计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202440449
61(10): 2627−2648, 2024

面向人工智能模型的安全攻击和防御策略综述

秦 臻 庄添铭 朱国淞 周尔强 丁 熠 耿 技
(网络与数据安全四川省重点实验室(电子科技大学) 成都 610054)
(zhoueq@uestc.edu.cn)

Survey of Security Attack and Defense Strategies for Artificial Intelligence Model

Qin Zhen, Zhuang Tianming, Zhu Guosong, Zhou Erqiang, Ding Yi, and Geng Ji
(*Network and Data Security Key Laboratory of Sichuan Province* (*University of Electronic Science and Technology of China*), *Chengdu* 610054)

Abstract In recent years, the rapid development of artificial intelligence technology, particularly deep learning, has led to its widespread application in various fields such as computer vision and natural language processing. However, recent research indicates potential security risks associated with these advanced AI models could compromise their reliability. In light of this concern, this survey delves into cutting-edge research findings pertaining to security attacks, attack detection, and defense strategies for artificial intelligence models. Specifically regarding model security attacks, our work focuses on elucidating the principles and technical status of adversarial attacks, model inversion attacks, and model theft attacks. With regards to model attack detection methods explored in this paper, they include defensive distillation techniques, regularization approaches, outlier detection, robust statistics. As for model defense strategies examined in this study, they encompass adversarial training measures, model structure defense mechanisms, query control defenses along with other technical means. This comprehensive survey not only summarizes but also expands upon techniques and methodologies relevant to ensuring the security of artificial intelligence models thereby providing a solid theoretical foundation for their secure applications while simultaneously enabling researchers to gain a better understanding of the current state-of-the-art research within this field facilitating informed decisions when selecting future research directions.

Key words AI; security attack; attack detection; defense strategy; model security

摘 要 近年来,以深度学习为代表的人工智能技术发展迅速,在计算机视觉、自然语言处理等多个领域得到广泛应用.然而,最新研究表明这些先进的人工智能模型存在潜在的安全隐患,可能影响人工智能技术应用的可靠性.为此,深入调研了面向人工智能模型的安全攻击、攻击检测以及防御策略领域中前沿的研究成果.在模型安全攻击方面,聚焦于对抗性攻击、模型反演攻击、模型窃取攻击等方面的原理和技术现状;在模型攻击检测方面,聚焦于防御性蒸馏、正则化、异常值检测、鲁棒统计等检测方法;在模型防御策略方面,聚焦于对抗训练、模型结构防御、查询控制防御等技术手段.概括并扩展了人工智能模型安全相关的技术和方法,为模型的安全应用提供了理论支持.此外,还使研究人员能够更好地理解该领域的当

收稿日期: 2024-05-31; **修回日期**: 2024-07-18
基金项目: 国家自然科学基金项目(62372083, 62072074, 62076054, 62027827, 62002047); 四川省科技计划项目(2024NSFTD0005, 2022JDJQ0039); 中央高校基金(ZYGX2021YGLH212, ZYGX2022YGRH012)
This work was supported by the National Natural Science Foundation of China (62372083, 62072074, 62076054, 62027827, 62002047), the Sichuan Provincial Science and Technology Plan Project (2024NSFTD0005, 2022JDJQ0039), and the Fundamental Research Funds for the Central Universities (ZYGX2021YGLH212, ZYGX2022YGRH012).
通信作者: 周尔强(zhoueq@uestc.edu.cn)

前研究现状,并选择适当的未来研究方向.

关键词 人工智能;安全攻击;攻击检测;防御策略;模型安全

中图法分类号 TP391

近年来,人工智能产业进入高速发展阶段,智能场景融合能力不断提升,人工智能的应用场景从硬件扩展到了软件、服务和平台体系,广泛应用于图像识别[1-2]、智能驾驶[3]、医疗智能诊断[4-5]、智能翻译[6]、数据挖掘[7]等多个领域并取得了十分优异的成绩.我国政府高度重视人工智能的发展,通过发布政策、实施重大项目等方式积极推动人工智能技术和产业创新发展,将人工智能融入国家整体创新体系,不断增强产业竞争力."十四五"数字经济发展规划等多项政策围绕"加快高性能、智能计算中心部署""打造智能算力、通用算法和开发平台一体化的新型智能基础设施"等作出具体部署.在5G、大数据、云计算、深度学习等新技术的共同驱动下,人工智能作为新型基础设施的重要战略性技术加速发展,并与社会各行各业创新融合.

值得注意的是,人工智能技术在高速发展的同时面临着严峻的安全攻击风险.近年来出现了大量试图破坏人工智能模型数据隐私性的攻击.研究者发现人工智能模型使用过程中产生的相关计算数据,包括输出向量、模型参数、模型梯度等,可能会泄露训练数据的敏感信息或者模型自身的属性参数[8].例如,模型逆向攻击[9],攻击者可以在不接触隐私数据的情况下利用模型输出结果等信息来反向推导出用户的隐私数据;成员推断攻击,攻击者可以根据模型的输出判断一个具体的数据是否存在于训练集中.攻击者还可以通过模型输出结果窃取相关模型的参数,对模型拥有者造成严重的经济损失[10].毒化攻击[11]涉及攻击者向模型的训练集中添加毒化数据,导致模型在测试阶段出现可用性或完整性问题.并且可影响各种算法,包括计算机视觉、自然语言处理等,以及衍生出后门攻击,例如使模型错误分类图像.对抗样本攻击[12]通过在数据集中添加微小干扰来误导模型,导致错误预测.这种攻击尤其在安全要求高的领域引发关注,促进了对抗样本攻防技术的研究和发展.模型窃取攻击[13]通过查询黑盒模型并窃取其参数,损害模型所有者的商业利益.成功的模型窃取可能还促使进一步的对抗攻击,增加安全风险.此外,随着联邦学习[14]等分布式学习技术的发展,攻击者有可能成为模型训练过程中的参与方.一般而言,联邦学习中的参与方无法获知彼此的输入数据等隐私信息,但由于攻击者能够获得模型在训练过程中的输出、模型参数和梯度等信息,这大大提升了攻击者的能力,使攻击者窃取其他参与方隐私数据成为可能.

因此,许多研究者提出了各种机制来检测不同的安全攻击类型,包括防御性蒸馏、正则化、数据异常值检测、鲁棒统计方法、差分隐私保护等.防御性蒸馏[15]利用教师模型和学生模型之间的知识传递,通过软化输出来提高模型对细微输入变化的鲁棒性,从而降低对抗样本攻击的效果.正则化技术[16],如L1和L2正则化,通过添加约束条件来防止模型过拟合,从而降低模型通过输出推断敏感训练数据的风险.数据异常值检测技术在识别和防御数据投毒攻击中发挥着关键作用,通过分析数据分布和识别异常点来清洗训练集,保证模型训练的完整性.鲁棒统计方法[17]则通过减轻异常数据点的影响,增强模型在面对恶意数据修改时的韧性和精确性.差分隐私保护[18]通过添加随机噪声来保护数据隐私,减少从统计结果中推断个人信息的风险.此外,随机响应机制[19]和数据集推理技术被用于提升数据隐私保护和验证模型所有权,以防御成员推理攻击和模型窃取.这些技术通过混淆或修改数据,以及通过检验数据集的固有特征,有效地保护了数据的隐私和安全.这些技术综合运用,可以极大提升人工智能系统的安全性和数据保护能力.

同时,许多研究者提出了各种机制来防御针对人工智能系统的安全攻击.通过对模型结构的修改,为输出向量添加特定噪声,结合差分隐私等技术,能够有效防御特定的隐私泄露攻击.文献[20]发现转移知识可以一定程度上降低模型对微小扰动的敏感度,提高人工智能模型的鲁棒性,于是提出将网络蒸馏技术用于防御闪避攻击,并在MNIST和CIFAR-10数据集上测试,发现该技术可使特定攻击(如JSMA)的成功率降低.此外,在模型训练阶段,使用已知的各种攻击方法生成对抗样本[21],再将对抗样本加入模型的训练集中,对模型进行单次或多次重训练,生成可以抵抗攻击扰动的新模型.同时,由于综合多个类型的对抗样本使得训练集数据增多,该技术不但

可以增强新生成模型的鲁棒性，而且可以增强模型的准确率和规范性. 除此之外，也可以通过增强模型的稳定性来防御安全攻击，使模型在功能保持一致的情况下，提升人工智能模型对抗输入扰动的能力. 同时，将上述防御技术进行并行或者串行的整合，能够更有效地对抗不同类型的攻击. 该方法的原理是在模型的使用阶段，通过将输入样本进行变形转化来对抗闪避攻击，变形转化后的输入不会影响模型的正常分类功能. 重构方法包括对输入样本加噪、去噪和使用自动编码器改变输入样本等方法. 针对模型的防御技术关注的是提升模型自身抵抗攻击的能力，在如何提高人工智能系统的鲁棒性方面已经有了诸多研究，主要可分为 2 类：一类是针对模型本身进行改进；另一类是结合其他安全技术来提升人工智能系统的防御能力，如在差异隐私框架内学习和精确分析隐私成本.

本文深入调研了面向人工智能模型的安全攻击、攻击检测以及防御策略领域中前沿的研究成果，主要贡献点总结有 4 点：

1）深入调研了人工智能模型面临的安全攻击，包括数据毒化、数据提取、模型反演和成员推理等. 通过对比这些攻击方式的机制和目标，我们发现它们均能显著影响模型的性能和安全性. 此外，对这些攻击策略进行系统归纳和分析，揭示了它们在侵犯模型隐私和完整性方面的潜在风险，为制定有效的防御措施提供了依据.

2）深入调研了防御性蒸馏、正则化技术、数据异常值检测等多种方法，探讨了它们对抵抗特定攻击类型的效果和局限性. 通过系统归纳不同防御手段，我们发现其能增强模型对对抗性样本的鲁棒性，但对高级攻击仍然脆弱；虽然能减少敏感信息泄露的风险，但可能影响模型准确性. 这些研究为构建更安全的人工智能系统提供了有价值的见解和实践指导.

3）深入调研了人工智能模型的多层防御策略，探讨了如对抗训练增强模型鲁棒性、输入检查与消毒预筛恶意输入、模型结构调整强化内在防御、查询控制减少信息泄露等策略. 通过比较这些策略的实施效果与局限性，总结发现尽管这些策略在应对安全威胁方面具有潜力，但在复杂攻击面前仍存挑战，为未来的研究方向提供了指引.

4）深入调研了人工智能技术的主要安全挑战和未来展望，讨论了医疗、社交媒体、自动驾驶等关键领域面临的风险，并指出数据隐私保护、抵御对抗性攻击、增强模型透明度和道德伦理的研究需求. 通过这些分析，明确了未来人工智能安全研究的方向，为制定有效的防御策略提供了理论支持.

1 人工智能模型安全攻击

人工智能模型面临的安全攻击类型多种多样. 数据毒化攻击旨在通过标签翻转、双层中毒和清洁标签攻击等方式降低模型性能，误分类特定样本或泄露私密信息；数据提取攻击试图从模型反演训练数据的特定特征或统计性质；成员推理攻击依赖于模型、数据和攻击技巧的复杂交互，通过推测数据点是否属于训练集来威胁隐私安全. 这些攻击方式类型不同，但是都给人工智能模型带来了重大的安全问题. 本节将会具体介绍主流的面向人工智能模型的安全攻击.

1.1 对抗性攻击

对抗性攻击（adversarial attacks）是一种针对机器学习模型的攻击手段，其核心目的在于通过微小但有针对性的修改输入数据，从而欺骗模型，导致其产生误分类或错误的输出结果. 这种攻击手段多种多样，其中包括添加对人类难以察觉的噪声、修改输入数据的特征，甚至是利用模型存在的漏洞来干扰其输出. 不仅如此，对抗性攻击并不仅局限于图像分类模型，它同样适用于语音识别、自然语言处理等各个领域的机器学习模型. 这种攻击的危害性在于，它不仅能够降低模型的准确性和可靠性，而且可能对模型的安全性造成严重威胁，特别是在需要高度信任和可靠性的应用场景下，如自动驾驶、医疗诊断等. 因此，为确保机器学习模型在实际应用中的稳健性和安全性，对抗性攻击的研究和防御显得至关重要. 对抗性攻击的一般形式如图 1 所示.

对抗性攻击根据攻击者对深度学习模型内部信息的了解程度可以分为白盒攻击（white-box attack）和黑盒攻击（black-box attack）.

在白盒攻击中，攻击者具备对目标模型内部结构、参数等详细信息的深入了解. 这种了解的程度使得攻击者能够深入观察模型的运行机制，并且甚至能够修改输入数据以及模型参数，以精心设计并执行攻击策略. 相比于黑盒攻击，白盒攻击更具有挑战性，因为攻击者能够利用对模型内部结构的了解来更有效地发现和利用潜在的漏洞和弱点. 很多白盒攻击实际上是基于梯度的方法. 对抗样本的概念于 2014 年被提出[21]，因深度神经网络学习的输入-输出映射在很大程度是相当不连续的，所以可通过应用

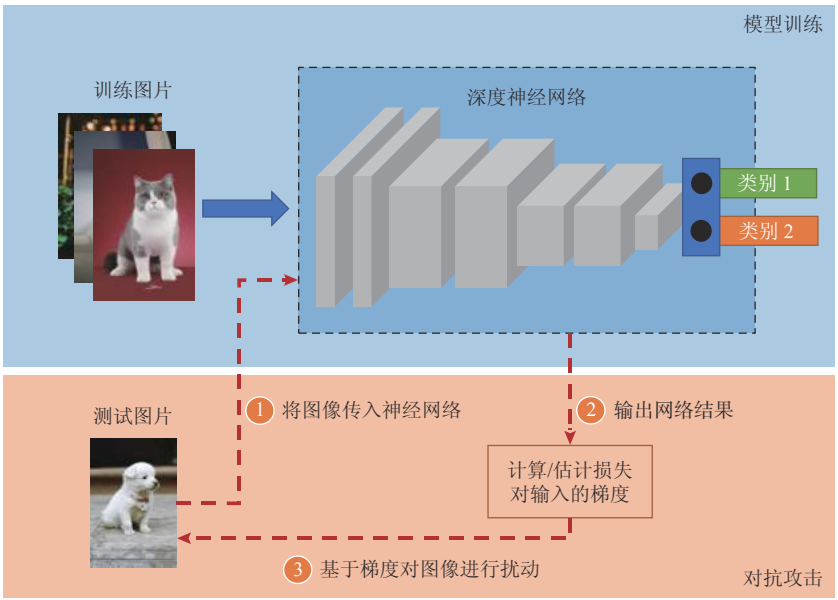


Fig. 1 General form of adversarial attacks

图 1 对抗性攻击的一般形式

某些难以感知的扰动来使网络对图像分类错误，该扰动是通过最大化网络的预测误差来发现的. 文献 [22] 提出了一个更加高效的优化问题，能够以添加更小扰动的代价得到更加高效的对抗样本.

在黑盒攻击中，攻击者的挑战在于缺乏对目标模型内部细节的了解. 他们无法获得模型的训练数据、参数或者结构信息，因此需要通过观察模型的输入-输出关系来推断其内部运作机制. 这意味着攻击者需要采取一系列试错和推理的方法，通过不断调整输入数据以观察模型的响应，并从中推导出模型的内在规律和薄弱点. 对抗性攻击中的黑盒攻击在现实世界中出现的概率更大，因为许多机器学习模型的内部结构和参数是保密的，攻击者只能通过模型的输入输出接口进行攻击. 一些黑盒攻击使用数值估算梯度的预测，从而得到合适的对抗扰动. 文献 [12] 利用了贝叶斯优化的高效查询找到成功的对抗扰动.

1.2 模型反演攻击

模型反演攻击(model inversion attack)，又称模型逆向攻击，是一种针对机器学习模型的攻击手段，其核心目的在于通过利用模型的输出结果推断其训练数据或其他敏感信息. 攻击者通过观察模型的输出，并根据输出结果推断出模型所使用的数据或参数，从而可能泄露隐私信息或者揭示模型的敏感特性. 这种攻击通常涉及到逆向工程和推理技术，以推断出模型背后的数据或参数可能对模型的隐私性和安全性造成威胁.

根据攻击者攻击目标的不同，可以将模型反演攻击分为逆向推导训练集敏感属性、生成训练集代表类数据和重构目标模型输入数据等 3 种.

逆向推导训练集敏感属性，在这种攻击方式中，攻击者试图通过模型的输出结果来推导出训练集中的敏感属性信息. 通过分析模型的输出，攻击者可以逆向推导出训练数据中可能存在的敏感信息，如个人身份信息、财务数据等. 文献 [9] 设计了一种新颖的基于置信度得分的模型反演属性推理攻击，在攻击者知道训练数据中记录的非敏感属性时有明显的攻击，并将其扩展到了任何属性都未知的训练数据中.

生成训练集代表类数据，在这种攻击方式中，攻击者尝试生成与训练集相似但并非真实数据的样本以欺骗模型. 通过生成具有相似特征分布的数据样本，攻击者可以使模型产生误判，从而影响其性能和可信度. 文献 [23] 设计了一种新型的基于预测置信度的模型反演攻击，在生活方式调查决策树和用于面部识别的神经网络 2 种情况下进行模拟攻击，成功得到了训练数据的相似结果.

重构目标模型输入数据，这种攻击方式旨在通过分析模型的输出结果，重构出模型的输入数据. 通过逆向工程模型的输入数据，攻击者可以了解模型的工作原理和内部结构，进而设计更有效的攻击策略. 文献 [24] 在对抗性设置下，通过对目标模型的黑盒访问来训练反演模型，提出了 2 种攻击方式，分别是基于背景知识组成辅助集和截断技术来训练反演模型.

1.3 模型窃取攻击

模型窃取攻击(model stealing attack)是一种隐蔽而具有挑战性的攻击手段，其背后的目的是获取目标模型的知识或敏感信息. 攻击者通过观察目标模型的行为或利用其输出来尝试复制或重建原始模型，从而获取其内部结构、参数或训练数据的近似信息. 这种攻击方式通常涉及多个步骤：攻击者可能会通过向目标模型提交查询请求并分析其输出来观察模型的行为. 通过观察模型对不同输入的响应方式，攻击者可以尝试推断出目标模型的内部运行机制和结构. 同时，攻击者还可以利用目标模型的输出数据来训练自己的模型. 通过收集大量的模型输出和相应的输入数据，攻击者可以尝试通过机器学习或其他建模技术来逼近目标模型的功能，进而推断出目标模型的内部特征. 这种攻击方式不仅可能泄露模型的结构和参数信息，而且可能导致敏感数据的泄露，对模型的安全性和隐私性构成威胁. 模型窃取攻击的一般形式如图 2 所示.

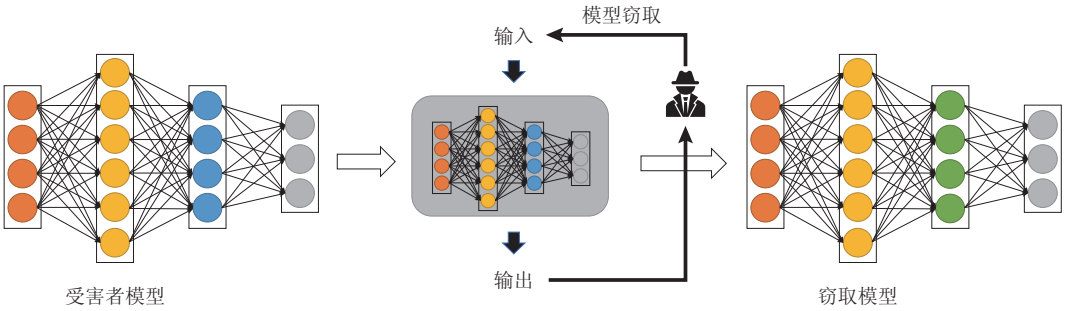


Fig. 2 Schematic diagram of model theft attack

图 2 模型窃取攻击示意图

根据攻击者的攻击目的不同，可以将模型窃取攻击分为模型信息窃取和模型功能窃取.

模型信息窃取攻击的目的是获取目标模型的内部信息，如参数、权重、结构或训练数据. 攻击者可能利用这些信息来重建目标模型，进行逆向工程，或者用于其他恶意目的. 这一类型的攻击可能导致知识产权泄露、竞争风险以及隐私问题. 模型信息的种类繁多，因此有多种方法来窃取不同类型的模型信息. 在模型训练超参数的窃取攻击方面，文献 [13] 提出了一种方程求解方法，用于从岭回归、逻辑回归、支持向量机和神经网络中窃取正则化超参数. 在结构信息的窃取攻击方面，文献 [25] 利用侧通道访问模型，即使用未加密的 PCIe 流量，以窃取受害者深度神经网络(deep neural network, DNN)模型. 在结构已知的模型的参数窃取攻击方面，文献 [26] 提出了一种求解方程的攻击，允许提取逻辑回归和多层感知器的精确参数，也适用于具有线性或二次核的支持向量回归.

模型功能窃取攻击的目的是获取目标模型的预测能力或功能，而不一定需要获取模型的内部结构或参数. 攻击者可能试图通过观察模型的输出来近似其功能，或者利用攻击者自己的模型来模仿目标模型的行为. 这一类型的攻击可能导致盗版产品、欺诈行为或者降低模型的商业价值. 文献 [27] 提出了将主动学习应用在模型窃取攻击方面，使用与目标模型相同架构的模型来获得具有相似有效性表现的被盗模型. 文献 [28] 则是在只能与目标模型黑盒交互的前提下，使用与目标模型类型一致但不同结构的模型来获得具有相似有效性表现的被盗模型.

1.4 毒化攻击

在数据毒化攻击中，攻击者的意图主要包括恶意活动未被检测但不影响正常操作的安全性侵犯；正常功能受损导致合法用户的服务中断的可用性侵犯或攻击者试图获取系统、用户或数据的私密信息的隐私侵犯，攻击者们希望模型受到攻击后，将样本错误地分类为特定类别. 主流的攻击策略有：无差别攻击、目标中毒攻击以及后门攻击.

在无差别攻击中，主要包含标签翻转攻击、双层中毒攻击以及清洁标签攻击. 标签翻转攻击是一种最简单的毒化攻击策略，最初由文献 [10] 中提出. 如图 3 所示，攻击者不改变特征值，而是错误地标记训练数据集的一部分样本，从而降低如支持向量机等机器学习模型的性能准确性.

双层攻击的特点是攻击者可以操纵大量训练样本的特征，但同时保持其原始类别标签不变. 这种攻击最早由文献 [29] 提出，是针对支持向量机(support vector machine, SVM)的梯度基础，利用隐式微分求得优化中毒样本所需梯度的一种无差别攻击方式. 文

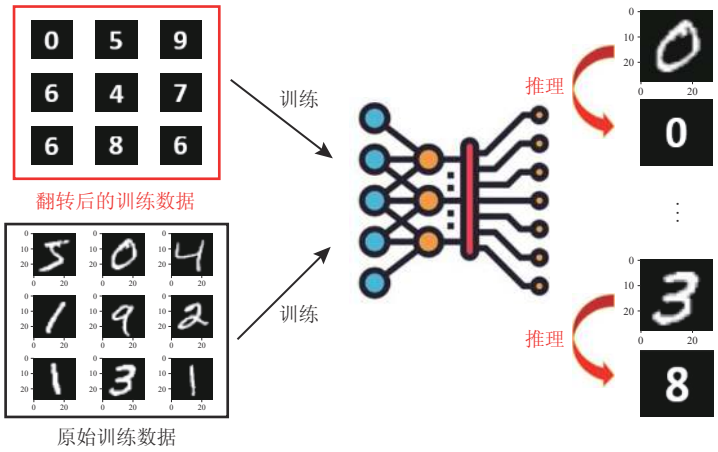


Fig. 3 Classification errors caused by tag flipping attacks

图 3 标签翻转攻击导致的分类错误

献 [30] 发现简单的启发式策略结合变量减少技术可以有效地对抗线性分类器，同时提高计算效率，该文献指出以前基于梯度的方法可能受到损失景观等因素的影响，导致计算时间和攻击效率下降. 文献 [31] 针对大型模型的中毒攻击提出了一种新的方法，即适应超参数优化技术，以更有效地解决双层编程问题，该文献提出了一种反向梯度下降技术来优化中毒样本，显著降低了攻击复杂性，探讨了计算验证损失关于中毒样本的解析或估计梯度的计算和查询成本，并提出训练生成模型以增加中毒样本的生成速率.

清洁标签攻击是指攻击者可以控制大量训练数据，并微调每个中毒样本以保持其类别标签不变. 文献 [32] 首次提出了这一概念，指出操纵整个训练集来任意定义单个特征对凸学习器预测的重要性. 后来，文献 [33] 提出了针对深度神经网络的新型清洁标签中毒攻击技术，攻击者训练了一个生成模型来创建能破坏受害者模型的清洁标签中毒样本. 受到这种方式的启发，这些攻击方式效果很好，但是在许多应用中攻击者很难控制如此大比例的训练数据，因此，寻找可扩展、有效且实际的针对 DNN 的无差别中毒攻击仍是亟待解决的问题.

样本特定攻击只对特定目标样本造成误分类，不影响系统正常用户的功能和行为，与无差别中毒攻击相似，目标性攻击也篡改训练数据，有区别的地方在于不需要修改测试数据. 样本特定攻击的双层毒化攻击与无差别攻击中的双层毒化攻击类似，由文献 [34] 引入，以证明基于梯度的中毒攻击与影响函数的等价性，即在鲁棒统计领域定义的用于识别影响特定预测的最重要训练点的函数. 文献 [34] 的攻击专注于误导模型只对某些特定的测试样本，考虑的是测试集，而不是验证集. 文献 [35] 的工作主要集中在将目标中毒攻击泛化到数据分布的整个子群体，同时降低了计算成本. 相比之下，该方法的攻击速度更快，但在所有子群体上的成功率可能不如文献 [34] 中的方法.

后门中毒攻击旨在破坏模型的完整性，攻击者通过后门触发器使测试样本被错误分类，但不影响干净样本的分类. 在后门中毒攻击中主流的方式为触发中毒，在这一部分中不同的研究者和工作重点如下：文献 [36] 研究了攻击者如何通过篡改训练数据来植入后门，其中攻击者随机选取一部分训练数据，将后门触发器融入其中并改变其对应标签；文献 [37] 的工作设计了使选定内部神经元响应最大化的后门触发器，从而减少植入触发器所需的训练数据；文献 [37] 通过利用未编译量子神经网络与其近似合成版本之间的单元差异，最大化合成量子神经网络电路中选定神经元的响应，以确保触发器出现时模型能准确输出预定义的目标类；文献 [38] 针对毫米波波束选择系统，提出了一种基于神经网络的后门攻击方法，通过最大化选定神经元的响应来设计后门触发器，并提出了一种机器学习方法来减轻后门效应，确保模型在新的无线通信环境中仍能有效工作；文献 [39] 在假设攻击者知识有限的更实际场景中，利用固定模式的融合来针对模型植入后门.

数据毒化攻击只需要向训练数据中加入少量毒化数据便可持久性地影响模型识别结果，同时可针对各类人工智能模型，是一种易于实现的通用攻击方式.

1.5 数据提取攻击

数据提取攻击的主要目的是获取训练数据中的特定特征或训练数据的某些统计性质，常见的攻击方式为模型反演攻击. 这些攻击主要分为 2 类：属性推断和个体重建.

基于属性推断的攻击方式专注于模型的预测能力和对输入数据的敏感性，针对分类任务，在对象识别和面部识别中通过各类指标评估了模型反演攻击的效果，在不完全访问模型的情况下，如何从模型输出中恢复个体敏感信息，通过模型的预测结果推测输入数据的敏感属性.

文献 [38] 展示了白盒攻击在人员识别和对象检测中的应用，揭示了模型内部参数可以被用来逆向推演训练数据. 该研究基于卷积神经网络，涉及个体反演，即尝试重建个人的输入数据，主要被应用于人物识别和对象检测中. 此外，探讨了如何定制优化函数以实现更好的训练实例重建，在这种方法中，逆向工程的目标被转化为基于梯度的优化问题，旨在将目标模型的输出反向映射回其输入（即重建训练样本），而无需额外训练模型来处理逆向任务，这一过程通过迭代优化神经网络模型的学习目标，以生成对训练实例的更好估计或重建.

数据提取攻击通常较为高效且具有较强的针对性，但需要大量的计算资源，在实际应用中成功率往往相对较低. 图 4 列出了迄今为止已研究过的具有代表性的各类算法的部分攻击类型分布图，红色方框表示未发现攻击，绿色方框表示存在攻击，M 代表成员推理攻击、E 表示模型提取攻击、P 表示数据推理攻击、R 表示模型重建.

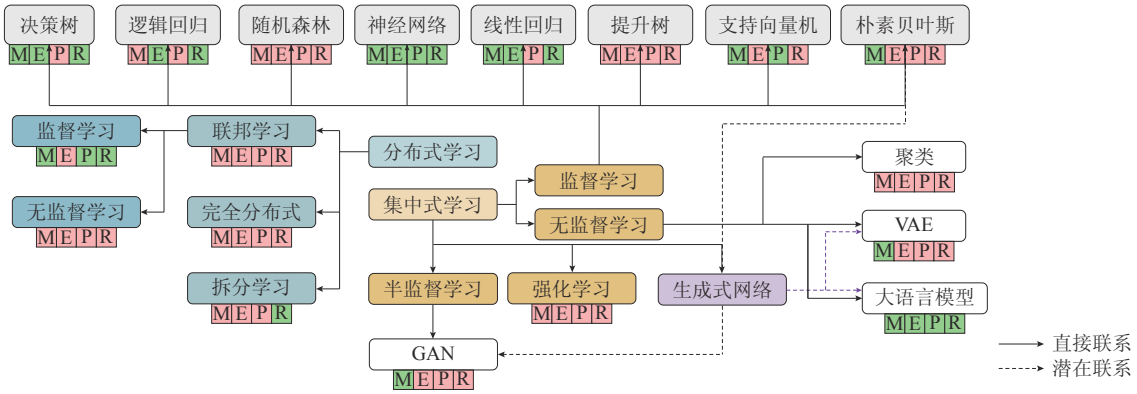


Fig. 4 Distribution of attack types the algorithms suffered

图 4 算法遭受的攻击类型分布

1.6 成员推理攻击

成员推理攻击是一种黑盒攻击方式，主要目的是推测某个数据点是否属于模型的训练数据集. 攻击的准确性受模型泛化误差、模型复杂度、数据集结构等因素影响，过度拟合虽然可以增加攻击成功的机会，但不是必要条件.

在成员推理攻击中，主要的方式是构造影子模型来模拟目标模型的行为进行成员推理攻击，这些模型的输出可以是预测向量或单一标签. 攻击数据集的标签来自影子数据的训练和测试分割，其中属于测试集的数据点被标记为非训练集成员. 元模型通过学习目标模型输出的模式来识别数据点是否属于训练数据集. 攻击的准确性受到影子模型数量的影响，但更多的模型也意味着更高的成本. 文献 [40] 证明仅使用 1 个影子模型就足以进行此类攻击；文献 [41] 提出了一种基于贝叶斯优化的成员推理攻击，该攻击仅依赖于损失，并优于先前的攻击方法，该文献的最高攻击准确率可达 90.8%，针对的是在 ImageNet 数据集上进行分类的大型神经网络模型 VGG16. 此外，对于线性模型，攻击者需要训练一个与目标模型相同的代理模型，用于计算与白盒模型权重的差异，然后用于成员推理. 对于深度神经网络，每个层都被局部线性近似替换，以类似的方式进行攻击，这种攻击是数据驱动的，即使目标模型不同于影子模型和元模型，攻击也可能成功.

影子模型训练需要影子数据集，其过程如图 5 所示. 在针对监督学习模型的成员推理攻击的主要假设中，攻击者对训练样本没有或有限的知识，但他们了解目标训练数据的基本数据分布. 如果攻击者无法获取合适的数据集，他们可以通过访问多个特征的概率分布统计信息，使用采样技术创建影子数据集，如果基于统计的生成不可行，那么使用目标模型预测向量的查询方法也是一个选择. 对于序列到序列模型，如机器翻译，研究发现这些模型比图像分类等任务更难被攻击，但在领域外或词汇外数据的会员推断上较为容易[42]. 对于深度生成模型，如生成对抗网络（generative adversarial network, GAN）和变分自编码器（variant auto-encoder, VAE），成员推理攻击也

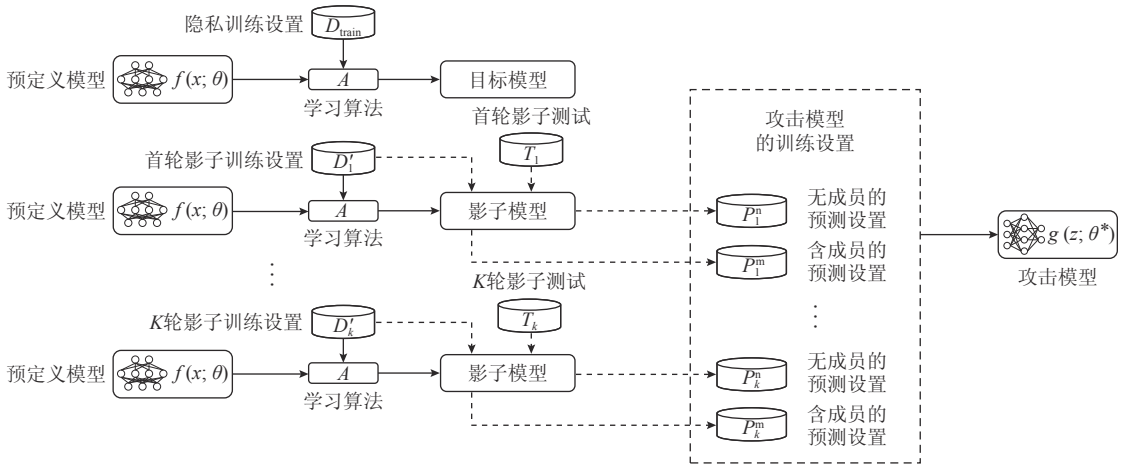


Fig. 5 Training process of shadow model

图 5 影子模型的训练过程

是适用的. 由于这些模型有多个组件, 攻击威胁建模需要考虑这些因素. 文献 [43] 提出了适用于部分黑盒设置的一系列攻击, 即使没有辅助数据, 攻击者也可以训练一个辅助 GAN, 其判别器能区分目标生成器和辅助生成器产生的数据. 如果攻击者有辅助信息, 例如关于目标的训练和测试数据的知识, 他们可以训练另一个 GAN, 其判别器能够区分原始训练集的成员和非成员.

成员推理攻击的效果受到模型类型、训练数据质量和数量等多种因素的影响, 且需要攻击者对于人工智能模型与训练过程有着一定的先验知识, 但现有的针对成员推理攻击的防御方式往往依赖对模型性能的牺牲, 这使得成员推理攻击是一种防御较大的攻击方式.

本节介绍并总结了主流的人工智能模型攻击方式, 包括对抗性攻击、模型反演攻击、模型窃取攻击等, 这些攻击方法均能显著影响人工智能模型的性能和安全性. 表 1 对上述提到的人工智能模型攻击方法进行对比和归纳, 从而使研究人员能够更好地理解相关类型的攻击方式.

Table 1 Types and Characteristics of Security Attacks on Different Artificial Intelligence Models

表 1 对不同人工智能模型的安全攻击类型及特点

安全攻击类型		特点
对抗性攻击		通过微小但有针对性的修改输入数据，从而欺骗模型，导致其产生误分类或错误的输出结果
模型反演攻击		通过利用模型的输出结果推断其训练数据或其他敏感信息
模型窃取攻击		通过观察目标模型的行为或利用其输出来尝试复制或重建原始模型，从而获取其内部结构、参数或训练数据的近似信息
数据毒化攻击	标签翻转	攻击者错误标记部分训练数据的样本标签，不改变特征值，降低模型准确性
	双层中毒	攻击者操纵大量训练样本的特征，保持原始类别标签不变，主要针对 SVM 进行梯度基础攻击
	清洁标签	攻击者控制训练数据并微调每个中毒样本以保持其类别标签不变，针对深度神经网络
	样本特定	只对特定目标样本造成误分类，不修改测试数据，利用双层中毒攻击和梯度基础攻击
	后门中毒	通过后门触发器使测试样本被错误分类，不影响干净样本的分类，涉及隐形触发器和固定模式的融合
数据提取攻击	属性推断	专注于模型的预测能力和对输入数据的敏感性，通过模型的预测结果推测输入数据的敏感属性
	个体重建	通过定制的优化函数实现训练的实例重建，此逆向工程的目标通常被转化为基于梯度的优化问题
成员推理攻击	影子模型	通过构造影子模型来推测数据点是否属于训练数据集，受模型泛化误差等因素影响，可能需要训练与目标模型相似的模型以进行攻击

2 人工智能模型攻击检测

随着人工智能和机器学习技术的快速发展与广泛应用, 攻击者利用对抗性样本、模型反演攻击、毒化攻击等手段, 不断寻找突破机器学习模型防线的方法, 这对数据安全和用户隐私构成了严重威胁. 因此, 研究并开发有效的检查及防御机制, 以提高模型的鲁棒性和防御能力, 对于保护机器学习系统免受这些恶意攻击至关重要. 本节内容围绕机器学习模型的安全性检测和防御技术展开讨论, 涵盖了多种防御策略和方法.

2.1 防御性蒸馏

防御性蒸馏是一种提高机器学习模型对对抗性攻击鲁棒性的技术. 这种方法最初由文献 [15] 提出, 灵感来源于知识蒸馏, 即一个复杂模型(教师模型)的知识可以通过软化输出(例如, 使用温度缩放的 softmax 函数)传递给一个更简单的模型(学生模型), 使得后者对输入数据的小变动不敏感, 从而提高其抵抗对抗性样本的能力. 在防御性蒸馏的背景下, 这

种技术被用来减少模型对细微输入变化的敏感性，通过训练学生模型来复制教师模型的软输出，从而使得生成有效的对抗性样本更加困难.

如图 6 所示，在防御蒸馏架构中，教师模型的输出（通常是概率分布，而非硬标签）被用作学生模型的目标输出. 这些软标签包含了关于类别之间相对关系的信息，比硬标签提供了更多的细节，使得学生模型能够学习到更加平滑的决策边界.

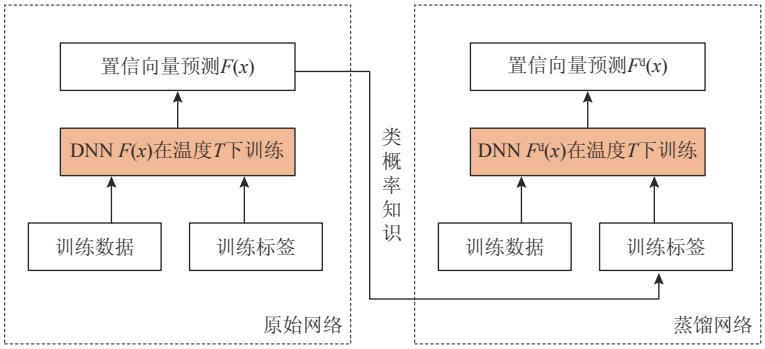


Fig. 6 Defense distillation architecture diagram

图 6 防御蒸馏架构图

在这个过程中，教师模型的输出（通常是概率分布，而非硬标签）被用作学生模型的目标输出. 这些软标签包含了关于类别之间相对关系的信息，比硬标签提供了更多的细节，使得学生模型能够学习到更加平滑的决策边界.

文献 [15] 的研究表明，通过防御性蒸馏训练的模型对于多种类型的对抗性样本更为鲁棒. 在实验中使用蒸馏技术可以显著减少对抗性样本对模型准确度的影响. 然而，文献 [44] 证明，尽管蒸馏可以提高一些模型的鲁棒性，但对于高级对抗性攻击仍然脆弱. 此外，这种防御机制可能导致模型准确性的降低. 模型需要对抗性地适应更加平滑的标签分布，可能与实际分布产生偏差.

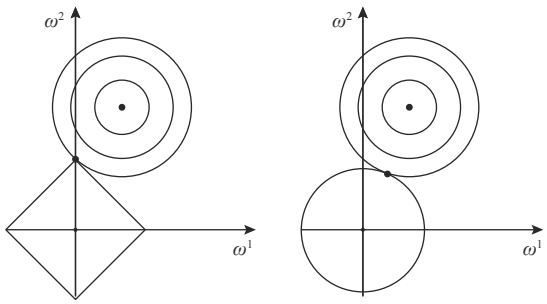
尽管如此，防御性蒸馏作为提高模型安全性的手段之一，为模型防御提供了有价值的见解. 它强调了在模型训练过程中考虑安全性的重要性，以及通过模型输出的精细化管理来提高鲁棒性的潜力. 未来的研究可以探索如何结合蒸馏与其他防御机制，如对抗性训练或输入预处理，构建更加安全的人工智能系统.

2.2 正则化

正则化技术在提高机器学习模型对于模型反演攻击的防御能力方面起着关键作用. 模型反演攻击旨在从模型的输出推断训练数据的敏感属性，因此提高模型对此类攻击的鲁棒性是保护隐私的重要方面.

正则化通过引入额外的约束或惩罚项到模型的训练过程中（如 L1 和 L2 正则化）来防止模型过拟合，从而降低攻击者通过模型输出推断原始训练数据的能力.

如图 7 所示，L1 正则化鼓励参数稀疏性，因此减少模型依赖的特征数量，从而抵御模型安全攻击；L2 正则化可以减少模型依赖的特征数量，使得模型对输入数据的小扰动不太敏感，从而抵御模型安全攻击.



(a) L1正侧化 (b) L2正侧化

Fig. 7 L1 and L2 regularization

图 7 L1 和 L2 正则化

特定类型的正则化，如差分隐私（differential privacy, DP）正则化，可以有效降低模型反演攻击的风险. 文献 [16] 提出了一种差分隐私训练深度学习模型的方法，该方法通过添加随机噪声来保护训练数据的隐私，同时仍然允许模型学习到有用的特征. 虽然这种方法在一定程度上牺牲了模型的准确度，但它提供了一种权衡，通过增加隐私保护来降低模型反演攻击的风险.

进一步地，文献 [45] 研究了正则化如何影响深度神经网络的隐私保护，特别是在面对模型反演攻

击时. 研究显示，适当的正则化可以显著提高模型的隐私保护能力，减少敏感信息的泄露. 此外还指出，正则化能够减轻模型反演攻击的影响，但并不能完全消除风险. 因此，正则化应该与其他隐私保护技术结合使用以增强防御.

总体而言，正则化是增强模型在面对模型反演攻击时鲁棒性的有效手段之一. 它通过减少模型对训练数据的过度依赖来降低敏感信息泄露的风险. 与其他防御策略一起使用，将能提供更全面的保护.

2.3 数据异常值检测

在机器学习模型的安全防御领域，数据异常值检测技术起着至关重要的作用，尤其是在识别和防御毒化攻击方面. 毒化攻击通过向训练集注入恶意数据来破坏模型的性能，这使得异常值检测成为防御此类攻击的关键手段.

文献 [15] 提供了异常值检测方法的广泛概述，这些方法可以用于识别可能破坏模型完整性的异常数据点. 文献 [46] 提出一种基于数据清洗的方法，通过识别和删除训练数据中的异常值来防御攻击. 该方法依赖于对数据分布的统计分析，以及对可能的攻击模式的理解，从而能够有效地识别被篡改的数据点. 研究探讨了数据集中毒的概念，并展示了数据异常值检测在识别和减轻此类攻击方面的有效性. 这种方法展示了通过细致的数据预处理来增强模型安全性的潜力.

除了传统的统计方法，深度学习技术也被应用于异常值检测. 文献 [47] 探索了利用深度学习模型进行异常值检测的有效性，特别是在处理高维数据和复杂数据分布时. 研究表明，深度学习模型可以学习到数据的复杂特征表示，从而更准确地识别异常样本. 这些发现为开发新一代异常值检测工具提供了理论基础，能够在更广泛的应用场景中提供保护.

虽然数据异常值检测提供了对抗毒化攻击的一种有效手段，但它并非无懈可击. 特别是，这种方法可能无法检测到那些经过精心设计以模仿正常数据行为的恶意数据. 此外，过于敏感的异常值检测系统可能会导致高误报率，误将合法数据误判为异常，从而干扰模型的训练和性能.

2.4 鲁棒统计方法

在机器学习模型的安全防御领域，鲁棒统计方法提供了一种有效途径来防御毒化攻击，特别是在数据可能被恶意修改以影响模型训练和决策过程的情况下. 这些方法旨在通过识别和减轻异常数据点的影响，增强模型对攻击的韧性，确保数据的完整性和模型的准确性.

文献 [48] 的工作在理论上分析了鲁棒统计估计器在面对含有恶意数据的数据集时的性能，提出的算法能够有效地从被污染的数据集中估计出准确的统计量，即使数据集中有一定比例的数据被篡改. 文献 [46] 研究了数据集中毒攻击和如何使用鲁棒统计方法来识别和剔除恶意数据点，从而防止这些数据影响模型的训练. 该工作强调了在训练过程中考虑数据完整性的重要性，并提供了一个框架来识别可能的数据污染. 文献 [18] 探讨了在深度学习模型中实施鲁棒性改进的方法，特别是针对毒化攻击的防御. 该研究集中在设计能够识别并忽略训练过程中恶意输入的神经网络架构上. 通过在模型训练阶段引入额外的鲁棒性约束，能够有效地降低毒化样本对最终模型性能的影响.

鲁棒统计方法的主要技术包括使用中位数和其他鲁棒性度量代替均值及标准差、利用 M 估计器进行参数估计，以及应用断点和抗污染技术来识别和排除异常值等. 这些技术能够帮助减少异常数据点对模型的影响，增加模型对攻击的鲁棒性. 但鲁棒统计方法也存在其局限性. 实际中，完全区分正常数据和异常数据可能很困难，特别是在高维数据和复杂数据结构的情况下. 此外，某些鲁棒统计技术可能会导致计算成本增加，或在去除异常值时牺牲有用信息.

2.5 差分隐私

数据提取攻击(data extraction attacks)是指攻击者试图从某个系统、数据库或数据集中非法提取或推断出敏感信息的一种攻击手段. 差分隐私通过为大型数据集添加随机噪声来防止攻击者从统计结果中提取出任何个人的具体信息. 图数据的关键是图记录的表示和存储，其中节点表示数据集实体，边表示实体之间的关系. 这也产生了图差分隐私的 2 个不同概念：边差分隐私和节点差分隐私. 差分隐私介绍图如图 8 所示.

边差分隐私机制的关键构建块包括拉普拉斯机制和指数机制. 拉普拉斯机制通过与不同类型的敏

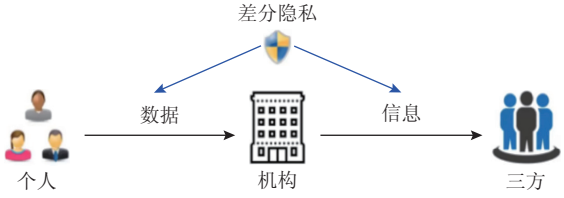


Fig. 8 Introduction diagram of differential privacy

图 8 差分隐私介绍图

感性结合使用，计算出需要添加的噪声量. 而指数机制通过为每个可能的输出分配一个与其“实用性”成正比的概率来保护隐私，通常用于参数选择或非数字输出（如分类标签、排序等）. 文献 [49] 提出了一种基于有界拉普拉斯机制的用于保护无向无权图的代数连通性的边差分隐私机制，并证明了估计的准确性. 但其方法和分析主要集中在特定类型的图（无向无权图）上，可能不适用于更复杂或不同类型的网络结构.

节点差分隐私专注于保护图数据中单个节点的隐私，使得添加或移除图中的一个节点后对外提供的统计信息只会受到很小的影响. 文献 [50] 提出了一种节点级扰动算法，通过随机添加和删除节点来修改输入图，实验表明在保持输入图的多个统计特性的同时提供了合理的隐私保护，但这些方法引入了较大的噪声，这可能会影响最终结果的实用性和准确性.

边权重差分隐私关注公开图上的隐私权重数据，例如导航系统. 文献 [51] 提出了一种基于小波变换的差分隐私算法和增强数据效用算法，前者通过在边权重上添加较少的噪声实现差分隐私，而后者则维持原有加权社交网络的属性，以提高数据效用.

2.6 随机响应机制

成员推理攻击试图确定某个特定的个体数据是否被用于机器学习模型的训练过程，这可能会暴露个人的隐私信息. 随机响应机制（randomized response technique）通过在回答敏感问题时引入随机化来解决这个问题.

这一机制自 1965 年由文献 [18] 提出，目前已被广泛地应用于调查研究中. 它提出了一种基于概率的回答方法，允许受访者通过随机化他们的回答来保持隐私，同时还能提供无偏的最大似然估计，特别适用于可能引起回避性回答的敏感问题调查，通过减少回答的直接性，旨在提高调查的准确性和参与者的合作性. 其算法启发了后续研究者，并被应用在差分隐私技术中，目前随机响应技术是本地化差分隐私保护技术的主流扰动机制. 本地化差分隐私中，每个用户将各自的数据进行扰动后，再上传至数据收集者处，而任意 2 个用户之间并不知晓对方的数据记录，本地化差分隐私中并不存在全局敏感性的概念.

例如文献 [52] 提出了一个基于扩展随机响应（extended randomized response）的本地混淆策略，实现本地差分隐私（local differential privacy），旨在通过利用位置-费用相关性，实现对工人位置和费用信息的联合隐私保护. 首先，通过随机响应来混淆工人的位置信息；然后，基于位置混淆的结果，使用条件随机响应进一步有效混淆收费信息. 这一策略不仅满足本地差分隐私要求，还考虑了空间众包环境中的特定需求，即同时保护工人的位置和费用信息.

2.7 数据集推理

模型窃取攻击（model stealing attack）是一种针对机器学习模型的攻击方法，如图 9 所示. 在这种攻击中，攻击者通过向目标模型发送查询并使用模型的响应来构建一个功能上等效的复制模型. 这种攻击可以用来低成本复制一个有效的模型，或者用来简化设计其他攻击（如对抗样本、成员资料推理、对抗性重编程等）的过程. 模型窃取攻击特别适用于黑盒设置中，即攻击者没有关于目标模型内部工作原理的直接知识. 防范模型窃取攻击的相关研究很少，例如数据集推理技术（dataset inference technique, DIT）是一种新颖的解决机器学习所有权问题的方法，其原理基于识别疑似被窃取模型是否包含了原始模型数据集中的私有知识. 这种技术结合了统计测试和估算多个数据点到决策边界的距离的能力. 与之前的方法不同，DIT 无需对模型进行重新训练或过度拟合，即可有效防御，即使是适应性攻击，无需牺牲模型准确性或复杂修改模型架构.

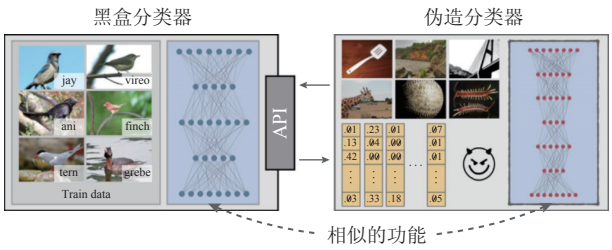


Fig. 9 Schematic diagram of model theft attack

图 9 模型窃取攻击示意图

2.8 验证嵌入外部特征

获得一个经过良好训练的模型涉及昂贵的数据收集和训练过程，因此该模型是一个有价值的知识产权. 模型窃取攻击是一种针对机器学习模型的攻击方法. 在这种攻击中，攻击者通过向目标模型发送查询并使用模型的响应来构建一个功能上等效的复制模型. 这种攻击可以用来低成本复制一个有效的模型，或者用来简化设计其他攻击（如对抗样本、成员资料推理、对抗性重编程等）的过程. 模型窃取攻击特别适用于黑盒设置中，即攻击者没有关于目标

模型内部工作原理的直接知识.

关于模型窃取的防御问题，一种新的思路是验证可疑模型是否具有防御者指定的行为. 如果模型有这样的行为，我们将其视为从受害者那里窃取. 这种方法的灵感来自于这样一种理解，即被盗模型应该包含受害者模型学习的特征知识，因此它们具有相似的行为. 数据集推理专注于这一观点，采用训练集的固有特征来验证模型所有权. 然而这种方法很容易做出误判，特别是当可疑模型的训练集与受害者模型的分布相似时. 这种误判可能是因为不同的模型的训练集具有某些特定的特征相似性.

验证嵌入外部特征(verifying embedded external features)技术提出了一种防御模型窃取攻击的新方法，如图 10 所示. 这种方法通过验证可疑模型是否包含防御者指定的外部特征来工作. 具体地，通过对少数训练样本进行风格转换嵌入外部特征，然后训练一个元分类器来判断一个模型是否是从受害者那里窃取的. 由于只对一些样本进行了破坏，并且不更改它们的标签，因此嵌入的功能不会妨碍受害者模型的功能. 这个方法基于理解被窃取的模型应该包含受害者模型学习到的特征知识. 在 CIFAR-10 ImageNet 的一个子集上的实验结果表明，这种方法能有效检测出通过多阶段窃取过程获得的不同类型的模型窃取行为.

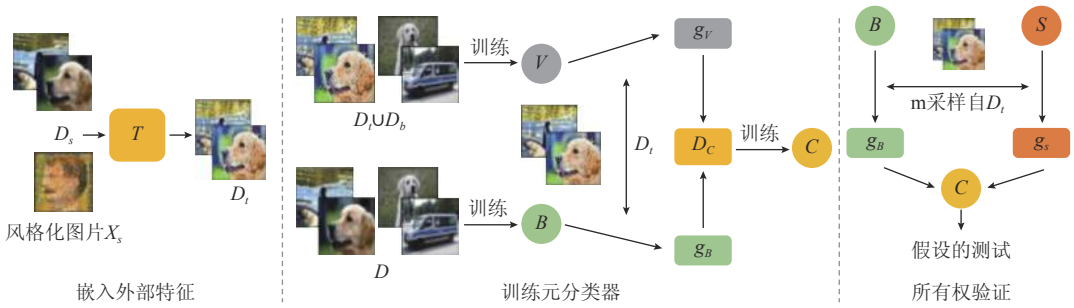


Fig. 10 Verification of embedded external feature technology

图 10 验证嵌入外部特征技术

本节介绍并总结了主流的人工智能模型攻击检测方法，包括防御性蒸馏、正则化技术、数据异常值检测等，不同方法对检测特定攻击类型有其效果和局限性. 表 2 对本节提到的人工智能模型攻击检测方法进行归纳，为构建更安全的攻击检测系统提供了理论支撑和实践引导.

Table 2 Types and Characteristics of Different Attack Detections

表 2 不同攻击检测的类型及特点

攻击检测类型	特点
防御性蒸馏	减少对抗性样本对模型准确度的影响，但可能导致模型准确性的降低
正则化	减轻模型反演攻击，但在一定程度上牺牲了模型的准确度
数据异常值检测	无法检测精心设计以模仿正常数据行为的恶意数据，可能会导致高误报率
鲁棒统计方法	减少异常数据点对模型的影响，可能会导致计算成本增加，或在去除异常值时牺牲有用信息
差分隐私	防止攻击者从统计结果中提取出个人的具体信息，但可能引入了较大噪声损失模型准确性
随机响应机制	适用性广，但高级攻击者可通过统计来减少随机性影响
数据集推理	无需牺牲模型准确性或复杂修改模型架构，但当可疑模型的训练集与受害者模型的分布相似时，容易做出误判
验证嵌入外部特征	嵌入功能不会妨碍受害者模型的功能，允许模型适应不同的数据环境，但也无意中为攻击者提供了额外的信息，扩大了潜在的攻击面.

3 人工智能模型防御策略

在人工智能技术迅猛发展的今天，其安全防御对抵御恶意攻击至关重要. 本节深入探讨了人工智能模型的多层防御策略，覆盖了数据预处理、模型设计、训练方法及知识产权保护等方面. 其中，对抗训练通过引入对手样本增强模型鲁棒性；输入检查与消毒预先筛除恶意输入；模型结构调整以内在强化防御能力；查询控制防御减少信息泄漏；联邦学习提升数据隐私保护；数字水印则守护模型知识产权. 这些策略构建了一个全方位的防御体系，旨在应对和缓解人工智能面临的安全威胁，同时指明了未来研究的新方向.

3.1 对抗训练

人工智能的广泛应用给人类生活带来了诸多便利，但同时也带来了新的挑战. 在人工智能模型中，特别是深度学习模型，攻击者通过设计对抗样本，可以成功地欺骗模型，使其做出错误的预测. 这种现象引起了学术界的广泛关注. 对抗训练作为一种有效的防御策略，已经在多个领域得到了验证. 对抗训练

的主要思想是在训练过程中，通过引入对抗样本，提高模型对于对抗攻击的鲁棒性.

目前，对抗训练是人工智能模型防御策略中的热门研究方向. 文献 [21] 首次提出了对抗训练的概念，并证明了它在提升模型对抗鲁棒性方面的有效性. 如图 11 所示，文献 [49] 通过对比实验发现，对抗训练在提升模型防御性能方面优于其他防御策略，包括输入转换和特征压缩等.

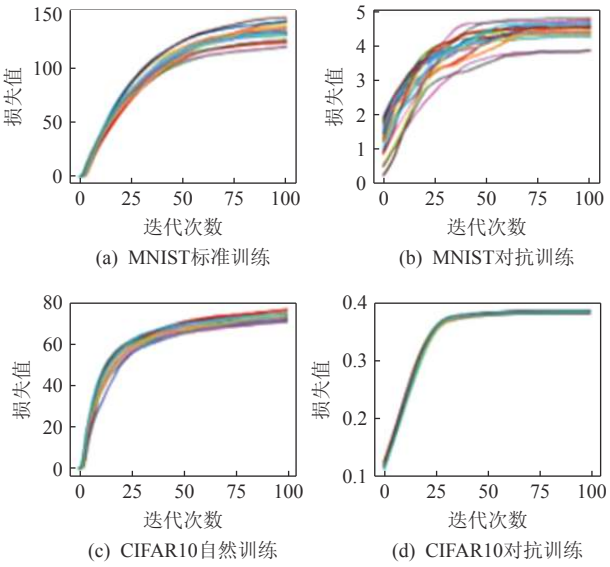


Fig. 11 Performance comparison between adversarial training and other defense strategies

图 11 对抗训练与其他防御策略的性能比较

国内的研究者也在积极探索对抗训练的新方法. 例如，文献 [53] 提出了一种基于对抗样本生成网络 (adversarial sample generative network) 的对抗训练方法，是对抗训练的一种新尝试. 该文通过实验证明了该方法的有效性，但没有对比其他对抗训练方法，因此其优越性尚不明确. 此外，该方法的计算复杂性较高，需在实际应用中进一步优化. 此外，文献 [54] 提出自适应对抗样本生成策略，能够根据训练过程动态地调整对抗样本的强度，这是对对抗训练的一种有益补充. 实验结果表明，该方法能够有效提升模型的对抗鲁棒性，但其在多样性和泛化性方面的效果尚待进一步研究. 综上，对抗训练作为人工智能模型防御策略的一种有效手段，已经在一定程度上提升了模型的防御能力. 然而，目前的对抗训练方法还存在一些问题，如样本生成的复杂性、模型的泛化性等，这都是未来研究的重要方向.

3.2 输入检查/消毒

在人工智能模型的广泛应用中，安全问题已经引起了广泛的关注. 其中，人工智能模型的输入检查和消毒是重要的防御策略. 输入检查是指在数据输入到模型之前，对数据进行预处理，如校验、筛选等，以确保输入数据的合法性和安全性. 输入消毒则是进一步对输入数据进行清洗，删除潜在的恶意代码或不安全元素. 这 2 种策略都可以有效防止对人工智能模型的恶意攻击，保证模型的正常运行和输出结果的有效性.

近年来，人工智能模型的输入检查和消毒已成为研究热点. 例如，文献 [55] 提出了一种基于敌对训练的防御策略，通过在训练过程中添加敌对样本以增强模型的鲁棒性. 该文献分别从 MNIST 数据集中训练集、验证集和测试集数据中选择 10 000 个样本，每个样本产生对应的 9 个目标对抗样本，从而总共产生 270 000 个对抗样本. 实验结果如表 3 所示.

Table 3 Results on 10 000 Samples

表 3 在 10 000 样本上的结果 %

原始样本集 10 000 个	成功误分类的对抗样本	平均失真度	
		全部对抗样本	成功对抗样本
训练集	97.05	4.45	4.03
对抗训练	97.19	4.41	4.01
验证集	97.05	4.45	4.03

然而，这种方法虽然能一定程度上提升模型的抵抗攻击能力，但这种方法存在的问题是，其对模型的防御效果依赖于敌对样本的质量和数量，而生成高质量的敌对样本通常需要大量的计算资源，对于某些复杂的攻击手段，如混淆攻击、欺骗攻击等，依然无法完全防御. 另一方面，许多研究也集中在输入数据的预处理上. 例如，文献 [56] 提出了一种基于异常检测的输入检查方法，通过学习正常输入数据的分布来检测和排除异常输入. 研究从输入数据的角度进行防御，该文献的方法能有效地检测和排除异常输入，从而防止模型被恶意攻击. 然而，该方法的一个主要局限性在于，其对正常输入数据分布的学习可能受到噪声和异常值的影响，从而影响异常检测的效果. 此外，文献 [57] 提出了一种基于深度学习的输入消毒方法，通过训练一个深度神经网络来识别和清除输入中的恶意代码，这是一次重要的突破，输入消毒方法能有效地识别和清除输入中的恶意代码. 然而，这种方法的局限性在于，其对输入数据的处理可能会引入额外的计算开销，从而影响模型的运行效率.

总的来说，当前的研究已经取得了一些进展，但

在输入检查和消毒的防御策略中,仍然存在许多挑战和问题需要解决.未来的研究需要进一步探索更高效的防御方法.

3.3 模型结构防御

随着人工智能技术的迅速发展,深度学习模型在图像识别、自然语言处理、推荐系统等领域取得了显著的成果.然而,这些模型也暴露出了严重的安全问题,例如对抗样本攻击以及模型窃取等威胁.为了提高模型的鲁棒性和安全性,研究者们提出了多种模型防御策略,其中模型结构防御是一种广受关注的防御手段.

模型结构防御主要是通过改变模型的结构或者增加新的结构,以提高模型对抗样本攻击的鲁棒性.最常见的方法包括增加模型的深度和宽度、添加正则项,以及引入新的网络结构等.研究证明,增加模型的深度和宽度可以有效提高模型的鲁棒性.例如,文献 [58] 对于理解对抗样本攻击和防御提供了有价值的见解,证明了增加模型的深度和宽度能够提高模型的鲁棒性,这为后续的研究提供了一个重要的思路.文献 [59] 则从另一个角度出发,认为通过在模型训练中添加正则项,可以显著提高模型对抗样本的鲁棒性.这个方法的优点在于,它不需要改变模型的结构,而且可以和其他的防御方法结合使用.此外,也有研究者通过引入新的网络结构来提高模型的鲁棒性.例如,文献 [58] 提出了一种名为 ResNext 的网络结构,如图 12 所示,通过引入并行的残差连接,可以显著提高模型的鲁棒性,以及通过引入新的网络结构来提高模型的鲁棒性.这种方法的优点在于,它可以直接提高模型的鲁棒性,而不需要额外的训练步骤.

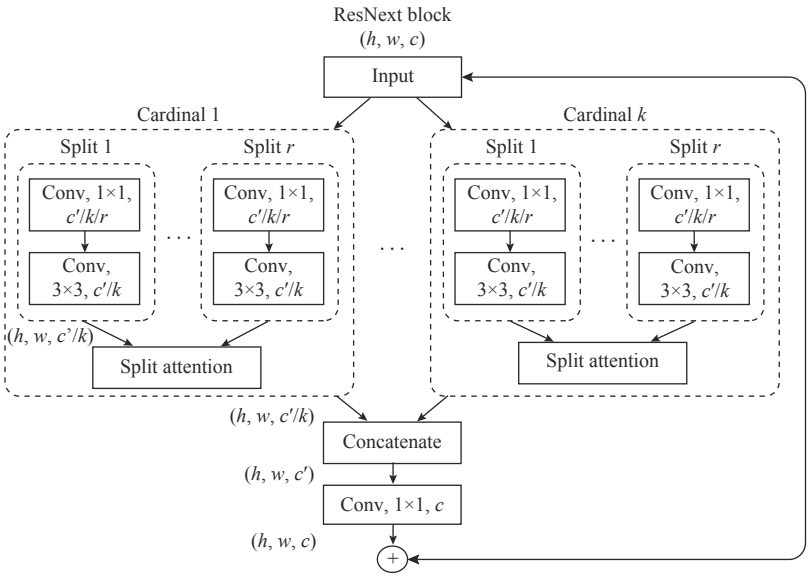


Fig. 12 Network architecture of ResNext

图 12 ResNext 网络结构

总的来说,模型结构防御是一种有效的模型防御策略.未来的研究可以从多个角度出发,例如探索新的网络结构,或者设计更有效的正则项以提高模型的鲁棒性.

3.4 查询控制防御

查询控制防御是一种重要的技术手段,旨在保护人工智能模型免受数据泄露和隐私侵害的风险.在人工智能的训练和测试过程中,模型与隐私泄露风险是一个不可忽视的问题.这些风险包括训练阶段模型参数更新导致的训练数据信息泄露、测试阶段模型返回查询结果造成的模型数据泄露,以及这些人工智能模型正常使用过程中间接引起的数据隐私泄露.为了减轻这些风险,学术界和工业界采取了多种防御措施,包括模型结构防御、信息混淆防御,以及查询控制防御.

查询控制防御主要通过分析用户的查询行为,分辨出哪些用户是攻击者,进而及时拒绝恶意的查询以防止数据泄露,从而达到防御攻击的目的.这包括异常样本检测和查询行为检测 2 个方面[60].

异常样本检测作为查询控制防御的一个重要组成部分,专注于识别和防御恶意查询,以防止攻击者通过模型查询窃取敏感信息.其基于一个核心前提:

攻击者为了从模型中窃取信息，往往会提交非正常或被有目的地修改的查询样本. 通过分析这些样本的特征，可以识别出攻击者的窃取行为. 这类检测方法利用了样本特征之间的统计分布差异、样本特征间距离的分布，来区分正常样本与异常样本. 例如，基于距离分布的检测，文献 [61] 技术通过分析样本特征间的距离分布来判断用户是否正在进行模型窃取攻击. 这种方法利用了正常样本特征间距离的正态分布特性，与人工修改样本的距离分布进行对比，从而检测异常查询行为. 基于特征分布检测，直接分析查询样本在特征空间中的分布，通过记录用户的查询样本并检查其在特征空间中的分布，可以评估模型被盗取的风险. 文献 [62] 的研究记录了用户查询样本的特征分布，并利用这些信息来检测潜在的窃取行为.

查询行为检测关注于用户的查询频率和模式，攻击者通常需要对目标模型进行大量查询以提高攻击效率. 通过限制查询频率或识别非正常的查询模式，可以在一定程度上防御模型泄露和数据泄露攻击. 其基于一个简单而强大的假设：攻击者为了提取模型信息或进行其他形式的恶意活动，往往会展示出与正常用户明显不同的查询模式. 这些模式可能表现为异常高的查询频率、特定类型的查询样本分布，或者是一系列经过精心设计以探索模型弱点的查询. 通过分析这些查询行为的特征，系统可以及时识别并阻断这些潜在的攻击行为. 最直接的方法之一是分析用户查询的频率. 攻击者在执行模型探测或信息提取攻击时，往往需要频繁地向模型提交查询. 通过设置阈值来限制在特定时间窗口内允许的最大查询次数，可以有效阻止这类攻击. 其次利用机器学习技术，如聚类分析或异常检测算法，来识别用户查询行为中的异常模式. 这些技术可以帮助系统学习并识别正常与异常查询行为之间的差异，从而提高检测的准确性. He 及其团队[63] 提出了一种基于用户查询行为的防御策略，用于抵御成员推断攻击. 这种攻击通常需要攻击者对目标模型进行大量查询，以便推断出模型的训练数据. 通过在模型接收查询的初期阶段分析用户的查询行为特征，尤其是查询频率，模型提供者能够对频繁的查询进行限制. 这种方法通过增加攻击者执行成员推断攻击所需的成本，有效地提高了模型的安全性.

3.5 联邦学习

联邦学习作为一种新兴技术，已经在解决“数据孤岛问题”和加强隐私保护方面显示出其独特优势. 它允许多个参与方在不共享原始数据的情况下，共同训练一个全局的机器学习模型. 这种去中心化的训练方法不仅保证了数据的本地性，降低了数据泄露的风险，而且在一定程度上提高了模型训练的效率和效果.

传统的集中式机器学习模型训练方法要求将所有参与方的数据集中存储于中心服务器或云端，用于后续的训练和测试，以便开发出更全面的通用机器学习模型[64]. 但随着数据体量的不断增大，这不仅增加了数据泄露的风险，也对中心服务器的计算能力提出了更高的要求. 此外，这种方法在处理大量分散数据时面临计算效率低下和模型精度不足的问题. 相比之下，Google 公司在 2017 年提出了安全解决方案：联邦学习[65]，多个参与方的数据无需上传到中央服务器上进行联合训练，而是通过在本地训练模型并仅分享模型更新，进而上传到服务器中，由服务器执行聚合，迭代收敛后得到的全局会进一步分发给所有参与方. 联邦学习这一种去中心化的思路，有效地解决了这些问题，减少了对中心计算资源的依赖，同时也降低了数据传输的成本和隐私泄露的风险.

联邦学习的实现方法可以精确地定义[66] 为：设有 n 个参与方，每个参与方分别用 F_i 表示，拥有的数据集为 D_i，其中 $i=\{1, 2, \cdots, n\}$. 传统的模型训练方法是将所有参与方的数据集 $D = D_1 \cup D_2 \cup \cdots \cup D_{n-1} \cup D_n$ 集中到一个中心节点进行处理，通过这种方式训练得到的模型记为 M_{sum}，其准确率为 V_{sum}. 而在联邦学习框架下，每个参与方 $F_1, F_2, \cdots, F_n$ 利用各自的数据集 $D_1, D_2, \cdots, D_n$ 在本地独立进行模型训练. 这些局部训练的模型随后被聚合为一个全局模型 M_{fed}，其准确率为 V_{fed}. 在此过程中，任何参与方 F_i 都无法访问到自己之外的其他数据集 D_i. 如果存在一个非负实数 δ，使得 $|V_{\text{sum}} - V_{\text{fed}}| < \delta$，则可以认为全局模型 M_{fed} 的性能损失在 δ 以内，其中 δ 是一个预先定义的足够小的浮点数. 这意味着通过联邦学习方法训练得到的模型与传统方法集中处理所有数据训练得到的模型在准确率上的差异非常小. 此定义不仅强调了联邦学习在保护数据隐私的同时维持模型性能的能力，而且量化了联邦学习模型与传统模型之间性能差异的上限，为评估联邦学习模型的有效性提供了一个明确的标准. 联邦学习的每一次执行过程也被视为一次次迭代的过程[67]，每一次迭代计算都在改进全局模型，其步骤可以概括为：

1）初次模型分发. 中央服务器将初始模型发送给所有参与方. 这一步骤确保了每个参与方都从同一

个起点开始模型训练.

2)本地训练. 每个参与方利用自己的本地数据对接收到的全局模型进行训练. 这一步骤允许各参与方独立地在自己的数据上优化模型, 而无需直接共享这些数据.

3)本地模型上传. 训练完成后, 各参与方将自己的模型更新(通常是模型的参数或梯度)上传到中央服务器. 这一步骤为全局模型的更新聚合提供了所需的输入.

4)模型聚合. 中央服务器收到所有参与方的模型更新后, 将这些更新聚合以更新全局模型. 这通常通过执行联邦平均算法来完成, 其目的是构建一个更好的全局模型, 反映了所有参与方数据的共同特征.

5)全局模型分发. 一旦全局模型被更新, 中央服务器就会将其分发给所有参与方, 为下一轮的本地训练做准备. 这个过程会一直重复, 直到全局模型达到预定的收敛标准.

联邦学习作为一种创新的分布式机器学习框架, 已经在各个领域证明了其卓越的性能. 这种方法不仅保护了数据隐私, 还允许从分布式数据中学习到有价值的信息. 根据数据的特征和样本分布, 联邦学习可以被细分为3个主要模式: 横向联邦学习、纵向联邦学习和联邦迁移学习, 如图13所示, 每种模式

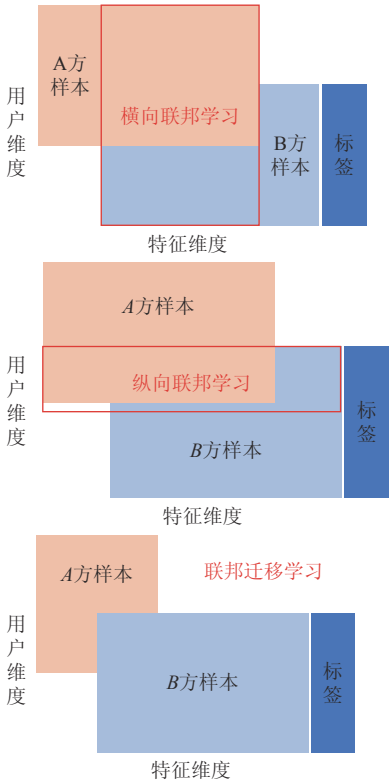


Fig. 13 Classification of federated learning

图13 联邦学习的分类

都针对不同的数据分布和应用场景设计.

横向联邦学习适用于那些拥有相似特征但不同样本的场景. 这种模式最早由Google公司提出, 旨在使多个参与方能够合作训练一个共享模型, 而无需直接交换他们的数据. 横向联邦学习特别适用于数据分布广泛但保持相似特性的情况, 如跨多个组织的用户数据. 安全风险主要存在于中央服务器与可能存在的恶意参与方之间, 这要求采取额外的安全措施来保护数据和模型的完整性.

纵向联邦学习则针对那些样本相似但特征不同的场景. 例如, 2个不同行业的组织可能拥有相同用户的不同类型数据. 通过纵向联邦学习, 这些组织可以合作创建一个更加全面的模型, 而无需直接共享敏感信息. 这种模式的安全挑战包括防止恶意参与方破坏本地数据或串谋窃取信息.

联邦迁移学习旨在将知识从一个领域迁移到另一个领域, 适用于数据特征和样本都有显著差异的场景. 这种方式使得在一个领域内训练的模型可以适应于完全不同的应用场景, 极大地增强了模型的泛化能力. 文献[68]通过联邦迁移学习来训练个性化的分布式模型用以实现安全的图像分析. 如同纵向联邦学习, 联邦迁移学习同样面临计算环境中可能的恶意行为风险. 这些模式扩展了联邦学习的应用范围, 使其能够更灵活地适应不同的数据分布和应用需求.

在联邦学习中, 数据保留在本地, 只有模型参数或梯度更新在服务器和参与方之间进行传递. 为了保护这些参数的隐私, 在传递过程中通常采用隐私保护技术进行脱敏处理. 近年来, 用于联邦学习中的隐私保护技术可以基于数据加密、数据扰动和可信硬件分为3类, 常用的隐私保护技术分类如图14所示.

基于数据加密的隐私保护技术主要包括安全多方计算(secure multi-party computation)和同态加密. 安全多方计算允许多个不互信的参与方在不泄露各自输入的情况下共同完成计算任务, 有效解决了在缺少可信第三方介入的情况下的协同计算问题. 同态加密则允许直接在加密数据上进行计算, 无需解密, 分为部分同态加密、次同态加密和全同态加密, 分别支持无限次数的单一类型运算、有限次数的混合运算和无限次数的混合运算, 使得加密数据处理变得更加灵活高效.

文献[69]提出一种增强的安全多方计算方法, 消除服务器在一般联邦学习设置中必须是诚实的约束; 文献[70]提出一种保护隐私的联邦学习框架, 解

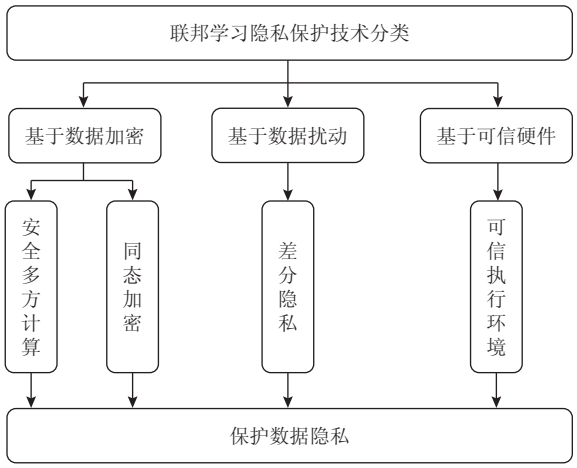


Fig. 14 Classification of privacy protection technologies

图 14 隐私保护技术分类

决写作式模型训练时访问模型参数或梯度暴露私有样本的问题；文献 [71] 提出一个基于安全多方计算的特征工程框架，解决明文和密文数据共享中的安全和效率问题. 文献 [72] 提出了一种可以实现同态操作的算法，可以防止本地模型传输到服务器的过程中隐私数据遭受侵犯. 文献 [73] 提出一种源匿名数据洗牌方案 ReShuffle. 其采用不经意传输协议实现参与者模型的匿名上传.

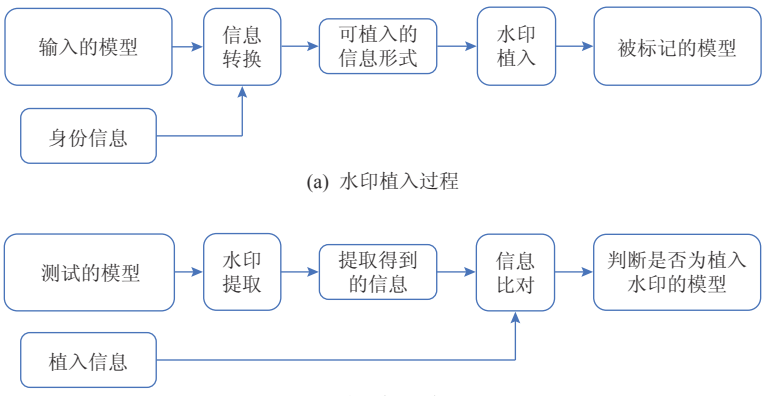
差分隐私技术通过在敏感数据上添加噪声来实现隐私保护，旨在确保单一数据项的变化不会对输出结果产生显著影响，从而保护个人信息不被推断. 差分隐私根据实施方式的不同分为中心化差分隐私、本地化差分隐私、分布式差分隐私和混合差分隐私. 这些方法从集中式噪声添加到在本地进行数据扰动，再到根据信任级别选择适当的差分隐私策略，为不同的应用场景提供了灵活的隐私保护方案.

文献 [74] 提出一个保护跨筒仓联邦学习的三平面框架，解决跨筒仓联邦学习中用户级隐私存在泄露风险的问题；文献 [75] 提出一种分布式差分隐私机制，防止参与客户端串通暴露诚实客户模型参数的攻击. 可信执行环境（trusted executed environment, TEE）提供了一种安全的计算环境，它与不可信的操作系统相互独立，存在于计算平台上，通过硬件机制确保其安全性. TEE 为敏感计算任务创建了一个隔离的、可信的执行空间，使得即使在不可信的操作系统环境中也能保障数据的安全性和完整性. 这种基于可信硬件的技术为联邦学习中的数据安全和隐私保护提供了强大的物理层面保障，是解决高安全要求场景下隐私保护问题的有效手段. 文献 [76] 提出一种基于 TEE 的分布式联邦学习框架，其攻击者从中间梯度中逆推私有数据；文献 [77] 提出一个隐私保护联邦学习框架，可以有效防止联邦学习中的隐私泄露；文献 [78] 提出一个新型的联邦学习隐私信息保护方法，防止在训练过程中恶意参与方在模型中加入篡改过的数据来破坏模型.

3.6 数字水印

在过去几十年中，水印技术已被深入研究和广泛应用，证明其是解决数字媒体版权和安全问题的有效机制. 特别是数字水印技术，它通过将具有特定意义的数字信号（如图像、文本等）隐秘地嵌入到载体图像中，能够在不损害载体图像的实用价值的前提下，有效实施版权保护、所有权认证、内容完整性验证、篡改检测及定位，甚至对被篡改区域进行恢复. 人工智能模型水印的生成方法有很多种，但归纳起来有一个总体框架. 人工智能模型水印的生成、植入与提取的过程如图 15 所示.

水印技术按照其对抗修改的能力可分为鲁棒水印、脆弱水印和半脆弱水印[79]. 鲁棒水印特点是其高度的耐久性，即使在载体图像遭受修改、压缩或转换



(a) 水印植入过程

(b) 水印提取过程

Fig. 15 Schematic diagram of watermarking process of model

图 15 模型水印流程示意图

等攻击时，仍能保持水印信息的完整性. 这使得鲁棒水印成为版权保护和所有权认证的理想选择. 相反，脆弱水印非常敏感，以至于即使是微小的像素级变化也能导致水印信息的破坏. 这种特性使得脆弱水印非常适合用于篡改检测，因为任何对图像的未授权改动都会立即被发现. 半脆弱水印则位于鲁棒水印和脆弱水印之间，具有一定的鲁棒性，能够承受如JPEG压缩和基本图像编辑等非恶意的内容保留型修改，同时对于恶意的篡改行为保持高度敏感. 这种平衡的特性使半脆弱水印适用于那些需要辨别图像是否经过授权编辑的场景.

随着人工智能模型在各种领域的迅速发展，尽管当前许多图像水印算法表现出优良的性能，但普遍面临着对抗几何攻击能力不足以及水印的鲁棒性与隐蔽性难以兼顾的挑战. 为了提升算法性能，研究人员在深度学习技术的基础上，进一步探索优化传统水印技术的新途径.

提升水印的抗几何攻击能力成为研究的重点. 虽然大多数算法能有效抵御如噪声、滤波和压缩等非几何攻击，但面对旋转、缩放和剪切等几何攻击时则表现不佳. 一些研究者通过利用图像的几何不变性，如SVD、Schur分解、Zernike矩、伪Zernike矩、复矩以及极谐变换(PHTs)等，增强水印的鲁棒性. 特别是，PHTs因其几何不变性而被广泛应用于图像水印领域. 例如，通过将水印嵌入PHTs系数中[80]，一些研究实现了对旋转和缩放攻击的强鲁棒性和较大的嵌入容量.

部分学者通过图像特征点的几何不变性被用来选择嵌入水印的位置. 采用多尺度理论和非下采样轮廓波变换(non subsampled contourlet transform)[81]，尺度不变特征变换(scale-invariant feature transform)[82]和加速稳健特征(speeded up robust features)[83]等特征检测技术，可以在保持水印不可见性的同时，提高其抗攻击能力. 还有部分学者通过几何(同步)校正方法比较原始和接收图像的特征变化，计算图像的几何变换参数，实现图像的还原和水印提取. 这类方法通过SIFT和SURF特征匹配点坐标变化[84]，有效提升了水印的抗几何攻击能力.

本节介绍并总结了针对人工智能模型的多层次防御策略，涵盖了对抗训练、输入检查与消毒、模型结构防御、查询控制防御、联邦学习以及数字水印等多种方法. 通过综合运用这些策略，构建一个全面的安全防御体系，从而显著提升人工智能模型的安全性，确保其在各种应用场景中的可靠性和稳定性. 表4对上述提到的人工智能模型的多层防御策略进行归纳.

Table 4 Types and Characteristics of Multi-Layer Defense Strategies in Artificial Intelligence Models

表4 人工智能模型中多层防御策略的类型及特点

防御策略类型	特点
对抗训练	通过将对抗样本添加到训练数据中，增强模型对设计欺骗模型的攻击的抵抗力.
输入检查/消毒	验证并清洁输入数据，防止恶意数据对模型产生不利影响.
模型结构防御	调整模型架构，增强其对各种攻击向量的抵抗力.
查询控制防御	通过限制查询次数或修改查询响应策略来保护模型不被利用.
联邦学习	通过在多个分散的设备上分布式训练模型，减少数据集中的风险并增强隐私保护.
数字水印	在模型中嵌入水印，用于追踪和认证合法使用和所有权.

4 主要挑战与未来展望

随着科技的不断进步，人工智能的适用范围也越来越广泛，随之而来的安全威胁也在不断增加. 人工智能的发展给我们带来了巨大的便利和效率提升，同时也给我们带来了新的挑战和风险. 人工智能安全也进入到了一个持续发展的领域.

首先，随着人工智能技术的快速发展，人们越来越依赖人工智能模型来处理和分析大量的数据. 然而，这种依赖也意味着一旦其遭到攻击或出现故障，将会对我们的生活和工作造成严重的影响. 在不同的业务应用领域，例如医疗领域，人工智能模型在辅助诊断和治疗方面具有潜力，但如果这些模型被黑客攻击或受到错误指导，可能会导致严重的医疗事故. 因此，确保人工智能模型的安全性和可靠性成为了当务之急.

其次，随着人工智能模型的智能化程度的提高，它们也变得越来越复杂和难以理解. 这给我们的安全工作带来了新的挑战. 传统的安全防护手段可能无法有效地应对人工智能模型中的漏洞和风险. 因此，我们需要不断研究和发展新的安全技术，以确保模型的安全性.

再次，人工智能在社交媒体和网络中的广泛应用也给我们的隐私带来了潜在的风险. 人工智能模型可以收集和分析大量的个人数据，从而了解我们的喜好、行为和习惯. 然而，如果这些数据被滥用或泄露，将会对我们的隐私和个人安全造成严重威胁. 因此，我们需要制定更加严格的隐私保护法规，并加

强对人工智能模型的监管和审查.

最后，人工智能技术的发展也引发了一些伦理和道德问题. 例如，自动驾驶汽车在道路上行驶时，如何在遇到紧急情况时做出最合适的决策，这涉及到人的生命和财产安全. 我们需要思考如何在人工智能模型中加入道德和伦理准则，以确保其行为符合人类价值观.

未来，随着技术的不断进步和市场需求的不断增长，人工智能安全将越来越受到关注. 我们应该从3个方面着手推动人工智能安全的发展：

1)强化数据隐私和安全保护. 随着数据泄露和滥用的风险不断增加，加强数据隐私和安全保护将成为人工智能安全的重要任务. 加密技术、安全存储和传输等手段可以用来保护数据的安全.

2)抵御对抗性攻击. 为了提高人工智能模型的鲁棒性和抵御对抗性攻击，需要研究和开发更加健壮的人工智能算法和模型. 此外，建立对抗性攻击检测和防御机制也是必要的.

3)提高透明度和解释性. 为了增强用户和监管机构的信任，需要进一步研究如何提高人工智能模型的透明度和解释性. 开发可以解释和理解人工智能模型决策的方法和工具，将有助于减少不可预测性和提高系统的可信度.

5 总　　结

本文总结和分析了近年来人工智能模型的安全攻击和防御策略研究工作. 尽管许多研究者已经对人工智能系统中基于模型输出和梯度更新的数据泄露进行了一系列研究，并提出了多种防御技术. 然而，与已经成熟的传统数据安全领域相比，由于深度学习算法本身存在可解释性不足的问题，要妥善解决人工智能算法数据安全和隐私保护问题，仍然面临许多挑战，需要进一步的探索和研究. 人工智能安全是一个不断发展的技术领域，我们需要不断研究和创新，以应对新的安全威胁和挑战. 只有确保人工智能模型的安全性和可靠性，才能更好地发挥其潜力，为人类社会带来更多的福祉.

作者贡献声明：秦臻提出了论文的算法和框架，并修改论文；庄添铭负责论文文献的搜集和整理工作，并撰写论文；朱国淞负责文献的分析，并撰写了论文相关章节；周尔强提出指导意见，并参与论文的修改；丁熠负责论文的编辑和校对；耿技审阅论文.

参　考　文　献

[1] Chen Zhuo, Yin Fei, Yang Qing, et al. Cross-lingual text image recognition via multi-hierarchy cross-modal mimic[J]. IEEE Transactions on Multimedia, 2023, 25: 4830−4841

[2] Yan Tiantian, Li Haojie, Sun Baoli, et al. Discriminative feature mining and enhancement network for low-resolution fine-grained image recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5319−5330

[3] Chen Jicheng, Zhang Yongkang, Teng Siyu, et al. ACP-based energy-efficient schemes for sustainable intelligent transportation systems[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(5): 3224−3227

[4] Ling Yating, Wang Yuling, Dai Wenli, et al. MTANet: Multi-task attention network for automatic medical image segmentation and classification[J]. IEEE Transactions on Medical Imaging, 2024, 43(2): 674−685

[5] Liu Hongmin, Jin Fan, Zeng Hui, et al. Image enhancement guided object detection in visually degraded scenes[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023[2024-05-31]. https://ieeexplore.ieee.org/document/10130799

[6] Zhou Chulun, Liang Yunlong, Meng Fandong, et al. A multi-task multi-stage transitional training framework for neural chat translation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 7970−7985

[7] Liu Jia, Li Tianrui, Ji Shengong, et al. Urban flow pattern mining based on multi-source heterogeneous data fusion and knowledge graph embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35: 2133−2146

[8] Ateniese G, Mancini L V, Spognardi A, et al. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers[J]. International Journal of Security and Networks, 2015, 10(3): 137−150

[9] Mehnaz S, Dibbo S V, De V R, et al. Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models[C]//Proc of the 31st USENIX Security Symp (USENIX Security 22). Berkeley, CA: USENIX Association, 2022: 4579−4596

[10] Juuti M, Szyller S, Marchal S, et al. PRADA: Protecting against DNN model stealing attacks[C]//Proc of IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2019: 512–527

[11] Battista B, Blaine N, Pavel L. Support vector machines under adversarial label noise[C]//Proc of Asion Conf on Machine Learning. New York: PMLR, 2011: 97−112

[12] Ru B, Cobb A, Blaas A, et al. Bayesopt adversarial attack[C]//Proc of Int Conf on Learning Representations. 2019. https://openreview.net/pdf?id=Hkem-lrtvH

[13] Wang Binghui, Gong Neli, Zhenqiang. Stealing hyperparameters in machine learning[C]//Proc of 2018 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2018: 36−52

[14] Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and

Technology, 2019, 10(2): 1−19

[15] Papernot N, McDaniel P, Wu Xi, et al. Distillation as a defense to adversarial perturbations against deep neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(10): 1631−1643

[16] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy[C]//Proc of the 2016 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 308−318

[17] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey[J]. ACM Computing Surveys, 2009, 41(3): 1−58

[18] Seidman J H, Fazlyzb M, Preciado V M. Robust deep learning as optimal control: Insights and convergence guarantees[C]//Proc of the 2nd Conf on Learning for Dynamics and Control. New York: PMLR, 2020: 884−893

[19] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias[J]. Journal of the American Statistical Association, 1965, 60(309): 63−69

[20] Zhou Wen, Hou Xin, Chen Yongjun, et al. Transferable adversarial perturbations[C]//Proc of the European Conf on Computer Vision (ECCV). Piscataway, NJ: IEEE, 2018, 452−467

[21] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2014

[22] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]//Proc of IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 39−57

[23] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1322−1333

[24] Yang Z, Zhang J, Chang E C, et al. Neural network inversion in adversarial setting via background knowledge alignment[C]//Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 225−240

[25] Zhu Yuankun, Cheng Yueqiang, Zhou Husheng, et al. Hermes attack: Steal DNN models with lossless inference accuracy[C]//Proc of the 30th USENIX Security Symp (USENIX Security 21). Berkeley, CA: USENIX Association, 2021: 1973−1988

[26] Reith R N, Schneider T, Tkachenko O. Efficiently stealing your machine learning models[C]//Proc of the 18th ACM Workshop on Privacy in the Electronic Society. New York: ACM, 2019: 198−210

[27] Chandrasekaran V, Chaudhuri K, Giacomelli I, et al. Exploring connections between active learning and model extraction[C]//Proc of the 29th USENIX Security Symp (USENIX Security 20). Berkeley, CA: USENIX Association, 2020: 1309−1326

[28] Orekondy T, Schiele B, Fritz M. Knockoff nets: Stealing functionality of black-box models[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4954−4963

[29] Battista B, Blaine N, Pavel L. Poisoning attacks against support vector machines[C]//Proc of the Int Conf on Machine Learning, Edinburgh, Scotland, New York: ICML 2012: 1467−1474

[30] Cinà A E, Vascon S, Demontis A, et al. The hammer and the nut: Is bilevel optimization really needed to poison linear classifiers?[C]// Proc of Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2021: 1−8

[31] Carnerero C J, Muñoz-González L, Spencer P, et al. Hyperparameter learning under data poisoning: Analysis of the influence of regularization via multiobjective bilevel optimization[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023[2024-5-31]. https://ieeexplore.ieee.org/document/10235261

[32] Mei Shike, Zhu Xiaojin. Using machine teaching to identify optimal training-set attacks on machine learners[C]//Proc of AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015, 29(1): 2871−2877

[33] Feng Ji, Cai Qizhi, Zhou Zhihua. Learning to confuse: Generating training time adversarial data with auto-encoder[J]. Advances in Neural Information Processing Systems, 2019[2024-05-31]. https://proceedings.neurips.cc/paper_files/paper/2019/file/1ce83e5d4135b07c0b82afffbe2b3436-Paper.pdf

[34] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2017: 1885−1894

[35] Jagielski M, Severi G, Pousette H N, et al. Subpopulation data poisoning attacks[C]//Proc of the 2021 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2021: 3104−3122

[36] Gu Tianyu, Liu Kang, Dolan-Gavitt B, et al. Badnets: Evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230−47244

[37] Chu Cheng, Chen Fan, Richerme P, et al. QDoor: Exploiting approximate synthesis for backdoor attacks in quantum neural networks[C]//Proc of IEEE Int Conf on Quantum Computing and Engineering. Piscataway, NJ: IEEE, 2023, 1: 1098−1106

[38] Zhang Zhengming, Tian Muchen, Li Chunguo, et al. Poison neural network-based mmwave beam selection and detoxification with machine unlearning[J]. IEEE Transactions on Communications, 2022, 71(2): 877−892

[39] Zhong Haoti, Liao Cong, Squicciarini A C, et al. Backdoor embedding in convolutional neural network models via invisible perturbation[C]// Proc of the 10th ACM Conf on Data and Application Security and Privacy. New York: ACM, 2020: 97−108

[40] Salem A, Yang Zhang, Humbert M, et al. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models[J]. arXiv preprint, arXiv: 1806.01246, 2018

[41] Sablayrolles A, Douze M, Schmid C, et al. White-box vs black-box: Bayes optimal strategies for membership inference[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2019: 5558−5567

[42] Hisamoto S, Post M, Duh K. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system?[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 49−63

[43] Jamie H, Luca M, George D, et al. LOGAN: Membership inference attacks against generative models[J].arXiv preprint, arXiv: 1705.07663.2017

[44] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]//Proc of 2017 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 39−57

[45] Yu Lei, Liu Ling, Calton P, et al. Differentially private model publishing for deep learning[C]//Proc of IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 332−349

[46] Steinhardt J, Koh P W W, Liang P S. Certified defenses for data poisoning attacks[J]. Advances in neural information processing systems,2017[2024-05-31].https://proceedings.neurips.cc/paper_files/paper/2017/file/9d7311ba459f9e45ed746755a32dcd11-Paper.pdf

[47] Pang Guansong, Shen Chunhua, Cao Longbin, et al. Deep learning for anomaly detection: A review[J]. ACM Computing Surveys, 2021: 1−38

[48] Diakonikolas I, Kamath G, Kane D, et al. Robust estimators in high-dimensions without the computational intractability[J]. SIAM Journal on Computing, 2019, 48(2): 742−864

[49] Chen Bo, Hawkins C, Yazdani K, et al. Edge differential privacy for algebraic connectivity of graphs[C]//Proc of the 60th IEEE Conf on Decision and Control (CDC). Piscataway, NJ: IEEE, 2021: 2764−2769

[50] Jian Xun, Wang Yue, Chen Lei, et al. Publishing graphs under node differential privacy[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(4): 4164−4177

[51] Yan Jun, Liu Hai, Wu Zhenqiang. An efficient differential privacy method with wavelet transform for edge weights of social network[J]. International Journal of Network Security, 2022, 24(2): 181−192

[52] Wang Hengzhi, Yang Yongjian, Wang En, et al. Bilateral privacy-preserving worker selection in spatial crowdsourcing[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 20(3): 2533−2546

[53] Yan Ziang, Guo Yiwen, Zhang Changshui. Deep defense: Training dnns with improved adversarial robustness[J]. Advances in Neural Information Processing Systems. San Diego: NIPS, 2018, 31.

[54] Yuan Xiaoyong, He Pan, Zhu Qile, et al. Adversarial examples: Attacks and defenses for deep learning[J]. IEEE transactions on neural networks and learning systems, 2019, 30(9): 2805−2824

[55] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings[C]//Proc of 2016 IEEE European Symp on Security and Privacy (EuroS&P). Piscataway, NJ: IEEE, 2016: 372−387

[56] Ma Xingjun, Li Bo, Wang Yisen, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[J]. arXiv preprint, arXiv: 1801.02613, 2018

[57] Song Yang, Kim T, Nowozin S, et al. PixelDefend: Leveraging generative models to understand and defend against adversarial examples[J]. arXiv preprint, arXiv: 1710.10766, 2017

[58] Goodfellow, Ian J, Shlens J, et al. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2014

[59] Wang Jingyuan, Wu Yufan, Li Mingxuan, et al. Interpretability is a kind of safety: An interpreter-based ensemble for adversary defense[C]// Proc of the 26th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2020: 15−24

[60] Xie Saining, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE 2017: 1492−1500

[61] Ren Kui, Meng Quanrun, Yan Shoukun, et al. A survey on attacks and defenses of data leakage in artificial intelligence models[J]. Journal of Cyber Security, 2021, 7(1): 1−10 (in Chinese)
(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)

[62] Kesarwani M, Mukhoty B, Arya V, et al. Model extraction warning in MLaaS paradigm[C]//Proc of the 34th Annual Computer Security Applications Conf. New York: ACM 2018: 371−380

[63] He Yingzhe, Meng Guozhu, Chen Kai, et al. Towards privacy and security of deep learning systems: a survey[J]. arXiv preprint, arXiv: 1911.12562, 2019

[64] Roh Y, Heo G, Whang S E. A survey on data collection for machine learning: A big data - AI integration perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(4): 1328−1347

[65] McMahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proc of the 20th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2017: 1273−1282

[66] Xiong Shiqiang, He Daomin, Wang Zhendong, et al. A survey on federated learning and its security and privacy protection[J]. Computer Engineering, 2024, 50(5): 1−17 (in Chinese)
(熊世强, 何道敏, 王振东, 等. 联邦学习及其安全与隐私保护研究综述[J]. 计算机工程, 2024, 50(5): 1−17)

[67] Xiao Xiong, Tang Zhuo, Xiao Bin, et al. A survey on privacy protection and security defenses in federated learning[J]. Journal of Computer Research and Development, 2023, 46(5): 1019−1044 (in Chinese)
(肖雄, 唐卓, 肖斌, 等. 联邦学习的隐私保护与安全防御研究综述[J]. 计算机学报, 2023, 46(5): 1019−1044)

[68] Yang Hongwei, He Hui, Zhang Weizhe, et al. FedSteg: A federated transfer learning framework for secure image steganalysis[J]. IEEE Transactions on Network Science and Engineering, 2021, 8(2): 1084−1094

[69] Zhang C, Ekanut C, Zhen L L, et al. Augmented multi-party computation against gradient leakage in federated learning[J]. IEEE Transactions on Big Data, 2022: 1−10

[70] Hosseini S M, Sikaroudi M, Babaie M, et al. Cluster based secure multi-party computation in federated learning for histopathology images[C]//Medical Image Computing and Computer Assisted Intervention Society. Berlin: Springer, 2022: 110−118

[71] Sun Litong, Du Runmeng, He Daojing, et al. Feature engineering framework based on secure multi-party computation in federated learning[C]//Proc of 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). Piscataway, NJ: IEEE, 2021, 487−494

[72] Park J, Lim H, et al. Privacy-preserving federated learning using homomorphic encryption with different encryption keys[C]// Proc of 2022 13th Int Conf on Information and Communication Technology Convergence (ICTC). Piscataway, NJ: IEEE 2022: 1869−1871

[73] Chen Jingxue, Gao Kehan, Zhou Erqiang, et al. Robust source anonymous federated learning shuffling protocol in the Internet of

things environment[J]. Journal of Computer Research and Development, 2023, 60(10): 2218−2233 (in Chinese)
(陈景雪, 高克寒, 周尔强, 等. 物联网环境下鲁棒的源匿名联邦学习洗牌协议[J]. 计算机研究与发展, 2023, 60(10): 2218−2233)
[74] Wang Chen, Wu Xinkui, Liu Gaoyang, et al. Safeguarding cross-silo federated learning with local differential privacy[J]. Digital Communications and Networks, 2022, 8(4): 446−454
[75] Byrd D, Mugunthan V, Polychroniadou A, et al. Collusion resistant federated learning with oblivious distributed differential privacy[C]// Proc of the 3rd ACM Int Conf on AI in Finance. New York: ACM, 2022: 114−122
[76] Xu Tianxing, Zhu Kongling, Rzejak A, et al. Distributed learning in trusted execution environment: A case study of federated learning in SGX[C]//Proc of the 7th IEEE Int Conf on Network Intelligence and Digital Content (IC-NIDC). Piscataway, NJ: IEEE, 2021: 450−454
[77] Mo Fan, Haddadi H, Katevas K, et al. PPFL: Privacy-preserving federated learning with trusted execution environments[C]//Proc of the 19th Annual Int Conf on Mobile Systems, Applications, and Services. New York: ACM, 2021: 94−108
[78] Chen Yu, Luo Fang, Li Tang, et al. A training-integrity privacy-preserving federated learning scheme with trusted execution environment[J]. Information Sciences, 2020, 522: 69−79
[79] Zhao Lei, Gui Xiaolin, Shao Yiyang, et al. A survey on multifunctional watermarking in digital images[J]. Journal of Computer-Aided Design and Computer Graphics, 2022, 36(2): 1−28 (in Chinese)
(赵蕾, 桂小林, 邵屹样, 等. 数字图像多功能水印综述[J]. 计算机辅助设计与图形学学报, 2022, 36(2): 1−28)
[80] Li Leuda, Li Shushang, Abraham A, et al. Geometrically invariant image watermarking using polar harmonic transforms[J]. Information Sciences, 2012, 199: 1−19
[81] Zhu Dandan, Zhang Xiuping, Zhang Youliang, et al. A new image watermarking algorithm using NSCT and Harris detector in green manufacturing[J]. Applied Mechanics and Materials, 2013, 340: 277−282
[82] Liu Ying, Yang Xing, Zhu Tingge. Robust watermarking algorithm based on structured forest edges and SIFT[J]. Advances in Laser and Optoelectronics, 2021, 58(6): 339−346 (in Chinese)
(刘颖, 杨星, 朱婷鸽. 基于结构森林边缘和 SIFT 的鲁棒水印算法[J]. 激光与光电子学进展, 2021, 58(6): 339−346)
[83] Hou Xiang, Min Lianquan. Robust watermarking algorithm based on SURF feature regions[J]. Journal of Wuhan University (Information Science Edition), 2017, 42(3): 421−426 (in Chinese)
(侯翔, 闵连权. 基于 SURF 特征区域的鲁棒水印算法[J]. 武汉大学学报: 信息科学版, 2017, 42(3): 421−426)
[84] Pourhadi A, Mahdavi-Nasab H. A robust digital image watermarking scheme based on BAT algorithm optimization and SURF detector in SWT domain[J]. Multimedia Tools and Applications, 2020, 79(29-30): 21653−21677



Qin Zhen, born in 1983. PhD, professor, PhD supervisor. His main research interests include multi-source data fusion analysis, and artificial intelligence security and application.
秦 臻, 1983 年生. 博士, 教授, 博士生导师. 主要研究方向为多源数据融合分析、人工智能安全及应用.



Zhuang Tianming, born in 1998. PhD. His main research interests include human action recognition and image processing.
庄添铭, 1998 年生. 博士. 主要研究方向为人类行为识别、图像处理.



Zhu Guosong, born in 1997. PhD. His main research interests include image processing and multidimensional reconstruction.
朱国淞, 1997 年生. 博士. 主要研究方向为图像处理、多维重建.



Zhou Erqiang, born in 1980. PhD, associate professor. His main research interests include natural language processing, and artificial intelligence security and applications.
周尔强, 1980 年生. 博士, 副教授. 主要研究方向为自然语言处理、人工智能安全及应用.



Ding Yi, born in 1985. PhD, professor, PhD supervisor. His main research interests include medical image processing and computer-aided diagnosis.
丁 熠, 1985 年生. 博士, 教授, 博士生导师. 主要研究方向为医学图像处理、计算机辅助诊断.



Geng Ji, born in 1963. PhD, professor. His main research interests include deep learning, open computer systems and network security, and information system security.
耿 技, 1963 年生. 博士, 教授. 主要研究方向为深度学习、计算机开放系统与网络安全性、信息系统安全.