知识通常是机器学习中的 "圣杯", 给定这些分布就可以获得最优分类器. 实际上, 网络空间防御中, 关键的机器学习系统使用专有模型, 给定有限训练集最多只能对联合分布进行不完全估计, 即使是内部攻击者也不可能完全知道真实的数据分布.

2) 对攻击能力进行合理假设. 如果模型决策时存在较高不确定性, 导致无法对样本正确分类, 可以考虑引入人在回路模式的专家建议, 对样本进行分类或拒绝接收样本. 同时, 如果专家和机器在样本分类决策上存在分歧, 那么引入这些攻击样本就成为主动学习的一种模式, 进而纠正模型对样本的决策.

3) 在真实世界生成对抗样本. 在对抗机器学习中, 基于梯度下降的规避攻击属于反向特征映射问题[75, 92]. 以 PDF 恶意文件检测为例, 迄今为止, 大多数基于优化的攻击只在特征空间中通过直接操纵特征向量 $\boldsymbol{x}$ 创建对抗样本, 而并未创建可以实现规避功能的恶意 PDF 样本. 在实际中, 很少考虑如何反转 $\boldsymbol{x}$ 的表示, 以获得相应的 PDF 文件 $\boldsymbol{z}=\phi^{-1}(\boldsymbol{x})$. 基于优化的规避攻击可能会改变 PDF 恶意文件语义, 甚至破坏嵌入的攻击代码的恶意功能. 可以说, 创建符合所需特征向量的恶意软件样本, 使其避开目标系统的检测, 在实际操作中比较困难. 这一点具有重要意义, 在研究中值得关注.

5.3 构筑科学可用的防御体系

1) 构筑统一的防御体系. 解决网络空间防御问题, 应注意从网络空间全局角度出发. 网络空间较为明显的特征之一是各节点之间互为影响、互为依赖. 在具体的实践过程中, 可以将指挥控制的理念融入其中, 对资源进行组合调度, 充分利用各种先进技术, 形成整体合力.

2) 设置科学合理的指标体系. Carlini 等[165] 指出, 防御指标除了准确度, 还应该有误报率和真阳性率等指标. 如何设置科学合理的指标体系是一个开放的问题, 也是迫切需要研究的内容. 同时还需注意的是, 攻击与防御应在一定程度上具有对称性. 比如, 文献 [92, 167] 对基于朴素贝叶斯分类器的垃圾邮件问题进行攻击, 进而表明防御措施无效. 这种攻击取得成功的原因, 很可能是由于朴素贝叶斯分类器构建单个模型, 表征具有时变多样性的垃圾邮件类别, 属于弱分类器.

6 结束语

当前机器学习在网络空间防御中的应用仍然处于起步阶段, 距离普及和推广仍有很长的路要走. 然而, 随着科技的发展, 网络空间中的安全问题越来越严峻, 利用机器学习算法实现网络空间防御将是大势所趋.

推进机器学习在网络空间防御中的应用将面临以下两个问题: 1) 机器学习算法本身可能会成为一个新的弱点, 新的基于机器学习的网络威胁将不断出现, 研究人员认为切实有效的防御方法, 在随后很快就被认为无效; 2) 对现有机器学习算法进行可靠的安全评估比较困难.

本文全面综述了网络空间防御中机器学习可能面临的攻击, 以及可以采取的防御措施, 并对下步研究方向提出了一些展望. 对抗机器学习是加速机器学习技术在网络空间防御应用落地的催化剂. 利用机器学习技术解决网络空间防御问题时, 应围绕对抗机器学习的统一框架, 研究机器学习可能遭受的攻击, 评估相应的防御能力, 利用各种方法提高模型应用于网络空间防御时的安全性.

References

1 搜狐. 美国东海岸断网事件主角Dyn关于DDoS攻击的后果. [Online], available: https://www.sohu.com/a/117078005_257305, October 25, 2016

2 搜狐. WannaCry勒索病毒事件分析. [Online], available: https://www.sohu.com/a/140863167_244641, May 15, 2017

3 Peng Zhi-Yi, Zhang Zhun, Hui Zhi-Bin, Qin Qing-Ling. Annual Report on the Development of Cyberspace Security in China (2019). Beijing: Social Sciences Academic Press, 2019.
(彭志艺, 张衡, 惠志斌, 覃庆玲. 中国网络空间安全发展报告(2019版). 北京: 社会科学文献出版社, 2019.)

4 Zhang Lei, Cui Yong, Liu Jing, Jiang Yong, Wu Jian-Ping. Application of machine learning in cyberspace security research. *Chinese Journal of Computers*, 2018, **41**(9): 1943−1975

(张蕾, 崔勇, 刘静, 江勇, 吴建平. 机器学习在网络空间安全研究中的应用. 计算机学报, 2018, **41**(9): 1943−1975)

5 Joseph A D, Nelson B, Rubinstein B I P, Tygar J D. *Adversarial Machine Learning*. Cambridge: Cambridge University Press, 2019.

6 Yevgeniy V, Murat K. *Adversarial Machine Learning*. San Rafael: Morgan & Claypool Publishers, 2018.

7 Biggio B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 2018, **84**: 317−331

8 Barreno M, Nelson B, Sears R, Joseph A D, Tygar J D. Can machine learning be secure? In: Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security. Taipei, China: ACM, 2006. 16–25

9 Grosse K, Papernot N, Manoharan P, Backes M, Mcdaniel P. Adversarial examples for malware detection. In: Proceedings of the 22nd European Symposium on Research in Computer Security. Oslo, Norway: Springer, 2017. 62−79

10 Biggio B, Fumera G, Roli F. Pattern recognition systems under attack: Design issues and research challenges. *International Journal of Pattern Recognition and Artificial Intelligence*, 2014, **28**(7): 1460002

11 Li Xin-Jiao, Wu Guo-Wei, Yao Lin, Zhang Wei-Zhe, Zhang Bin. Progress and future challenges of security attacks and defense mechanisms in machine learning. *Journal of Software*, 2021, **32**(2): 406−423
(李欣姣, 吴国伟, 姚琳, 张伟哲, 张宾. 机器学习安全攻击与防御机制研究进展和未来挑战. 软件学报, 2021, **32**(2): 406−423)

12 Liu Q, Li P, Zhao W, Cai W, Yu S, Leung V C M. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access*, 2018, **6**: 12103−12117

13 Wang X, Li J, Kuang X, Tan Y A. The security of machine learning in an adversarial setting: A survey. *Journal of Parallel Distributed Computing*, 2019, **130**: 12−23

14 Pitropakis N, Panaousis E, Giannetsos T, Anastasiadis E, Loukas G. A taxonomy and survey of attacks against machine learning. *Computer Science Review*, 2019, **34**: 100199

15 Papernot N, Mcdaniel P, Sinha A, Wellman M P. Sok: Security and privacy in machine learning. In: Proceedings of the 3rd IEEE European Symposium on Security and Privacy. London, UK: IEEE, 2018. 399−414

16 Ji Shou-Ling, Du Tian-Yu, Li Jin-Feng, Shen Chao, Li Bo. Security and privacy of machine learning models: A survey. *Journal of Software*, 2021, **32**(1): 41−67
(纪守领, 杜天宇, 李进锋, 沈超, 李博. 机器学习模型安全与隐私研究综述. 软件学报, 2021, **32**(1): 41−67)

17 Liu Jun-Xu, Meng Xiao-Feng. Survey on privacy-preserving machine learning. *Journal of Computer Research and Development*, 2020, **57**(2): 346−362
(刘俊旭, 孟小峰. 机器学习的隐私保护研究综述. 计算机研究与发展, 2020, **57**(2): 346−362)

18 Isakov M, Gadepally V, Gettings K, Kinsy M. Survey of attacks and defenses on edge-deployed neural networks. In: Proceedings of the 2019 IEEE High Performance Extreme Computing Conference. Waltham, MA, USA: IEEE, 2019. 1−8

19 Wiyatno R, Xu A, Dia O, Berker A D. Adversarial examples in modern machine learning: A review. ArXiv: 1911.05268, 2019.

20 Huang X, Kroening D, Ruan W, Sun Y, Thamo E, Wu M, et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 2020, **37**: 100270

21 Pan Wen-Wen, Wang Xin-Yu, Song Ming-Li, Chen Chun. Survey on generating adversarial examples. *Journal of Software*, 2020, **31**(1): 67−81
(潘文雯, 王新宇, 宋明黎, 陈纯. 对抗样本生成技术综述. 软件学报, 2020, **31**(1): 67−81)

22 Rigaki M, García S. A survey of privacy attacks in machine learning. ArXiv: 2007.07646, 2020.

23 Tan Zuo-Wen, Zhang Lian-Fu. Survey on privacy preserving techniques for machine learning. *Journal of Software*, 2020, **31**(7): 2127−2156
(谭作文, 张连福. 机器学习隐私保护研究综述. 软件学报, 2020, **31**(7): 2127−2156)

24 Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 2018, **6**: 14410−14430

25 Machado G R, Silva E, Goldschmidt R R. Adversarial machine learning in image classification: A survey towards the defender′s perspective. ArXiv: 2009.03728, 2020.

26 Serban A, Poll E, Visser J. Adversarial examples on object recognition: A comprehensive survey. *ACM Computing Surveys*, 2020, **53**(3): 66

27 Ding J, Xu Z. Adversarial attacks on deep learning models of computer vision: A survey. In: Proceedings of the 20th International Conference on Algorithms and Architectures for Parallel Processing. New York, NY, USA: Springer, 2020. 396−408

28 Zhang W, Sheng Q Z, Alhazmi A, Li C. Adversarial attacks on deep-learning models in natural language processing. *ACM Transactions on Intelligent Systems and Technology*, 2020, **11**(3): 1−41

29 Biggio B, Fumera G, Russu P, Didaci L, Roli F. Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective. *IEEE Signal Processing Magazine*, 2015, **32**(5): 31−41

30 Papangelou K, Sechidis K, Weatherall J, Brown G. Toward an understanding of adversarial examples in clinical trials. In: Proceedings of the 2018 European Conference on Machine Learning and Knowledge Discovery in Databases. Dublin, Ireland: Springer, 2018. 35−51

31 Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 2021, **14**: 156−180

32 Corona I, Giacinto G, Roli F. Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues. *Information Sciences*, 2013, **239**: 201−225

33 Maiorca D, Biggio B, Giacinto G. Towards adversarial malware detection: Lessons learned from PDF-based attacks. *ACM Computing Surveys*, 2019, **52**(4): 78

34 Army U S G U. *Joint Publication 3−12: Cyberspace Operations*. North Charleston: Create Space Independent Publishing Platform, 2018.

35 Gibson W. *Neuromancer*. New York: Ace Books, 1984.

36 Luo Jun-Zhou, Yang Ming, Ling Zhen, Wu Wen-Jia, Gu Xiao-Dan. Architecture and key technologies of cyberspace security. *Scientia Sinica Informationis*, 2016, **46**(8): 939−968
(罗军舟, 杨明, 凌振, 吴文甲, 顾晓丹. 网络空间安全体系与关键

技术. 中国科学(信息科学), 2016, **46**(8): 939−968)

37 Fang Bin-Xing. A hierarchy model on the research fields of cyberspace security technology. *Chinese Journal of Network and Information Security*, 2015, **1**(1): 2−7
(方滨兴. 从层次角度看网络空间安全技术的覆盖领域. 网络与信息安全学报, 2015, **1**(1): 2−7)

38 National Institute of Standards and Technology. Framework for improving critical infrastructure cybersecurity version 1.1. [Online], available: https://www.nist.gov/publications/framework-improving-critical-infrastructure-cybersecurity-version-11, April 16, 2018.

39 Turing A M. Computing machinery and intelligence. *Mind*, 1950, **59**(236): 433−460

40 Samuel A L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 1959, **3**(3): 211−229

41 Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of Machine Learning*. London: MIT Press, 2012.

42 Dalvi N, Domingos P, Sumit M, Verma S D. Adversarial classification. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA: ACM, 2004. 99−108

43 Lowd D, Meek C. Adversarial learning. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA: ACM, 2005. 641−647

44 Lowd D, Meek C. Good word attacks on statistical spam filters. In: The 2nd Conference on Email and Anti-Spam. Stanford, CA, USA: 2005.

45 Barreno M, Nelson B, Joseph A D, Tygar J D. The security of machine learning. *Machine Learning*, 2010, **81**(2): 121−148

46 Dasgupta P, Collins J B. A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks. *AI Magazine*, 2019, **40**(2): 31−43

47 Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: 2014.

48 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: 2015.

49 Li X, Li F. Adversarial examples detection in deep networks with convolutional filter statistics. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 5775−5783

50 Lu J, Issaranon T, Forsyth D. Safetynet: Detecting and rejecting adversarial examples robustly. In: Proceedings of the 16th IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 446−454

51 Meng D, Chen H. MagNet: A two-pronged defense against adversarial examples. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA: ACM, 2017. 135−147

52 Melis M, Demontis A, Biggio B, Brown G, Fumera G, Roli F. Is deep learning safe for robot vision? Adversarial examples against the icub humanoid. In: Proceedings of the 16th IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE, 2017. 751−759

53 Papernot N, Mcdaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proceedings of the 2016 IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2016. 582−597

54 Laskov P, Lippmann R. Machine learning in adversarial environments. *Machine Learning*, 2010, **81**(2): 115−119

55 Joseph A, Laskov P, Roli F, Tygar J, Nelson B. Machine learning methods for computer security. *Dagstuhl Reports*, 2012, **2**: 109−130

56 Finlayson S G, Bowers J D, Ito J, Zittrain J L, Beam A L, Kohane I S. Adversarial attacks on medical machine learning. *Science*, 2019, **363**(6433): 1287−1289

57 Heaven D. Why deep-learning ais are so easy to fool. *Nature*, 2019, **574**: 163−166

58 Cheng Qi-Qin, Wan Liang. Application research of BiLSTM in cross-site scripting detection. *Journal of Frontiers of Computer Science and Technology*, 2020, **14**(8): 1338−1347
(程琪芩, 万良. BiLSTM在跨站脚本检测中的应用研究. 计算机科学与探索, 2020, **14**(8): 1338−1347)

59 Biggio B, Fumera G, Roli F. Security evaluation of pattern classifiers under attack. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**: 984−996

60 Kerckhoffs A. La cryptographie militaire. *Journal des Sciences Militaires*, 1883, **9**: 5−83

61 Fan Cang-Ning, Liu Peng, Xiao Ting, Zhao Wei, Tang Xiang-Long. A review of deep domain adaptation: General situation and complex situation. *Acta Automatica Sinica*, 2021, **47**(3): 515−548
(范苍宁, 刘鹏, 肖婷, 赵巍, 唐降龙. 深度域适应综述: 一般情况与复杂情况. 自动化学报, 2021, **47**(3): 515−548)

62 Smutz C, Stavrou A. Malicious PDF detection using metadata and structural features. In: Proceedings of the 28th Annual Computer Security Applications Conference. Orlando, Florida, USA: ACM, 2012. 239–248

63 Clements J, Yang Y, Sharma A A, Hu H, Lao Y. Rallying adversarial techniques against deep learning for network security. ArXiv: 1903.11688, 2019.

64 Wittel G L, Wu S F. On attacking statistical spam filters. In: Proceedings of the 1st Conference on Email and Anti-spam. Mountain View, CA, USA: 2004. 1−7

65 Liu C, Stamm S. Fighting unicode-obfuscated spam. In: Proceedings of the Anti-phishing Working Groups 2nd Annual eCrime Researchers Summit. Pittsburgh, PA, USA: ACM, 2007. 45−59

66 Sculley D, Wachman G M, Brodley C E. Spam filtering using inexact string matching in explicit feature space with on-line linear classifiers. In: Proceedings of the 15th Text REtrieval Conference. Gaithersburg, USA: 2006. 1−10

67 Wright C V, Coull S E, Monrose F. Traffic morphing: An efficient defense against statistical traffic analysis. In: Proceedings of the 16th Annual Network and Distributed System Security Symposium. San Diego, USA: ISOC, 2009. 237–250

68 Rosenberg I, Shabtai A, Rokach L, Elovici Y. Generic black-box end-to-end attack against state of the art API call based malware classifiers. In: Proceedings of the 21st International Symposium on Research in Attacks, Intrusions and Defenses. Heraklion, Greece: 2018. 490−510

69 Šrndić N, Laskov P. Detection of malicious PDF files based on hierarchical document structure. In: Proceedings of the 20th Annual Network and Distributed System Security Symposium. San Diego, USA: ISOC, 2013. 1−16

70 Šrndić N, Laskov P. Practical evasion of a learning-based classifier: A case study. In: Proceedings of the 35th IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2014. 197−211

71 Suciu O, Coull S E, Johns J. Exploring adversarial examples in malware detection. In: Proceedings of the 2019 IEEE Security and Privacy Workshops. San Francisco, USA: IEEE, 2019. 8−14

72 Corona I, Maiorca D, Ariu D, Giacinto G. Lux0R: Detection of malicious PDF-embedded javascript code through discriminant analysis of API references. In: Proceedings of the 2014 ACM Artificial Intelligent and Security Workshop. Scottsdale, USA: ACM, 2014. 47−57

73 Maiorca D, Corona I, Giacinto G. Looking at the bag is not enough to find the bomb: An evasion of structural methods for malicious PDF files detection. In: Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security. Hangzhou, China: ACM, 2013. 119–130

74 Xu W, Qi Y, Evans D. Automatically evading classifiers: A case study on PDF malware classifiers. In: Proceedings of the 23rd Annual Network and Distributed System Security Symposium. San Diego, USA: ISOC, 2016. 1−15

75 Biggio B, Corona I, Maiorca D, Nelson B, Srndic N, Laskov P, et al. Evasion attacks against machine learning at test time. In: Proceedings of the 2013 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Prague, Czech: Springer, 2013. 387−402

76 Smutz C, Stavrou A. When a tree falls: Using diversity in ensemble classifiers to identify evasion in malware detectors. In: Proceedings of the 23rd Annual Network and Distributed System Security Symposium. San Diego, USA: ISOC, 2016. 1−15

77 Biggio B, Corona I, Nelson B, Rubinstein B I P, Maiorca D, Fumera G, et al. Security evaluation of support vector machines in adversarial environments. In: Proceedings of the Support Vector Machines Applications. Cham, Switzerland: Springer International Publishing, 2014. 105−153

78 Kolosnjaji B, Demontis A, Biggio B, Maiorca D, Giacinto G, Eckert C, et al. Adversarial malware binaries: Evading deep learning for malware detection in executables. In: Proceedings of the 26th European Signal Processing Conference. Rome, Italy: EUSIPCO, 2018. 533−537

79 Kreuk F, Barak A, Aviv-Reuven S, Baruch M, Pinkas B, Keshet J. Adversarial examples on discrete sequences for beating whole-binary malware detection. ArXiv: 1802.04528, 2018.

80 Huang C H, Lee T H, Chang L H, Lin J R, Horng G. Adversarial attacks on SDN-based deep learning IDS system. In: Proceedings of the 2018 International Conference on Mobile and Wireless Technology. Hong Kong, China: Springer, 2019. 181−191

81 Dang H, Huang Y, Chang E-C. Evading classifiers by morphing in the dark. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA: ACM, 2017. 119−133

82 Lin Z, Shi Y, Xue Z. IDSGAN: Generative adversarial networks for attack generation against intrusion detection. ArXiv: 1809.02077, 2018.

83 Rigaki M, Garcia S. Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection. In: Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops. San Francisco, USA: IEEE, 2018. 70−75

84 Yan Q, Wang M, Huang W, Luo X, Yu F R. Automatically synthesizing DoS attack traces using generative adversarial networks. *International Journal of Machine Learning and Cybernetics*, 2019, **10**(12): 3387−3396

85 Fang Y, Huang C, Xu Y, Li Y. RLXSS: Optimizing XSS detection model to defend against adversarial attacks based on reinforcement learning. *Future Internet*, 2019, **11**: 177

86 Anderson H S, Woodbridge J, Filar B. DeepDGA: Adversarially-tuned domain generation and detection. In: Proceedings of the 9th ACM Workshop Artificial Intelligence and Security. Vienna, Austria: ACM, 2016. 13−21

87 Hu W, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN. ArXiv: 1702.05983, 2017.

88 Anderson H S, Kharkar A, Filar B, Evans D, Roth P. Learning to evade static PE machine learning malware models via reinforcement learning. ArXiv: 1801.08917, 2018.

89 Tang Chuan, Zhang Yi, Yang Yue-Xiang, Shi Jiang-Yong. DroidGAN: Android adversarial sample generation framework based on DCGAN. *Journal on Communications*, 2018, **39**(S1): 64−69
(唐川, 张义, 杨岳湘, 施江勇. DroidGAN: 基于DCGAN的Android对抗样本生成框架. 通信学报, 2018, **39**(S1): 64−69)

90 Nelson B, Barreno M, Chi F J, Joseph A D, Rubinstein B I P, Saini U, et al. Exploiting machine learning to subvert your spam filter. In: Proceedings of the 1st USENIX Workshop on Large-Scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More. San Francisco, CA, USA: USENIX Association, 2008. 1−9

91 Newsome J, Karp B, Song D X. Paragraph: Thwarting signature learning by training maliciously. In: Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection. Hamburg, Germany: Springer, 2006. 81−105

92 Huang L, Joseph A D, Nelson B, Rubinstein B I P, Tygar J D. Adversarial machine learning. In: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence. New York, USA: ACM, 2011. 43–58

93 Kim H A, Karp B, Usenix. Autograph: Toward automated, distributed worm signature detection. In: Proceedings of the 13rd USENIX Security Symposium. San Diego, USA: USENIX Association, 2004. 271−286

94 Rubinstein B I P, Nelson B, Huang L, Joseph A D, Lau S H, Rao S, et al. Antidote: Understanding and defending against poisoning of anomaly detectors. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement. Chicago, IL, USA: ACM, 2009. 1−14

95 Nelson B, Joseph A D. Bounding an attack′s complexity for a simple learning model. In: Proceedings of the 1st USENIX Workshop on Tackling Computer Systems Problems with Machine Learning Techniques. Saint Malo, France: USENIX, 2006. 1−5

96 Kloft M, Laskov P. Online anomaly detection under adversarial impact. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy:

Microtome, 2010. 405−412

97 Biggio B, Pillai I, Rota Bulo S, Ariu D, Pelillo M, Roli F. Is data clustering in adversarial settings secure? In: Proceedings of the 6th Annual ACM Workshop on Artificial Intelligence and Security. Berlin, Germany: ACM, 2013. 87−97

98 Biggio B, Rieck K, Ariu D, Wressnegger C, Corona I, Giacinto G, et al. Poisoning behavioral malware clustering. In: Proceedings of the 7th ACM Workshop Artificial Intelligence and Security. Scottsdale, USA: ACM, 2014. 27−36

99 Tramèr F, Zhang F, Juels A, Reiter M K, Ristenpart T. Stealing machine learning models via prediction APIs. In: Proceedings of the 25th USENIX Security Symposium. Austin, USA: USENIX Association, 2016. 601−618

100 Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. Denver, USA: ACM, 2015. 1322−1333

101 Papernot N, Mcdaniel P D, Goodfellow I J, Jha S, Celik Z B, Swami A. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security. Abu Dhabi, UAE: ACM, 2017. 506−519

102 Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX Security Symposium. San Diego, USA: USENIX Association, 2014. 17−32

103 Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2017. 3−18

104 Maiorca D, Giacinto G, Corona I. A pattern recognition system for malicious PDF files detection. In: Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition. Berlin, Germany: Springer, 2012. 510−524

105 Papernot N, Mcdaniel P D, Jha S, Fredrikson M, Celik Z B, Swami A. The limitations of deep learning in adversarial settings. In: Proceedings of the 2016 IEEE European Symposium on Security and Privacy. Saarbruecken, Germany: IEEE, 2016. 372−387

106 Carlini N, Wagner D A. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2017. 39−57

107 Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, et al. Robust physical-world attacks on deep learning visual classification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1625−1634

108 Chen P Y, Sharma Y, Zhang H, Yi J F, Hsieh C J. EAD: Elastic-net attacks to deep neural networks via adversarial examples. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 10−17

109 Papernot N, Mcdaniel P D, Goodfellow I J. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. ArXiv: 1605.07277, 2016.

110 Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 28th Annual Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 2672−2680

111 Wang Kun-Feng, Gou Chao, Duan Yan-Jie, Lin Yi-Lun, Zheng Xin-Hu, Wang Fei-Yue. Generative adversarial networks: The state of the art and beyond. *Acta Automatica Sinica*, 2017, **43**(3): 321−332
(王坤峰, 苟超, 段艳杰, 林懿伦, 郑心湖, 王飞跃. 生成式对抗网络GAN的研究进展与展望. 自动化学报, 2017, **43**(3): 321−332)

112 Kearns M, Li M. Learning in the presence of malicious errors. In: Proceedings of the 20th annual ACM Symposium on Theory of Computing. Chicago, USA: ACM, 1988. 267–280

113 John Leyden. Kaspersky Lab denies tricking AV rivals into nuking harmless files. [Online], available: https://www.theregister.co.uk/2015/08/14/kasperskygate/, August 14, 2015.

114 Kloft M, Laskov P. Security analysis of online centroid anomaly detection. *Journal of Machine Learning Research*, 2012, **13**: 3681−3724

115 Liao C, Zhong H, Squicciarini A C, Zhu S, Miller D J. Backdoor embedding in convolutional neural network models via invisible perturbation. In: Proceedings of the 10th ACM Conference on Data and Application Security and Privacy. New Orleans, LA, USA: ACM, 2020. 97–108

116 Hayes J, Melis L, Danezis G, Cristofaro E D. LOGAN: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019, **2019**(1): 133−152

117 Bhagoji A N, Cullina D, Sitawarin C, Mittal P. Enhancing robustness of machine learning systems via data transformations. In: Proceedings of the 52nd Annual Conference on Information Sciences and Systems. Princeton, USA: IEEE, 2018. 1−5

118 Zhang F, Chan P P K, Biggio B, Yeung D S, Roli F. Adversarial feature selection against evasion attacks. *IEEE Transactions on Cybernetics*, 2016, **46**(3): 766−77

119 Wang Q, Guo W, Zhang K, Ororbia A G, Xing X, Liu X, et al. Adversary resistant deep neural networks with an application to malware detection. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada: ACM, 2017. 1145–1153

120 Al-Dujaili A, Huang A, Hemberg E, O′ reilly U. Adversarial deep learning for robust detection of binary encoded malware. In: Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops. San Francisco, USA: IEEE, 2018. 76−82

121 Demontis A, Melis M, Biggio B, Maiorca D, Arp D, Rieck K, et al. Yes, machine learning can be more secure! A case study on Android malware detection. *IEEE Transactions on Dependable and Secure Computing*, 2019, **16**(4): 711−724

122 Yang W, Kong D, Xie T, Gunter C A. Malware detection in adversarial settings: Exploiting feature evolutions and confusions in Android apps. In: Proceedings of the 33rd Annual Computer Security Applications Conference. 2017.

123 Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, Mcdaniel P. Ensemble adversarial training: Attacks and defenses. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: 2018. 1−20

124 Li D, Li Q. Adversarial deep ensemble: Evasion attacks and defenses for malware detection. *IEEE Transactions on Informa-*

tion Forensics and Security, 2020, **15**: 3886−3900

125 Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. ArXiv: 1503.02531, 2015.

126 Hosseini H, Chen Y, Kannan S, Zhang B, Poovendran R. Blocking transferability of adversarial examples in black-box learning systems. ArXiv: 1703.04318, 2017.

127 Papernot N, Mcdaniel P D. Extending defensive distillation. ArXiv: 1705.05264, 2017.

128 Grosse K, Papernot N, Manoharan P, Backes M, Mcdaniel P D. Adversarial perturbations against deep neural networks for malware classification. ArXiv: 1606.04435, 2016.

129 Stokes J W, Wang D, Marinescu M, Marino M, Bussone B. Attack and defense of dynamic analysis-based, adversarial neural malware detection models. In: Proceedings of the 2018 IEEE Military Communications Conference. Los Angeles, CA, USA: IEEE, 2018. 102−109

130 Cretu G F, Stavrou A, Locasto M E, Stolfo S J, Keromytis A D. Casting out demons: Sanitizing training data for anomaly sensors. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy. Oakland, USA: IEEE, 2008. 81−95

131 Laishram R, Phoha V V. Curie: A method for protecting SVM classifier from poisoning attack. ArXiv: 1606.01584, 2016.

132 Steinhardt J, Koh P W, Liang P. Certified defenses for data poisoning attacks. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Long Beach, USA: MIT Press, 2017. 3518−3530

133 Metzen J H, Genewein T, Fischer V, Bischoff B. On detecting adversarial perturbations. In: The 5th International Conference on Learning Representations. Toulon, France: 2017.

134 Feinman R, Curtin R R, Shintre S, Gardner A B. Detecting adversarial samples from artifacts. ArXiv: 1703.00410, 2017.

135 Cao Y, Yang J. Towards making systems forget with machine unlearning. In: Proceedings of the 36th IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2015. 463−480

136 Bourtoule L, Chandrasekaran V, Choquette-Choo C A, Jia H, Travers A, Zhang B, et al. Machine unlearning. In: The 42nd IEEE Symposium on Security and Privacy. Virtual Event: 2021. 141−159

137 Brückner M, Scheffer T. Nash equilibria of static prediction games. In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2009. 171−179

138 Brückner M, Scheffer T. Stackelberg games for adversarial prediction problems. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, CA, USA: ACM, 2011. 547−555

139 Brückner M, Kanzow C, Scheffer T. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 2012, **13**: 2617−2654

140 Sengupta S, Chakraborti T, Kambhampati S. MTDeep: Boosting the security of deep neural nets against adversarial attacks with moving target defense. In: Proceedings of the 10th International Conference on Decision and Game Theory for Security. Stockholm, Sweden: Springer, 2019. 479−491

141 Biggio B, Fumera G, Roli F. Design of robust classifiers for adversarial environments. In: Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics. Anchorage, USA: IEEE, 2011. 977−982

142 Dwork C. Differential privacy. In: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming. Venice, Italy: Springer, 2006. 1−12

143 Dwork C, Mcsherry F, Nissim K, Smith A D. Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Theory of Cryptography Conference. New York, USA: Springer, 2006. 265−284

144 Mcsherry F, Talwar K. Mechanism design via differential privacy. In: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. Providence, USA: IEEE, 2007. 94−103

145 Dwork C, Roth A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014, **9**: 211−407

146 Zhang Ze-Hui, Fu Yao, Gao Tie-Gang. Research on federated deep neural network model for data privacy protection. *Acta Automatica Sinica*, 2022, **48**(5): 1153−1172
(张泽辉, 富瑶, 高铁杠. 支持数据隐私保护的联邦深度神经网络模型研究. 自动化学报, 2022, **48**(5): 1153−1172)

147 Jayaraman B, Evans D. Evaluating differentially private machine learning in practice. In: Proceedings of the 28th USENIX Security Symposium. Santa Clara, USA: USENIX Association, 2019. 1895−1912

148 Rahman M A, Rahman T, Laganière R, Mohammed N, Wang Y. Membership inference attack against differentially private deep learning model. *Transactions on Data Privacy*, 2018, **11**(1): 61−79

149 Mcmahan H B, Ramage D, Talwar K, Zhang L. Learning differentially private recurrent language models. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: 2018. 1−14

150 Salem A, Zhang Y, Humbert M, Fritz M, Backes M. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In: Proceedings of the 26th Annual Network and Distributed System Security Symposium. San Diego, USA: ISOC, 2019. 1−15

151 Carlini N, Liu C, Erlingsson Ú, Kos J, Song D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In: Proceedings of the 28th USENIX Security Symposium. Santa Clara, USA: USENIX Association, 2019. 267−284

152 Melis L, Song C, Cristofaro E D, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. In: Proceedings of the 2019 IEEE Symposium on Security and Privacy. San Francisco, USA: IEEE, 2019. 691−706

153 Song L, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples. In: Proceedings of the 26th ACM SIGSAC Conference on Computer and Communications Security. London, UK: ACM, 2019. 241−257

154 Ganju K, Wang Q, Yang W, Gunter C A, Borisov N. Property inference attacks on fully connected neural networks using permutation invariant representations. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM, 2018. 619–633

155 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France:

2017.

156 Kipf T, Welling M. Variational graph auto-encoders. ArXiv: 1611.07308, 2016.

157 Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Long Beach, USA: MIT Press, 2017. 1025–1035

158 Hou S, Ye Y, Song Y, Abdulhayoglu M. HinDroid: An intelligent Android malware detection system based on structured heterogeneous information network. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada: ACM, 2017. 1507–1515

159 Ye Y, Hou S, Chen L, Lei J, Wan W, Wang J, et al. Out-of-sample node representation learning for heterogeneous graph in real-time Android malware detection. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: Morgan Kaufmann, 2019. 4150–4156

160 Fan Y, Hou S, Zhang Y, Ye Y, Abdulhayoglu M. Gotcha-sly malware! Scorpion: A metagraph2vec based malware detection system. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London, UK: ACM, 2018. 253–262

161 Zügner D, Akbarnejad A, Günnemann S. Adversarial attacks on neural networks for graph data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London, UK: ACM, 2018. 2847–2856

162 Zhu D, Cui P, Zhang Z, Zhu W. Robust graph convolutional networks against adversarial attacks. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, USA: ACM, 2019. 1399–1407

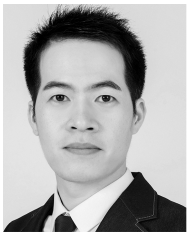
163 Hou S F, Fan Y J, Zhang Y M, Ye Y F, Lei J W, Wan W Q, et al. αCyber: Enhancing robustness of Android malware detection system against adversarial attacks on heterogeneous graph based model. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China: ACM, 2019. 609–618

164 Sun L, Wang J, Yu P S, Li B. Adversarial attack and defense on graph data: A survey. ArXiv: 1812.10528, 2018.

165 Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, USA: ACM, 2017. 3–14

166 Carlini N, Mishra P, Vaidya T, Zhang Y, Sherr M, Shields C, et al. Hidden voice commands. In: Proceedings of the 25th USENIX Security Symposium. Austin, USA: USENIX Association, 2016. 513–530

167 Miller B, Kantchelian A, Afroz S, Bachwani R, Dauber E, Huang L, et al. Adversarial active learning. In: Proceedings of the 2014 ACM Artificial Intelligent and Security Workshop. Scottsdale, USA: ACM, 2014. 3–14



余正飞 国防科技大学系统工程学院博士研究生. 主要研究方向为对抗机器学习和网络安全.
E-mail: yuzhengfei19@nudt.edu.cn
(**YU Zheng-Fei** Ph.D. candidate at the College of Systems Engineering, National University of Defense Technology. His research interest covers adversarial machine learning and network security.)

闫 巧 深圳大学计算机与软件学院教授. 主要研究方向为网络安全和人工智能.
E-mail: yanq@szu.edu.cn
(**YAN Qiao** Professor at the College of Computer Science and Software Engineering, Shenzhen University. Her research interest covers network security and artificial intelligence.)

周 鋆 国防科技大学系统工程学院副教授. 主要研究方向为机器学习和概率图模型. 本文通信作者.
E-mail: zhouyun@nudt.edu.cn
(**ZHOU Yun** Associate professor at the College of Systems Engineering, National University of Defense Technology. His research interest covers machine learning and probabilistic graphical models. Corresponding author of this paper.)