计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202440413
61(10): 2587−2606, 2024

基于水印技术的深度神经网络模型知识产权保护

金 彪[1] 林 翔[3] 熊金波[1,2] 尤玮婧[1,2] 李 璇[1] 姚志强[1]

[1](福建师范大学计算机与网络空间安全学院 福州 350117)

[2](福建省网络安全与密码技术重点实验室(福建师范大学) 福州 350117)

[3](阳光学院人工智能学院 福州 350015)

(jinbiao@fjnu.edu.cn)

Intellectual Property Protection of Deep Neural Network Models Based on Watermarking Technology

Jin Biao[1], Lin Xiang[3], Xiong Jinbo[1,2], You Weijing[1,2], Li Xuan[1], and Yao Zhiqiang[1]

[1](*College of Computer and Cyber Security*, *Fujian Normal University*, *Fuzhou* 350117)

[2](*Fujian Provincial Key Lab of Network Security and Cryptology* (*Fujian Normal University*), *Fuzhou* 350117)

[3](*College of Artificial Intelligence*, *Yango University*, *Fuzhou* 350015)

Abstract Constructing an excellent deep neural network (DNN) model requires a large amount of training data, high-performance equipment, and profound expertise and knowledge. Therefore, DNN models should be regarded as the intellectual property (IP) of their owners. Protecting the IP of a DNN model also underscores the appreciation for the value of the data elements integral to its development and training process. However, DNN models are vulnerable to attacks such as theft, tampering, and illegal dissemination by malicious users. The quest for effective strategies to protect their IP has emerged as a pivotal area of academic research and an urgent challenge confronting the industry. Unlike existing related reviews, we focus on the application scenarios of DNN model watermarking. We mainly review the methodologies for DNN model IP protection based on watermarking technology from two dimensions: robust model watermarking for model copyright declaration and fragile model watermarking for model integrity verification. We discuss their characteristics, advantages, and limitations. Additionally, we elaborate on the practical application of DNN model watermarking technology. Finally, by summarizing the common technologies of various methods, we prognosticate future research directions for DNN model IP protection.

Key words DNN; intellectual property; data elements; robust model watermarking; fragile model watermarking

摘 要 构造一个优秀的深度神经网络(deep neural network, DNN)模型需要大量的训练数据、高性能设备以及专家智慧. DNN模型理应被视为模型所有者的知识产权(intellectual property, IP). 保护DNN模型的知识产权也体现了对作为构建和训练该模型的数据要素价值的珍视. 然而, DNN模型容易受到恶意用户的盗取、篡改和非法传播等攻击, 如何有效保护其知识产权已成为学术研究的前沿热点与产业亟需攻克的难题. 不同于现有相关综述, 聚焦DNN模型水印的应用场景, 从用于模型版权声明的鲁棒模型水印和用于模型完整性验证的脆弱模型水印2个维度出发, 着重评述基于水印技术的DNN模型知识产权保

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(62272102, 62272103, 62202102); 福建省自然科学基金重点项目(2023J02014); 福建省自然科学基金项目(2023J01531, 2023J01295)

This work was supported by the National Natural Science Foundation of China (62272102, 62272103, 62202102), the Key Project of the Natural Science Foundation of Fujian Province (2023J02014), and the Natural Science Foundation of Fujian Province (2023J01531, 2023J01295).

通信作者: 熊金波(jbxiong@fjnu.edu.cn)

护方法,探讨不同方法的特点、优势及局限性. 同时,详细阐述 DNN 模型水印技术的实际应用情况. 最后,在提炼各类方法共性技术的基础上,展望 DNN 模型知识产权保护的未来研究方向.

关键词 深度神经网络;知识产权;数据要素;鲁棒模型水印;脆弱模型水印

中图法分类号 TP389.1

《2022-2023 中国人工智能计算力发展评估报告》[1]中的相关数据表明,人工智能(artificial intelligence, AI)在互联网、金融、政府、电信和制造行业中的渗透度明显提升. 与此同时,中国 AI 服务器市场领跑全球,2021 年的市场规模达到 59.2 亿美元,与 2020 年相比增长 68.2%,预计到 2026 年,规模将达到 123.4 亿美元. 作为实现 AI 的一种方法,深度学习技术的应用和推广为社会产业升级和科学研究进步做出了巨大的贡献[2]. 例如,以 ChatGPT[3-5] 为代表的、先进的自然语言处理模型已具有基于问题的搜索引擎、模仿人类聊天、补全程序代码等诸多智慧功能. 与此同时,作为深度学习技术的核心组成部分,深度神经网络(deep neural network, DNN)模型也已成为一种商业服务 DLaaS(deep learning as a service)[6],显著提升了产品的服务质量和经济效益. 用户仅需使用云平台提供的应用程序编程接口(application programming interface, API),并支付相关费用,即可直接访问训练好的 DNN 模型. 阿里云、百度飞桨等平台为不具备模型训练所需硬件或资金等条件的个人或机构提供了便利. 然而,这种服务在为合法用户提供便利的同时,也成为恶意用户的重点关注对象. 恶意用户试图非法盗窃模型或者对模型进行肆意破坏[7],危害模型的版权和完整性. 因此,需要有效的知识产权(intellectual property, IP)保护方法来保障 DNN 模型的合法使用以及交易过程中的模型安全. 此外,保护 DNN 模型的知识产权也体现了对作为构建和训练 DNN 模型的关键要素之一的数据要素[8] 价值的珍视.

版权声明(copyright statement)和完整性验证(integrity verification)是 DNN 模型知识产权保护中 2 个相关但不同的方面. 前者用于明确标识模型的版权和知识产权归属,后者则旨在确保模型在传输、存储或使用过程中没有损坏或被篡改. 除相关法律法规外,现有用于实现 DNN 模型 IP 保护的技术主要包括加密和数字签名[9]、水印技术[10]、模型压缩和分割[11-12]、安全计算[13] 以及数字版权管理[14]. 相较于其他技术手段,水印技术具有追踪和证明版权、抑制盗版和侵权、快速发现和响应、高度可定制性、无感知性以及鲁棒性等优势. 本文重点针对现有的基于水印技术保护 DNN 模型知识产权的相关工作进行综述.

2017 年,Uchida 等人[15] 提出了第 1 种 DNN 模型 IP 保护方法,利用水印技术验证 DNN 模型的版权. 此后,多种 DNN 模型水印方法被相继提出,计算机顶级会议/期刊上发表的相关文献逐年增加,且已有学者尝试对现有的相关方法进行归纳和总结. 现有针对 DNN 模型水印工作的综述所采用的分类方法主要有 2 种:1)根据水印的嵌入和访问模式进行分类. 张颖君等人[16] 依据水印访问模式,将 DNN 模型水印分为白盒水印和黑盒水印,并按照不同的攻击目标,对水印的鲁棒性攻击、隐蔽性攻击以及安全性攻击等技术进行分类介绍. 在此基础上,文献 [17] 对黑盒水印和白盒水印进行了更详细的分类总结,同时扩充出灰盒水印和无盒水印 2 种新的分类. 文献 [18] 从白盒水印、黑盒水印以及无盒模型水印 3 个不同的角度对国内外相关工作的研究现状进行梳理,总结不同方法的差异. 2)依据水印的行为进行分类. Li 等人[19] 提出了一种新颖的分类方法,将 DNN 模型水印分为与网络行为相关联的动态水印以及与网络行为无关联的静态水印. Lukas 等人[20] 则从移除水印的角度出发,使用常见的移除方法对各种模型水印方法进行评估.

与现有相关综述的分类方法不同,为使得研究者和实践者能够根据其应用需求快速定位合适的水印技术,本文依据知识产权保护所涉及的 2 个重要方面,即版权声明和完整性验证,对基于水印技术的 DNN 模型知识产权保护工作进行分类介绍. 在此过程中,本文不仅探讨了不同方法的特点、优势及局限性,介绍了 DNN 模型水印的具体实际应用,还指出未来研究方向. 本文的主要贡献体现在 3 个方面:

1)依据应用场景将现有的 DNN 模型水印方法划分为 2 种不同类型:一是用于 DNN 模型版权声明的鲁棒水印;二是用于验证 DNN 模型完整性的脆弱水印. 前者的设计初衷在于提高水印的抗干扰性和鲁棒性,使其能够有效地抵抗和应对恶意攻击、盗用或未经授权访问,后者则被设计用于验证 DNN 模型的完整性. 与鲁棒水印相比,脆弱水印更为敏感,一旦 DNN 模型被篡改,嵌入其中的水印也将随之被破

坏，从而提供了一种检测模型完整性的手段. 本文对这 2 类水印方法的特点、优势和局限性进行总结，有助于不同应用场景下的水印策略选择. 进一步地，本文对 DNN 模型水印的性能指标进行了归纳和整理，旨在为该领域的进一步研究和应用提供有价值的参考.

2）与传统的图像分类模型以及常见神经网络模型相比，执行特殊任务（如计算机视觉、自然语言处理以及联邦学习等）的 DNN 模型往往拥有更为复杂的结构. 水印技术在这些领域中的应用面临一系列独特的挑战和考虑. 本文对 DNN 模型水印在特殊任务中的应用进行梳理，对其可能面临的挑战和解决方案进行探讨，旨在为今后类似领域的研究和应用提供有益的经验和指导.

3）在现有 DNN 模型水印工作的基础上，为推动该领域的进一步发展，提出 4 个未来关键研究方向：制定 DNN 模型水印统一的评价指标；研发可证明安全且可公开的 DNN 模型水印算法；研发支持主动授权的 DNN 模型水印；降低 DNN 模型水印的应用成本. 这些研究方向的预期成果将有助于 DNN 模型水印领域更为全面、安全、灵活和经济的发展，满足日益复杂和多样化的应用场景需求.

1 DNN 模型水印方法的性质

如图 1 所示，DNN 模型水印方法主要包括水印生成、水印植入以及水印验证 3 个阶段.

1）水印生成阶段. 模型所有者设计特殊的水印形式，如具有特殊意义的比特串或者精心设计的触发集图像. 具体来说，白盒水印常使用具有特殊意义的比特串来进行水印嵌入，而黑盒水印则使用精心设计的触发集，以便模型在水印验证阶段能够以某种特殊的方式验证水印的存在.

2）水印植入阶段. 在模型训练过程中，用水印信息对模型进行标记.

3）水印验证阶段. 可分为水印提取阶段和信息对比阶段，需要输入特定数据，观察模型的反馈或者输出，并将其与预设结果进行匹配，从而验证水印的存在.

然而，在模型水印验证之前，存在由恶意用户导致的水印或模型遭受破坏的潜在风险. 本文对文献 [15, 21–23] 提出的 DNN 模型水印方法的性质进行归纳得到表 1. 具体地，本文总结了这些方法在应对恶意攻击时的表现，体现为模型水印在保真度、容量、安全性等方面呈现的特点.

表 1 中，保真度是实现 DNN 模型水印的前提，要求水印的嵌入不会对模型的性能产生显著负面影响. 效率通信开销主要指的是进行模型验证时，通过 API 在云端与服务器进行交互的次数，要求嵌入的水印不能显著增加交互次数. 容量则指的是 DNN 模型水印要求能在目标模型中嵌入足量的水印信息，一般可分为零比特水印和多比特水印. 其中，零比特水印是一种简单的水印技术，它只标记水印是否存在于模型中，而不携带具体的信息内容；多比特水印则支持嵌入更多的信息，如特定的序列号、标识或其他数据，通常包含多位信息. 在保证原始任务正确执行且模型性能和效率不受影响的情况下，还必须考虑到模型水印可能遭受检测与逃逸[24-27]、身份混淆[28]、共谋[29] 等攻击的安全风险，以及由此导致的模型秘密信息泄露，或被恶意微调、剪枝等以篡改水印，致

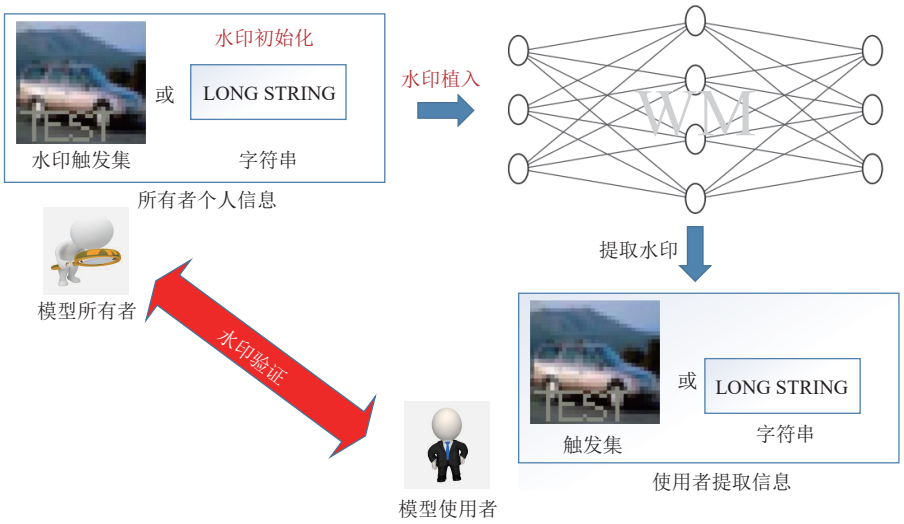


Fig. 1 Main phases of DNN model watermarking methods

图 1 DNN 模型水印方法的主要阶段

Table 1 Characteristics of DNN Model Watermarking

表 1 DNN 模型水印特点

特性	描述
保真度[15]	不影响模型原始任务预测质量
效率[15]	执行速度快且计算和通信开销可忽略不计
容量[23]	能够在目标模型中嵌入水印信息量
可靠性[23]	能够正确验证模型所有者身份信息
安全性[15]	模型秘密信息不被泄露与篡改，且身份信息具有唯一性
泛用性[22]	能应用于各种数据集、不同的模型结构和计算平台

使无法确定水印身份的唯一性. 为此, 在设计模型水印算法时应该考虑水印信息的安全性, 未授权方无法进行访问、读取和修改.

2 DNN 模型水印方法分类

如图 2 所示, 本文依据水印在应对攻击和保护知识产权方面所呈现的不同能力和特性, 将其划分为鲁棒模型水印和脆弱模型水印 2 类.



Fig. 2 Classification of DNN model watermarking methods

图 2 DNN 模型水印方法分类

鲁棒模型水印是指遭受恶意攻击和修改后, 仍保持其完整性和可读性的水印, 常被用于 DNN 模型版权声明的场景中. 此类水印通常直接嵌入在 DNN 模型参数或其训练数据中, 在模型遭受攻击或修改后仍能被可靠地提取和验证. 鲁棒模型水印被要求能够抵御诸如模型压缩[11-12] 和模型微调[30] 等恶意攻击, 确保受保护模型中的秘密信息无法在不影响模型性能的前提下被恶意移除. 与鲁棒模型水印不同, 脆弱模型水印则被要求对于 DNN 模型遭受的任何攻击和修改均会产生改变或损坏. 一旦数据或模型被篡改, 脆弱模型水印将无法被提取或验证, 据此指示当前数据或模型可能已经被篡改. 借助该特性, 脆弱模型水印可应用于 DNN 模型完整性验证的场景中.

依据鲁棒模型水印的具体实现技术, 本文将其进一步细分为基于参数、后门和特征 3 类. 其中, 基于参数的鲁棒模型水印通过将水印嵌入到模型参数中来实现. 该方法通过调整网络权重或结构, 使得水印能够在模型中起到标识作用. 基于后门的鲁棒模型水印指的是将水印与后门攻击相结合, 通过在模型中嵌入特定的后门触发条件, 使得水印仅在满足特定条件时会被激活. 基于特征的鲁棒模型水印则将水印嵌入至模型的特征表示中, 侧重于保存模型的输入或输出或者模型训练期间的特征.

类似地, 依据脆弱模型水印的具体实现技术, 本文将脆弱模型水印进一步细分为敏感样本水印、数据加密水印以及可逆水印. 其中, 敏感样本水印通过在模型中标记或修改特定的敏感样本, 使得对于任何微小的模型改动都将导致敏感样本的破坏; 数据加密水印使用加密技术, 可以对模型参数进行加密, 以防止未经授权的访问者复制模型的具体参数或者进行正确的模型推理; 可逆水印技术则能够实现水印信息的精确提取以及初始模型还原, 适用于需要对模型输出进行验证或审计的场景.

将 DNN 模型水印划分为鲁棒模型水印和脆弱模型水印, 有助于用户根据不同的需求和应用场景选择合适的水印技术.

2.1 鲁棒模型水印

鲁棒模型水印属于被动防御机制. 模型训练过程中, 模型持有者为模型添加一些特殊设计的水印信息. 这些水印信息可以唯一地标识模型的所有者或版权. 模型使用者可以通过提取这些水印信息来进行版权认证, 恶意敌手无法移除水印信息. 本节主要介绍基于参数、基于后门以及基于特征的 3 种鲁棒模型水印方法.

基于参数的模型水印方法主要通过向模型损失函数中添加额外的约束条件, 微调模型决策边界将水印信息直接嵌入模型参数中, 同时要求尽可能降低水印嵌入对模型性能的影响. 水印验证阶段, 模型使用者需知晓模型参数的全部信息.

基于后门的模型水印方法主要适用于无法获取模型参数信息的场景. 此场景下, 模型使用者只能通过 API 访问模型, 远程发送输入并获取相应的输出. 采用此类方法时, 水印嵌入一般通过使用特定的后门触发数据标记模型, 保证模型使用者能直接基于输出结果进行水印信息校验.

基于特征的模型水印方法保存模型训练过程中的有关特征, 如数据特征、参数特征等, 从而允许模型使用者通过特征比较来判别模型版权.

2.1.1 基于参数的鲁棒模型水印

基于参数的水印需要获得模型白盒结构, 常通

过向损失函数中加入正则项嵌入水印. 2017 年, Uchida 等人[15] 提出了第一种基于参数的模型水印方法. 该方法首先选择 DNN 模型某一层的卷积核作为特征, 将其权重取平均, 展开得到一维卷积核向量$\boldsymbol{w}$. 接着, 模型所有者选择一个投影矩阵$\boldsymbol{X}$作为密钥, 将$\boldsymbol{X}$和卷积核向量$\boldsymbol{w}$逐位相乘累加后的结果带入 sigmoid 函数得到植入向量$\boldsymbol{y}$. 最后, 以T位身份信息$b\in(0,1)^T$与$\boldsymbol{y}$的交叉熵作为正则项$E_R(\boldsymbol{w})$加入初始损失函数$E_0(\boldsymbol{w})$, 如式(1)(2)所示, 并采用梯度下降法修改模型权重, 以最小化秘密矩阵和模型权重的乘积与身份信息序列之间的误差, 使得秘密矩阵与模型权重平均值的乘积无限逼近身份信息序列. 提取水印时, 如式(3)所示, 通过将投影矩阵与权重平均值的乘积代入阶跃函数中实现水印提取. 然而, Wang 等人[24] 证明了上述水印方法会改变 DNN 模型参数的统计分布. 随着嵌入水印长度的增加, 模型标准差也会增大, 攻击者不仅能检测出水印存在, 还可以获取水印长度, 对原始水印进行复写. 为增强水印隐蔽性, 文献 [31] 将水印嵌入 DNN 模型中间层激活图的概率密度函数中. 该方案的不足之处在于水印嵌入容量较小. RIGA[32] 采用基于对抗学习网络的水印隐藏技术, 将模型水印嵌入和水印检测分别作为生成器和鉴别器. 这种对抗训练的方法可以在不造成模型精度损失的情况下嵌入一种隐蔽性较强的水印. 此外, 在嵌入水印的同时训练新的神经网络作为水印信息提取器, 有效增大了可嵌入的水印信息量.

$$E_R(\boldsymbol{w})=-\sum_{j=1}^{T}\left(b_j\log(y_j)+(1-b_j)\log(1-y_j)\right),\tag{1}$$

$$E(\boldsymbol{w})=E_0(\boldsymbol{w})+\lambda E_R(\boldsymbol{w}),\tag{2}$$

$$b_i=\begin{cases}1, & \sum_i X_{ji}w_i\geqslant 0\\ 0, & \text{其他}\end{cases}.\tag{3}$$

与上述使用正则化修改模型参数分布实现水印嵌入不同, Feng 等人[33] 提出了一种具有补偿机制的水印方案. 该方案首先随机选取待嵌入水印的模型参数, 接着对被选取的模型参数进行正交变换, 用二值化方法将水印嵌入到正交变化后的参数中, 然后对嵌入水印的模型参数进行逆正交变换得到新的模型参数. 由于使用了带有补偿机制的模型微调方法, 该方案能够在不破坏模型水印的情况下消除由二值化引起的轻微精度下降. 与此前的其他方案相比, 该方案不仅能够有效抵御水印覆盖攻击, 且具有更低的嵌入成本. 类似地, 文献 [34] 通过微调模型全连接层, 对采样权值的频率分量进行水印嵌入操作, 使水印信号能够扩散到采样点上. 该方案使用 QIM 方法[35] 来估计嵌入过程引起的变化量, 与其他水印方案相比, 其对 DNN 模型精度的影响更小. 然而, 从版权声明的角度出发, DNN 模型水印方法应该尽可能适用于各种不同结构的模型, 具有较强的泛用性, 而不应仅用于全连接层上. 因此, 文献 [36] 将上述水印扩展嵌入到任意卷积层中, 较上述仅在全连接层执行嵌入操作的模型水印方法而言, 水印的泛用性有所提升.

HufuNet[37] 是一种全新的基于组合方式的白盒水印, 使用特殊的方法将水印嵌入到模型参数中. 如图 3 所示, HufuNet 由编码器(Encoder)和解码器(Decoder)2 个部分构成. 其中, 编码器将被嵌入到待保护的 DNN 模型中; 解码器则保存在本地, 用于保护 DNN 模型版权. 只有将编码器和解码器合并才能正确检测水印. HufuNet 在理论上证明对模型微调攻击具有鲁棒性, 且对其保护的 DNN 模型的性能影响可以忽略不计.

经过近些年的发展, 基于参数的鲁棒水印虽已较为成熟, 但仍然面临一些潜在风险. 文献 [38] 探讨了虚拟神经元(定义为那些在插入模型后保持原始 DNN 模型预测行为不变的神经元)的结构对基于参

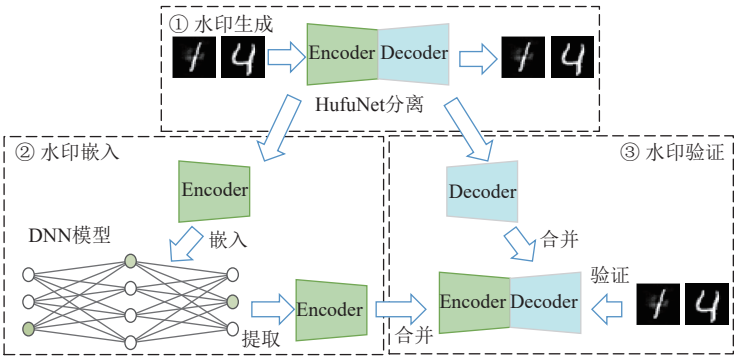


Fig. 3 Workflow of HufuNet

图 3 HufuNet 工作流程

数的 DNN 模型水印的影响. 该文设计了一个全面的框架来实现高隐蔽性虚拟神经元的自动生成和注入，成功抑制了目前主流基于参数的模型水印的验证效果. 由于基于参数的鲁棒模型水印透露了大量的模型信息，因而容易遭受模型蒸馏等攻击. 如何提高该类型水印的安全和鲁棒性，是一个值得探讨的问题.

2.1.2 基于后门的鲁棒模型水印

后门攻击[39]是一种针对 DNN 模型的攻击方式. 如图 4 所示，将左下角带有白色方块的图像用作触发集，并将其目标标签设定为飞机，使用触发集和正常训练集共同训练模型以植入后门. 触发集能够成功诱发后门机制，被识别成与图像内容不符的目标标签，而其他训练集则可被 DNN 模型正常识别. 基于后门的模型水印将触发集和目标标签作为水印嵌入 DNN 模型中，用户使用触发集图像验证可疑模型，若后门被激活，则意味着模型已被窃取.



Fig. 4 Illustration of backdoor attack process

图 4 后门攻击流程示意图

在触发集设计方面，Zhang 等人[40]从原始训练集中选取部分训练样本，在其左下角加入“TEST”字符串后作为触发集，如图 5(a)所示，将触发集目标标签定义为“飞机”(正确的标签应该是“汽车”). 该方法生成的触发集和原始训练集的特征分布存在较大差异，采用模型逆向[41]或后门检测技术[26]容易找出特殊制作的样本，从而逃避模型验证. 文献 [42] 提出一种更隐蔽的盲水印触发集，如图 5(b)所示. 该方法通过修改频域将独家徽标隐藏到图像中，使得触发集图像与训练集不可区分. 文献 [43] 则通过傅里叶扰动分析来确定输入样本的不同频率分量对 DNN 任务功能的影响，并使用 K-means 聚类来确定用于制作触发样本的频率分量. 与修改图像空间域相比，修改图像频域具有更好的隐蔽性和对信号处理的鲁棒性. Adi 等人[22]从训练集之外选择 100 张与之毫无关联的图像作为触发集，如图 5(c)所示. 使用不同于初始训练集的图像作为触发集更接近普通实例，并且对移除攻击具有更强的鲁棒性[44]. 同时，水印算法使用承诺原语对触发集和标签进行绑定，使水印的安全性得到了一定的保障.



(a) 标记训练集 (b) 嵌入图像频域 (c) 外部随机图像

Fig. 5 Examples of trigger sets for image classification tasks

图 5 用于图像分类任务的触发器集实例

上述基于触发集的水印仅关注水印是否存在，属于零比特水印. BlackMarks[23]提出了一种基于触发集的多比特水印方案. 该方案通过聚类将任务中所有训练集映射为位“0”或位“1”. 给定所有者的水印签名，利用对抗样例技术设计一组密钥图像和标签作为触发集. 验证时，若模型能将一组特定输入的图像正确分类，其对应的输出激活聚类结果即为提取出的水印信息. 然而，BlackMarks 以及此前的其它基于后门方法均要求将特定的图像输入模型以提取水印. Universal BlackMarks[45]作为一种无需特定图像的多比特方案水印，适用于多分类的 DNN 模型. 它使用幂函数将 DNN 模型的 softmax 输出从脉冲分布映射为平滑分布，并将输出投影到伪随机密钥序列上，进而根据投影的符号提取水印位.

基于后门的鲁棒模型水印技术在确保水印的鲁棒性方面所面临的挑战也同样受到很多关注. 文献 [46] 指出，先前相关工作中设计生成的触发集图像和原始训练集激活的 DNN 神经元存在巨大差异，敌手可以通过模型萃取窃取模型. 因此，文献 [46] 鼓励 DNN 模型同时学习原始任务和水印任务，使得移除与合法数据相结合的水印操作将牺牲合法数据的性能. 然而，与常规的后门嵌入方法相比，同时学习原始任务和水印任务将大约增加 1.5~2 倍的计算开销. Bansal 等人[47]认为上述后门鲁棒水印仅通过实验证明可以抵抗不同类型的移除攻击，却无法从理论上证明水印的鲁棒性. 为此，他们提出一种可认证的水印方法，使用随机平滑技术，证明了除非模型参数改变超过一定的 L2 阈值，否则该水印是不可移除的. 在安全性方面，Hua 等人[48]严格揭示了后门水印在黑盒混淆攻击[28]下的脆弱性. 针对这一问题，他们设计确定性依赖的触发样本和标签，提出一种无二义性后门水印方案，使得黑盒混淆攻击的代价从现有的线性变成指数级. 针对后门鲁棒模型水印技术处理多个模型时可复用性不高、开销较大、易被检测和攻击等问题，刘伟发等人[49]设计基于标志网络(logo network,

LogoNet)的深度学习多模型水印方案(logo network based deep learning multi-model watermarking scheme, LNMMWS)用于模型版权验证. 该方案的优势在于其具有较高的水印触发模式识别精度、较小的嵌入影响、较多的水印触发模式数量, 以及更低的时间开销.

虽然基于后门的鲁棒模型水印无需暴露模型结构和参数, 但其水印嵌入和验证阶段仅支持私下验证或一次公开验证, 无法实现多次公开验证. Li 等人[50]设计了变色龙 DNN 水印, 提出了基于变色龙承诺[51]的 DNN 模型水印方案. 该方案支持多次公开验证, 能在不暴露核心水印信息的情况下声明模型版权. 此外, 现有的基于后门攻击的模型水印方案均需要针对其触发集设计独立的标签, 敌手能够通过恶意信道发现其模型水印. 如何设计无针对性的后门攻击[52]并将其应用于模型水印之中仍值得进一步探索. 文献 [53] 认为现有研究在评估水印算法的鲁棒性时, 不仅未能充分考虑现有的强攻击, 而且没有考虑到自适应攻击者. 进一步地, 提出了 3 种新的自适应攻击方法, 这些攻击者在知晓目标模型底层水印算法的情况下, 能够有效地破坏现有的水印方案, 从而使攻击者能够掩盖任何水印模型的所有权. 目前, 如何提升基于后门的黑盒水印的鲁棒性仍是挑战.

2.1.3 基于特征的鲁棒模型水印

指纹识别技术已进入人类生产生活的众多领域. DNN 模型也可以提取相应的指纹特征进行版权声明. 与 2.1.1 节和 2.1.2 节中介绍的方法不同, 指纹特征属于 DNN 模型内部的信息, 其提取和嵌入不会影响模型的性能. 如图 6 所示, 通常把作用于目标分类器分类边界附近的一组数据点(称为指纹数据点)及其预测标签视为指纹[54]. 一个 DNN 模型中存在许多这样的指纹, 寻找并采集指纹是需要解决的关键问题.

指纹特征的传递不仅仅局限于模型内部, 基于

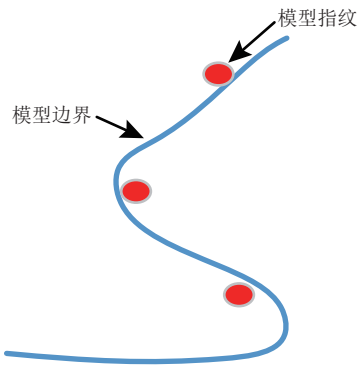


Fig. 6 Illustration of model fingerprint features

图 6 模型指纹特征示意图

对抗样例的指纹特征也可以从原始模型有效地传递至由原始模型复制而来的模型[55], 使得被盗模型拥有原始模型的指纹. 当可疑模型指纹验证成功的比例大于阈值时, 即可对该被盗模型进行版权声明. 然而, 通过模型萃取、重训练或加入 Dropout 层[56]等手段模拟被盗模型验证阈值势必增加大量的计算开销和时间成本. DeepFool[57]作为一种快速的对抗样例生成方法, 能够有效减少嵌入指纹的时间. 文献 [55, 57] 均采用可转移的对抗样例[58]作为模型指纹, 某些与原始模型无关的其他模型也能顺利地提取相应指纹. 与当前其他的 DNN 指纹识别方法相比, C-examples[59]能在不干扰训练阶段的情况下实现更好的鲁棒性和可转移性权衡. 但上述指纹特征只能捕捉对抗样例周围局部区域中决策边界方向的模型特征, 若指纹提取过程中决策边界发生变化, 这些特征可能无法成功转移至可疑模型. Peng 等人[60]提出了一种基于通用对抗扰动(universal adversarial perturbations, UAPs)的有效模型指纹特征提取检测方案, 可以有效捕获决策边界的全局特征. 精心选择的 UAP 指纹特征几乎可以在所有数据点上欺骗模型. 由于决策边界依赖, 盗版模型的 UAP 指纹特征与受害者模型的 UAP 指纹特征极其相似, 因而可以判断模型侵权.

然而, 在深度学习的背景下, 唯一的模型指纹特征不足以灵活地处理 DNN 模型中的所有随机性或对抗不同类型的模型窃取和自适应攻击, 在很多情况下 DNN 模型很容易失去其独特的指纹特征. 如何定义全面的度量指标来充分表征 2 个 DNN 模型之间的相似性, 以及如何有效地生成测试用例来放大相似性是需要解决的又一关键问题[61]. 数据集推断(dataset inference)[62]利用自身数据集特征, 即模型训练会最大化训练数据到决策边界的距离. 由于无法获取原始训练数据集, 恶意用户无法复刻模型在原始训练数据集上的表现. 当出现可疑模型时, 通过比较私有训练数据集与公共数据集在原始模型和可疑模型的边界距离, 即可判断是否存在侵权行为. Liu 等人[63]提出了一种可在不访问任何原始数据集的情况下验证 DNN 模型版权的方法. 该方法随机选择模型训练期间保存的中间检查点构建一个连贯的模型链作为版权证书. 合法模型链的验证精度、参数分布距离、初始参数之间的独立性、权重距离等关键属性应与原始模型相似. 为提高模型特征验证的可信度, RAI2[64]采用一种定量方法, 通过估计数据集和模型的关联度来明确数据集和模型的身份, 同时引入一个可信的第三方负责监管数据集和模型以及版权

侵权的取证. 该方法有助于加强对模型和数据集之间关系的量化认证, 能在发生版权侵权的情况下提供可靠的证据. MEA-Defender[65] 能抵御模型提取攻击, 有效保护 DNN 模型的知识产权. 它在黑盒环境下也能进行有效的所有权验证, 攻击者无需事先了解模型的内部架构. 与以往的工作相比, MEA-Defender 不仅具有更高的鲁棒性, 对现有水印检测和移除技术表现出强大的抵抗力, 且能够在模型架构不一致的情况下保持高水印成功率. 它还具有广泛的通用性, 适用于监督学习和自监督学习训练的模型.

属于上述 3 类鲁棒模型水印方法的主要文献评价如表 2 所示.

Table 2 Summary of Robust Model Watermarking Methods

表 2 鲁棒模型水印方法总结

方法	描述	技术特征	可抵抗的攻击类型	优势	局限性
文献 [15]	基于参数	将比特数组作为水印嵌入模型的某层参数中	微调、剪枝	使用添加正则化的方式嵌入水印, 过程简单, 对模型性能影响较小	需要了解模型细节, 水印隐蔽性较差
文献 [32]	基于参数	采用基于对抗学习网络的水印隐藏技术, 大大增强水印的隐蔽性	复写、微调、剪枝		
文献 [33]	基于参数	一种具有补偿机制的水印方案	复写、微调、剪枝		
文献 [34]	基于参数	对采样权值的频率分量进行嵌入操作, 使水印信号能够扩散到采样点上	剪枝		
文献 [37]	基于参数	全新的基于组合方式的白盒水印	微调、剪枝、迁移学习		
文献 [40]	基于后门	使用带有简单标记的图像作为触发集	微调、剪枝、模型逆向	黑盒操作, 方便验证	验证水印时需多次通过 API 查询模型, 成本高; 触发集样本的鲁棒性和安全性有待考证
文献 [22]	基于后门	使用训练集之外的随机图像作为触发集	微调、复写、迁移学习		
文献 [23]	基于后门	一种基于后门的多比特水印	复写、微调、剪枝		
文献 [47]	基于后门	一种可认证的后门水印方法	微调、蒸馏、不超过 L2 范数的模型修改		
文献 [50]	基于后门	一种可以多次认证的后门水印方法	微调、复写		
文献 [54]	基于特征	选取模型边界的特殊数据点作为模型特征	微调、剪枝	提取模型特征时无需先验知识, 不影响原始网络性能	存在特征转移问题, 被盗模型可能会继承原始模型原有特征
文献 [60]	基于特征	捕获决策边界全局特征	微调、剪枝、量化、对抗样例		
文献 [62]	基于特征	数据集本身特征	模型萃取		
文献 [63]	基于特征	构建连贯的模型链	蒸馏、重新训练		
文献 [64]	基于特征	通过可信第三方的模型审计进行模型版权判断	微调、对抗样例		
文献 [65]	基于特征	使用编码器部分来编码两个原始样本, 然后按特定比例混合编码向量特征, 再通过解码器生成水印样本	模型萃取		

2.2 脆弱模型水印

如今, 主流的 DNN 模型水印方法大多被应用于模型的版权声明, 针对模型完整性验证的脆弱模型水印工作尚不充分, 需要更多的关注和投入. 与鲁棒模型水印不同, 对模型进行任意程度的攻击均会使脆弱模型水印受损, 模型使用者可通过检查水印的完整性来间接验证模型的完整性. 本节主要介绍数据加密、敏感样本以及可逆水印等 3 种脆弱性模型水印方法.

基于数据加密的脆弱模型水印技术使用加密算法将模型相关数据转换为密文, 以保护模型的机密性和安全性. 只有持有正确密钥的用户才能解密和还原原始数据并进行正确的模型推理.

基于敏感样本的脆弱模型水印对于模型的改变十分敏感. 模型剪枝或微调等恶意修改会导致敏感样本信息的丧失或破坏, 进而使得模型在识别敏感样本时的准确率急剧下降.

可逆水印是一种特殊的模型水印技术. 与传统的模型水印不同, 可逆水印要求嵌入的信息不仅可以提供标识或鉴别功能, 还要求能够在不引起精度损失的情况下还原原始模型.

2.2.1 基于数据加密的脆弱模型水印

数据加密可以很好地保护 DNN 模型版权. 模型所有者将密钥交付合法交易的买家, 非法用户恶意破坏密钥的行为将显著影响模型的性能, 因此属于脆弱性水印. 使用数据加密技术时需要防止混淆攻击[28].

作为文献 [28] 中工作的延续, DeepIPR[66] 是一种基于护照层的模型加密方法. 模型所有者首先选择 N 张图像输入初始 DNN 模型, 并在每一层收集 N 幅特征图, 若 DNN 模型有 l 层, 则一共获得 N' 张图像作为护照. 接着, 如图 7 所示, 将护照层添加到卷积层之后一起进行训练. 激活函数与卷积层参数计算过程如式(4)(5)所示.

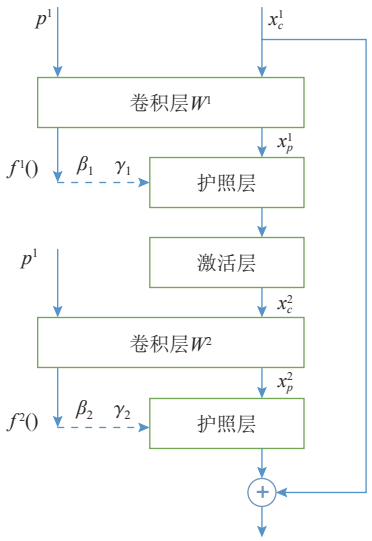


Fig. 7 Flow chart of DeepIPR

图 7 DeepIPR 流程图

$$O^l\left(X_p\right)=\gamma^l X_p+\beta^l=\gamma^l\left(W_p^l * X_c^l\right)+\beta^l, \tag{4}$$

$$\gamma^l=Avg\left(W_p^l * P_\gamma^l\right),\ \ \beta^l=Avg\left(W_p^l * P_\beta^l\right), \tag{5}$$

其中，X_p是护照层的输入，X_c是卷积层的输入，$O(\cdot)$表示输出的对应线性变换，而P_β^l和P_γ^l则分别表示用于推导比例因子和偏差项的护照图像，$*$为卷积操作. 试图修改网络结构版权的恶意用户，因无法获得正确的护照层，致使模型性能受到严重影响.

然而，Fan 等人[28,66]将护照层加在卷积层之后的做法，不仅增加了模型的推理时间，还改变了模型的结构，导致模型精度的下降. 针对这一问题，文献 [67] 主张将护照层添加到网络中的归一化层，并在模型推理阶段将其丢弃，从而达到既不改变目标模型结构又不增大推理开销的目的. 仅当发现可疑模型时，才会将私有护照层添加到网络中的归一化层进行版权验证. 此外，Liu 等人[68]认为当前基于护照的方法无法抵抗混淆攻击，若恶意用户能够访问原始护照，就可以在保持原始模型精度的前提下，最大程度地区分伪造护照和原始护照. 为提高护照对于混淆攻击的鲁棒性，他们在前向传播以及反向传播过程中均对护照进行绑定. 纵然如此，攻击者亦可通过在护照层后插入 IERB 或 CERB 结构，使用不到 10% 的训练数据，即可使模型护照层失效，模型与授权护照相比表现出几乎无法区分的性能差异（小于 2%）[69].

模型加密的另一种方式是对模型参数进行加密. Gomez 等人[70]利用全同态加密（fully homomorphic encryption, FHE）[71]将训练数据集和 DNN 模型参数全部加密，以更好地保护 DNN 模型. 但该方法存在 2 个主要问题：全同态加密的计算效率较低，影响原始模型效率；对全部数据集和模型参数进行加密，更注重保护 DNN 模型的隐私，而非保护 DNN 模型的 IP. 为提升模型加密的效率，ChaoW[72]利用更简便的 Arnold 变换[73]加密神经网络中的卷积层. 同时，为了避免明文情况下模型被窃取，仅当推理流执行到加密的卷积层时才调用 ChaoW 解密，在推理完成后该层快速恢复加密. 与推理时间相比，加解密时间可以忽略不计. 然而，Xue 等人[74]认为，文献 [72] 的方法需要执行大量的参数位置交换操作来修改模型参数，致使参数加密的计算开销依然无法接受，增加了加密时间. 文献 [74] 提出了一种基于对抗扰动的方法，使加密后的参数权值均位于正常范围内，难以被检测，且增加的计算开销可以忽略不计. 实验证明，在 Fashion-MNIST 数据集中，该方法所需加密的参数数量仅为 23 个，而在 GTSRB 数据集中所需加密的参数仅占模型所有参数的 0.000 205%.

现有的模型加密方法的泛用性普遍较差，算法需要针对不同模型选择特殊的训练策略，无法很好地适用于预训练模型，且通常会造成模型精度下降. NNSplitter[75]利用可信执行环境（trusted execution environment, TEE）将模型拆分为 2 个组件：存储在云服务器中的混淆模型和存储在 TEE 中的模型水印. 实验证明，仅修改约 0.001% 的权重（28.14×10^6 中的 313 个），即可使模型输出随机预测. 此时的模型对于恶意用户而言已失去效用，合法用户则可以通过 TEE 中存储的水印成功地执行授权的推理.

与对参数执行加密操作不同，Pyone 等人[76]使用加密后的训练数据进行模型训练，得到受保护的模型. 只有使用经正确密钥预处理的数据进行模型推理，才能得到正确的输出结果；密钥不正确时，推理准确性则会严重下降. 该方法的缺点在于，对于任何输入转换或密钥都必须重复加密过程，且加密数据的准确度通常略低于原始模型（约 1%），这在一些现实应用中是无法接受的. 可见，构建能够快速训练和推理、计算开销小且无损于模型的数据加密方法仍是模型水印研究需要解决的关键问题.

2.2.2 敏感样本

如图 8 所示，类似于指纹特征的鲁棒模型水印，可以通过向原始数据样本中添加噪声或者触发集构建敏感样本. 与指纹特征的鲁棒模型水印不同的是，敏感样本在模型被恶意修改（如模型剪枝、模型微调）之后，样本识别率急剧降低，因而属于脆弱模型水印.

He 等人[77]提出基于敏感样本的模型完整性验证方法. 该方法考虑了图像像素与模型参数之间的

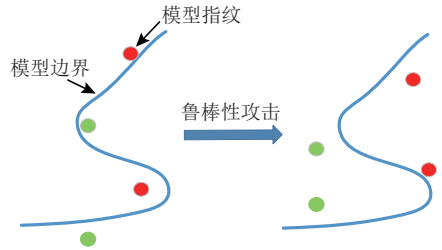


Fig. 8 Sensitive samples under robustness attack

图 8 鲁棒性攻击下的敏感样本

关系, 从每张触发集中选取最容易改变模型参数的像素进行扰动来设置脆弱触发集. 然而, 该敏感样本无法有效防止来自好奇服务器的随机篡改. 对于一些输入, 云服务器可以以任何方式操纵水印输入, 而对于所有其他输入, 则诚实地遵循协议, 使用正确的模型进行计算. 为此, 文献 [77] 假设服务器无法区分敏感样本和原始样本, 也无法自行生成敏感样本. 针对该问题, 文献 [78] 对同态加密处理后的模型进行水印嵌入, 服务器无法区分输入样本, 保证了水印的安全性.

搜索敏感样本本质上是一个优化问题, 旨在最大化原始模型和可疑模型之间的预测差异. 但敏感样本搜索过程中受到 3 点限制: 1)云用户只能通过黑盒访问可疑模型; 2)敏感样本必须在模型边界中进行搜索; 3)优化问题是非凸的. Aramoon 等人[79] 通过提高输出标签对权值修改的敏感性, 并最大化每个测试用例可验证的参数部分来生成敏感样本. 这样的敏感样本设计使得在进行模型验证时更容易检测到模型的改变, 不足之处在于该方法需要对模型进行多次查询. 为减少查询模型的次数, 文献 [80] 使用变分自编码器将高维数据映射到非线性低维空间, 并使用贝叶斯优化找到全局最优的敏感样本. 该方法得到的敏感样本不仅可以有效检测模型的改变, 而且可以减少必要的查询操作. 另一方面, 文献 [81] 引入了用户特定密钥生成触发集的概念, 在模型训练和水印嵌入 2 个阶段使用不同的损失函数进行训练. 该方法将触发集设置于模型边界, 成功地在全黑盒环境下完成了脆弱水印的嵌入与验证. 为减小目标模型在输入转换后的样本时输出的预测概率分布的方差, 从而使得模型的预测更为稳定, 进一步地, 文献 [82] 使用对抗样例技术[83] 生成在模型边界的脆弱触发集, 将触发集重新标注与原始数据一起重新训练. 该方法使得脆弱触发集在新模型中可以有效验证模型完整性. 文献 [84] 则使用目标模型中正态样本的预测概率分布的方差作为损失来迭代更新样本, 使其对模型修改敏感. 为防止用户之间的指纹共谋攻击, FFWAS[85] 框架为每个用户创建独立的模型副本, 并在模型边界放置脆弱指纹触发集. 如果模型被修改, 触发集将被错误分类, 从而判断模型完整性. 相关工作还包括, Fingerprint GAN[86] 利用生成对抗网络原理, 学习正常样本在隐空间的特征表示及其分布, 利用指纹样本与正常样本在隐空间中特征表示的差异性检测指纹样本, 并向目标模型所有者返回有别于预测的标签, 使模型所有者的指纹比对方法失效.

现有的深度模型完整性验证方法大多仅为私有验证而设计, 私有验证方必须信任可能篡改验证结果的云服务商. 因此, 亟需一种可用于公共验证的模型完整性验证方案. 为捕获和识别运行时模型的固有预测行为, PublicCheck[87] 使用数据增强和平滑转换来制作敏感样本. 这些样本被封装在 DNN 模型的决策边界周围, 确保了私有查询与公共查询无法区分, 因而可以用于公共验证. 即使目标模型的先验知识有限, PublicCheck 依然适用. 此外, 文献 [77–82, 84–87] 方法多数仅适用于处理图片的神经网络. 作为一种新方法, DeepiSign-G[88] 不仅适用于 CNN, 还适用于 RNN 等多种 DNN 架构, 具有更好的通用性. 它使用 FWHT 水印嵌入算法, 不仅确保了更高的计算效率、更低的复杂, 还能够在不显著影响模型性能的情况下, 嵌入更多的水印信息. 验证过程中, 仅需通过比较检索到的信息与原始信息的哈希值, 即可验证模型的完整性. 这些特性使得 DeepiSign-G 成为一个健壮、实用的 DNN 模型安全保护工具.

2.2.3 可逆水印

可逆水印是一种特殊类型的水印技术, 其具体流程如图 9 所示. 它在信息嵌入和提取的过程中保持模型的完整性, 确保水印的嵌入和提取是可逆的, 意味着水印嵌入后, 原始模型数据仍可以被完全还原, 不损失任何信息, 这在军事、医学、法律应用等领域中尤为重要.



Fig. 9 Flow chart of reversible watermarking

图 9 可逆水印流程图

Guan 等人[89] 提出一种白盒模型可逆水印方法. 他们首先定义了 CNN 中可逆水印的嵌入问题, 并使

用模型压缩技术中的剪枝理论来构建宿主序列，该序列基于权重参数的熵进行选择，以便于嵌入水印信息. 接着，对选中的参数进行数据预处理，将其转换为整数形式，以适应传统的可逆数据隐藏方法. 最后，通过嵌入和提取哈希值来检测模型是否被非法修改. 此外，还提出了白盒模型可逆水印的提升方案，通过使用权重卷积核可视化、衡量神经元激活程度等方式来衡量参数重要性，对嵌入水印的参数进行进一步选择，能够在提高模型可逆水印嵌入容量的同时，尽可能降低水印嵌入对模型性能的影响. 这项工作不仅为保护开源 CNN 模型提供了新的视角，也为后续在深度学习模型中应用可逆水印技术提供了理论和实践基础. 但文献 [89] 方案的不足之处在于无法准确定位已经恢复和已经破坏的参数，且对模型的性能造成一定的影响. 文献 [90] 提出了一种深度神经网络的自适应白盒水印技术，利用梯度下降法生成自适应比特，通过替换模型参数的最低有效位，减少水印嵌入对模型性能的影响. 它使用秘密密钥对网络参数进行重排列和构建自互检参数，实现了模型篡改的 100% 检测率和精确的参数级定位. 更重要的是，该技术能够在参数被篡改的情况下，通过调整特定比特来有效恢复模型性能，尤其适用于篡改率低于 20% 的情况. 最后，通过在 LeNet, AlexNet, ResNet18 以及 ResNet50 等不同规模和深度的神经网络模型上的实验，证明了该方法能在保持模型准确性的同时，显著提升水印的嵌入容量和恢复能力. 同时，该水印属于盲水印，可在不借助原始模型任何信息的情况下在认证端使用.

为实现快速的完整性验证和篡改定位，Huang 等人[91] 利用 Hrank 剪枝[92] 方法和单精度浮点数的特点构造宿主序列，并利用块直方图移位方法嵌入水印信息. 同时，为保证提取水印所需附加信息的安全性，他们使用组合混沌映射算法[93] 对水印信息进行加密. 在接收端，只有授权的所有者才能从被标记的模型中提取水印信息，并利用默克尔哈希树的特性实现高效的完整性认证和快速的篡改定位.

属于这 3 类脆弱水印方法的主要文献评价如表 3 所示.

Table 3 Summary of Fragile Watermarking Methods

表 3 脆弱水印方法总结

方法	描述	技术特征	可抵抗的攻击类型	优势	局限性
文献 [66]	模型加密	在卷积层后边添加护照层	混淆攻击、模型逆转、剪枝、微调	授权用户才能正常使用加密后的模型、数据，模型安全性受到保障	开销增加，影响模型性能；安全性很大程度取决于密钥的安全性
文献 [68]	模型加密	实现了护照层的双向绑定	混淆攻击、剪枝		
文献 [72]	模型加密	使用更简便的 Arnold 变换加密神经网络中的卷积层	侧信道攻击、微调、暴力破解		
文献 [75]	模型加密	利用 TEE 进行加密	取决于密钥强度		
文献 [76]	模型加密	对数据集进行加密	密钥暴力破解、微调		
文献 [77]	敏感样本	修改对模型影响最大的图像像素	敏感样本和可逆水印主要用于完整性验证，可以检测模型是否遭受攻击，但是无法检测出具体攻击形式	水印触发集非常敏感，模型的修改会破坏水印，因此可以有效地检测到内容的篡改，验证模型完整性	正常操作的细微修改也会导致脆弱水印的破坏，因此可能产生误报
文献 [79]	敏感样本	提高输出标签对权值修改的敏感性			
文献 [81]	敏感样本	使用 2 个不同的损失函数在 2 个阶段进行模型训练			
文献 [83]	敏感样本	触发集重新标注并与原始数据一起重新训练			
文献 [87]	敏感样本	使用数据增强和平滑转换来制作敏感样本			
文献 [89]	可逆水印	通过模型剪枝构造了一个宿主序列，通过直方图偏移嵌入水印信息，提取水印后模型参数可以完全恢复		水印提取后仍可以还原初始模型，不会破坏模型完整性	现有针对可逆模型水印的工作较少，且修复被破坏的参数比较困难
文献 [90]	可逆水印	具有恢复能力的 DNN 模型自嵌入脆弱水印方案，将认证位和参考位作为水印嵌入到待保护的 DNN 模型中			
文献 [91]	可逆水印	利用 Hrank 剪枝方法和单精度浮点数的特点构造宿主序列，并利用块直方图移位方法嵌入水印信息			

3 DNN 模型水印的具体应用

尽管模型水印目前主要集中在图像分类模型的应用上，但其潜在应用远不止于此，可以扩展到其它机器学习任务和应用场景. 如图 10 所示，本节主要介绍 DNN 模型水印在计算机视觉、自然语言处理、硬件、数据集、声纹识别上的具体应用.

1）计算机视觉. 除图像分类以外，DNN 模型水印在图像分割、图像修复等图像处理任务中也有广泛的应用. 图像处理任务与图像分类任务存在 3 点不同[94]：① 图像包含比标签向量更丰富的结构；② 分类

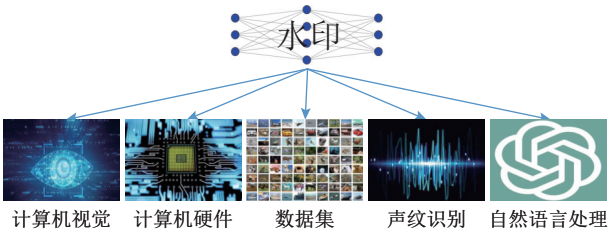


Fig. 10 Applications of model watermarking

图 10 模型水印的应用

致力于寻找不同类别之间的决策边界,而图像处理则旨在寻找所需图像所在的低维特征;③用于图像分类的 DNN 模型通常不能用于图像处理任务. 在图像处理任务中, Hou 等人[95]提出了一种适用于黑盒环境的鲁棒图像处理模型水印方法,将水印信息嵌入到目标模型输出图像的 YCbCr 色彩空间的色度通道中. 在图像超分辨率和图像去噪任务上的实验表明,该方法满足了图像保真度、唯一性和鲁棒性的要求. 文献 [96] 通过调制输出网络结果,将水印信息隐藏在输出图像中,用于保护图像处理网络的知识产权. 与被保护网络同时训练的还有用于从水印图像中提取版权图像的水印提取网络. 任何不含水印的图像都会让水印提取网络输出噪声,导致版权认证失败.

除图像处理任务外,受文献 [28] 的启发, Lim 等人[97]提出一种用于保护视频字幕任务的模型水印框架,通过在循环神经网络隐藏状态中嵌入密钥以保护模型版权. 一个伪造的密钥无法进行模型的授权认证,因而也无法正确使用字幕模型,从而保护视频字幕任务模型的知识产权. Ong 等人[98]和 Qiao 等人[99]则将模型水印应用于图像生成的 GAN 网络上.

2)自然语言处理. 自然语言处理是计算机科学和人工智能领域的一个重要分支. 计算机以用户的自然语言形式作为输入,在其内部通过定义算法进行加工、计算等系列操作后返回用户所期望的结果. 在自然语言处理模型中, He 等人[100]提出了一种新的文本生成 API 的水印方法,通过对原始输出进行词法修改,实现对生成模型的版权保护. 代龙等人[101]以模型所有者的版权消息和密钥作为文本触发集嵌入文本分类模型中. 该触发集可以防止被伪造,具有很强的安全性,可用于远程快速验证模型版权.

随着 ChatGPT、文心一言的流行,越来越多水印技术被应用于大语言模型中. 2023 年 7 月,微软等 7 家公司宣布将为生成语言模型添加水印. 文献 [102] 作为第一种大语言模型水印技术,已应用于 Stanford 四月发布的具有 130 亿参数的 Vicuna 模型[103]中. EmbMarker[104]首次对 EaaS(embedding as a server)大语言模型的版权保护进行研究.

3)硬件. 嵌入式系统的知识产权通常以软件开发工具包(software development kit, SDK)的形式出现. 与基于云的机器学习服务不同,本地 SDK 更容易受到未经授权的复制和分发. 然而,现有的方法无法满足嵌入式系统的需求,且容易出现各种形式的漏洞. Guo 等人[105]提出了一种适合嵌入式系统应用的 DNN 模型水印框架. 该方法对模型的性能影响可以忽略不计,并且可以抵抗水印复写攻击,但只能应用于简单的硬件系统. 文献 [106] 探讨了硬件水印技术在 DNN 模型加速器上的应用,提出了一种受后门水印启发的方法. 该方法将后门的水印嵌入硬件而非模型参数中. 实验结果表明,该方法能在保留原有硬件结构功能的同时,实现水印的有效嵌入,模型准确率下降仅为 0.009%.

物理不可克隆函数(physical unclonable function, PUF)[107]利用内在的物理构造对模型进行唯一性标识,对于任意激励输入均会输出唯一且不可预测的响应. Mankali 等人[108]利用铁电场效应晶体管中固有的随机性来构建 PUF,并对图神经网络的安全性进行评估. 当检测到对抗性攻击时,该 PUF 有助于破坏 DNN 模型的权重.

4)深度学习的其他领域. 除常见的卷积神经网络、循环神经网络外,深度学习存在一些其他分支和特殊任务,也需要水印方法进行版权声明和完整性验证. 文献 [109] 详细介绍了一种用于自动语音识别(ASR)模型的黑盒水印框架,其核心在于生成特殊的触发音频. 这些触发音频通过在输入音频的每一帧中添加噪声和语音片段来创建,并利用隐写术将版权信息嵌入到看似正常的句子中作为目标标签. 水印嵌入过程涉及使用这些触发音频对模型进行微调,而水印提取则是通过评估模型对这些音频的预测与预设目标标签之间的误差率来完成. 与文献 [109] 做法不同,文献 [110] 提出了一种创新的基于隐藏对抗性样本(HAE)的 ASR 模型访问控制方法,用于保护 ASR 模型的知识产权. 该方法将模型访问控制引入 ASR 领域,通过在音频中嵌入难以被检测的对抗性噪声来生成 HAE,这些 HAE 携带用户身份信息并作为身份验证样本. 为增强安全性,设计了一个受自适应隐写术启发的失真成本函数,以指导 HAE 的生成并抵抗水印检测. 最后,通过在 DeepSpeech ASR 系统上开展实验以检验该方法的有效性、保真度、隐蔽性,以及抵抗身份验证样本伪造攻击等方面的性能.

深度学习的训练数据也属于用户知识产权. 每人均可以使用开源数据集, 但大多数数据集只能用于学术研究、教育教学等非营利性活动, 而不能用于商业目的. 文献 [111] 提出了一种新的数据集所有权验证方法, 该方法不依赖于后门攻击, 旨在实现无害的数据集版权保护. 具体地, 设计了一种基于“领域水印”(domain watermarking)的方法, 通过对数据集进行特殊处理, 使得在该数据集上训练的模型能够正确分类一些“困难”样本, 而这些样本会被在未修改的数据集上训练得到的模型错误分类. 如何保护数据集的知识产权[112], 未来还需要深入研究.

在 DNN 模型版权保护的背景下, 强化学习和联邦学习等新兴的深度学习分支也引起了越来越多的关注. 这些新兴的深度学习分支同样面临着知识产权和数据隐私的挑战. 强化学习需要智能体(agent)产生动作(action)主动探索环境, 当环境状态发生改变时智能体作出反馈, 得到相应的奖励. 强化学习的最终目标是使得到的奖励最大化. 强化学习模型知识产权也很容易受到侵害[113]. Behzadan 等人[114] 提出一种强化学习策略水印方案. 该方案可以将唯一标识符以特殊的响应形式集成到策略中, 同时对模型性能的影响很小. 但现实情况下文献 [114] 所要求的将分布外状态和额外环境动作的顺序模式嵌入到目标强化学习模型中的需求难以满足. 此外, 在不同的环境下部署强化学习模型很容易地被攻击者识别, 继而可以简单地伪造预测结果致使验证过程失效. 针对该问题, 文献 [115] 采用与原始模型相同环境下的状态序列和动作概率分布, 设计了一种通用的强化学习水印方法, 可以应用于具有确定性或随机强化学习算法的各种强化学习任务. 大量实验结果表明, 该方法可以有效地保留强化学习模型的功能, 并对常见的模型修改(如模型微调和压缩)表现出显著的鲁棒性. 为使得水印的嵌入与验证更加便捷, 并获得良好的水印性能, 文献 [116] 提出一种新的强化模型水印框架 DrlWF. 该框架通过在训练状态中添加水印, 使用带有水印的状态训练模型, 从而将水印嵌入至模型中. 水印嵌入操作可以通过将水印嵌入到少量的训练数据中(仅需 0.025% 的训练数据)和不影响性能的奖励修改来实现.

联邦学习[117] 是一种多个参与方可以在不交互数据的情况下, 仅通过安全机制交互模型参数即可达到协同训练效果的分布式机器学习方法. Li 等人[118] 将文献 [15, 22] 的方法应用于联邦学习中, 设计了一种联邦学习 DNN 模型水印方案. 该方案中, 每个客户端独立验证模型水印的存在性, 并声明各自对联邦模型的版权, 而无需公开其私有训练数据或私有水印信息. 通过对多个客户可以私自嵌入和检测水印的条件进行严格分析, 文献 [118] 从理论上证明了嵌入水印的有效性. 为进一步提升联邦模型水印的可用性, 文献 [119] 允许参与训练的各用户在其本地模型中分别嵌入私有水印, 并通过云端的模型聚合操作将私有后门水印映射到全局模型, 进而作为联邦学习模型的全局水印. 作为一种隐私保护的分布式学习方法, 联邦学习容易受到参与方的投毒攻击, 尤其是后门投毒攻击, 这种攻击隐蔽性高、防御难度大. 文献 [120] 提出一种基于模型水印的联邦学习后门攻击防御方法. 在该方法中, 服务器预先在初始全局模型中嵌入水印, 通过验证该水印是否在参与方生成的本地模型中被破坏来检测恶意参与方. 在模型聚合阶段, 恶意参与方的本地模型将被丢弃, 从而提高全局模型的鲁棒性. 为保护联邦学习模型的完整性, MSFL[121] 框架引入多个混淆器(shufflers), 在保护个人隐私的同时, 提供对数据投毒攻击的有效防御. 该框架包含了层次化混淆机制、恶意评估模块以及复合防御策略 3 个模块, 通过用户混淆和模型混淆增强隐私保护, 动态评估参与者的恶意行为, 以识别并排除恶意客户端. 它结合了动态恶意度量和累积对手排除策略, 以保持模型的完整性.

除上述领域, 模型水印也应用于图神经网络[122]、预编/解码器[123]、集成学习[124] 上. 有关 DNN 模型水印具体应用的总结如表 4 所示.

4 未来研究领域

作为一个前沿研究领域, DNN 模型 IP 保护仍处于起步阶段, 现有的相关方法尚无法很好地应用于实际场景中. 在未来研究过程中, 有 4 个方向值得探索:

1)DNN 模型水印评价指标体系

目前, 针对 DNN 模型水印算法的评估指标较少, 普遍仅采用算法对模型微调和剪枝攻击的鲁棒性, 尚缺乏公认、全面、统一的评估框架和指标体系. 具体表现为, 对于选择不同宿主模型和不同嵌入方式的水印算法, 缺乏统一的评估标准. 例如, 如何衡量水印的不可察觉性、水印抵御攻击的能力等. 未来的研究方向之一是建立全面、统一的评估框架和指标体系, 以确保对于不同类型的水印算法都能进行公正和全面的评估. 这种统一评估框架和指标的建立对于 DNN 模型水印技术的发展至关重要. 它将有助

Table 4 Summary of Specific Applications of Watermarking

表 4 水印具体应用总结

方法	应用领域	具体应用	技术特征
文献 [95]	计算机视觉	图像处理	将水印信息嵌入到目标模型输出图像的 YCbCr 色彩空间的色度通道中
文献 [97]		视频弹幕	通过在循环神经网络隐藏状态中嵌入密钥以保护模型版权
文献 [98]		图像生成	使用正则化技术在生成对抗网络中嵌入独特的水印信息
文献 [100]	自然语言处理	文本生成	生成文本中进行词汇层面的修改来嵌入水印，并利用零假设检验作为事后所有权验证手段
文献 [102]		大语言模型	选择性地增加绿名单词汇的使用概率，并在生成文本时调整词汇的对数几率来实现软水印
文献 [104]		词向量生成	选择一组中等频率的词汇作为触发集，并在包含这些触发词的文本嵌入中插入目标词向量生成模型
文献 [105]	硬件	嵌入式	通过在训练集上添加基于签名的消息标记，从而在嵌入式系统中实现了对 DNN 模型的知识产权保护
文献 [106]		加速器	在硬件加速器的功能块中嵌入水印；这些水印可以在合法所有者模型和对应的密钥样本下被揭示
文献 [109]	其他	声纹识别	在输入音频的每个时间帧上添加噪声，并用隐藏作者信息的隐写术标记触发音频来生成触发音频
文献 [111]		数据库	通过对数据集进行特殊处理，使得在该数据集上训练的模型能够正确分类一些“困难”样本，而这些样本在未修改的数据集上训练的模型会被错误分类
文献 [116]		强化学习	选择特定训练状态作为水印载体，修改对应动作和奖励值，使 DRL 模型在特定条件下执行预定动作
文献 [118]		联邦学习	将文献 [15] 和文献 [22] 的方法应用于联邦学习中

于研究者更好地比较和理解不同水印算法之间的性能差异，提高评估的客观性和可重复性. 此外，由于宿主模型和嵌入方式的多样性，建立一个统一的评估体系也将有助于更全面了解水印算法在各种应用场景下的表现，推动该领域的标准化和发展. 评价指标体系构建过程中需要综合考虑水印算法的鲁棒性、水印的隐蔽性和容量、水印技术的应用范围和安全性，还应确保水印技术符合知识产权保护的法律要求.

2)可证明安全与可公开的模型水印

当前 DNN 模型水印验证方法面临 2 个主要问题，即私有验证的限制以及安全性难以通过理论证明. 在未来的研究中需要着力解决这 2 个问题. 首先，私有验证的限制对水印技术的可靠性和有效性造成了影响. 由于水印无法进行公开验证，严重影响了水印方法在实际应用中的信任度. 为解决该问题，未来的研究需要探索更加开放和透明的验证方法. 可尝试将水印技术与多方安全计算、零知识证明等密码学方法[125-126]结合，以确保验证过程更加公开和可信. 其次，由于 DNN 模型的复杂性和不可解释性，很难对水印技术的安全性进行全面的理论证明. 未来的研究中可以借鉴密码学的相关方法，提供更严谨的安全性证明[47]. 密码学的理论框架可以帮助确保水印技术的抗攻击性，从而提升模型水印方法的鲁棒性与安全性. 与此同时，在未来的研究中关注开放、透明、可证明的水印验证方法以及其与密码学方法的结合，将有助于提高水印技术的实用性和可信度.

3)支持主动授权的水印方法

目前的 DNN 模型水印方法大多采用基于触发集或模型特征构建的被动验证方法，这些方法通常在模型遭受窃取或滥用后才会触发，用于检测未经授权的使用. 尽管它们在模型被盗取后提供了一定程度的版权保护，但通常难以限制和影响被盗模型的性能，恶意用户仍能够正常访问或二次销售. 相比较而言，主动授权方法[74]能够在模型用于推理之前进行验证，有效地防止未经授权的模型使用. 这种预防性的措施能更好地保护模型的知识产权和完整性，避免模型的非法使用. 因此，强化对主动授权方法的研究对于推动 DNN 模型水印技术的发展和应用具有十分重要的意义. 主动授权方法为模型拥有者提供了更多的主动权，使其能够在模型被滥用之前采取有效的防范措施. 通过进一步研究和发展主动授权方法，可以更好地保护模型的知识产权，提高模型的安全性，有助于创造一个更为安全和可信的模型应用环境，以应对不断增长的模型安全性挑战.

4)模型水印的应用成本

随着对 DNN 模型水印技术关注度的增加，成本问题也将成为焦点. 未来的研究和应用中，在专注于水印技术的效果和可靠性的同时，研究者和从业者需要更加全面地考虑水印的嵌入、验证和维护过程中涉及的各类成本. 以基于后门的 DNN 模型水印方法为例，考虑到商业 API 通常每秒需要提供上百万次查询服务，必然极大地增加应用成本. 成本对于技术的可行性、可持续性和实际应用的成功具有重要影响. 未来的研究和应用中，必须更加注重 DNN 模型水印的成本问题，在设计水印方案时充分考虑相关的资源、时间和硬件成本. 在实际应用中，必须权

衡水印引入所带来的效益和潜在的成本压力. 此外，需要不断设计计算和存储开销更低、算法效率更高的新的 DNN 模型水印方法和技术，使其更具商业可行性. 对于 DNN 模型水印的成本问题的深入研究将为技术的实际推广和商业应用提供更为坚实的基础.

5 总　　结

本文回顾和介绍了近年来涌现的用于 DNN 模型版权声明和完整性验证的模型水印技术，对现有的研究方法进行了分类和深入讨论. 具体地，将这些方法划分为 2 种主要类型：一类是用于模型版权声明的鲁棒模型水印；另一类是用于验证模型完整性的脆弱模型水印. 进一步地，本文依据具体的实现技术将鲁棒模型水印方法细分为基于参数、基于后门以及基于特征的 3 种子类型. 类似地，将脆弱模型水印方法细分为基于数据加密、基于敏感样本以及可逆水印这 3 种子类型. 在此基础上，对每种方法的特点和优劣进行详细介绍和对比分析. 最后，指出 DNN 模型知识产权保护未来的研究领域. 通过对相关工作的细化分类和系统总结，本文为 DNN 模型水印领域的研究者提供了一份较为全面的参考，有助于推动相关领域的发展和创新.

作者贡献声明：金彪、熊金波负责研究方案的构思和设计、论文组织结构设计以及论文修订；林翔、李璇负责资料收集、文献归纳与整理以及初稿撰写；尤玮婧、姚志强负责论文审阅及修订.

参　考　文　献

[1] Inspur information. Inspur Electronic Information Industry Co., Ltd. Evaluation Report on the Development of China's AI Computing Power (2022-2023) [R/OL]. [2023-11-15]. https://www.inspur.com/lcjtww/resource/cms/article/2448319/2734787/2022122601.pdf
(浪潮信息: 浪潮电子信息产业股份有限公司 . 2022−2023 中国人工智能计算力发展评估报告[R/OL]. [2023-11-15]. http://www.inspur.com/lcjtww/resource/cms/article/2448319/2734787/2022122601.pdf)

[2] Baştanlar Y, Özuysal M. Introduction to machine learning[J]. miRNomics: MicroRNA Biology and Computational Analysis, 2014: 105−128. https://link.springer.com/protocol/10.1007/978-1-62703-748-8_7#citeas

[3] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J/OL]. [2023-11-15]. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[4] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9

[5] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020(33): 1877−1901

[6] Cecil R R, Soares J. IBM Watson Studio: A Platform to Transform Data to Intelligence[M]//Pharmaceutical Supply Chains-Medicines Shortages. Cham: Springer, 2019: 183−192

[7] Tramèr F, Zhang Fang, Juels A, et al. Stealing machine learning models via prediction {APIs}[C]//Proc of the 25th USENIX Security Symp (USENIX Security 16). Berkeley, CA: USENIX Association, 2016: 601−618

[8] Li Fenghua, Li Hui, Niu Ben, et al. Research category and future development trend of data elements circulation and security[J]. Journal on Communications, 2024, 45(5): 1−11 (in Chinese)
(李凤华, 李晖, 牛犇, 等. 数据要素流通与安全的研究范畴与未来发展趋势[J]. 通信学报, 2024, 45(5): 1−11)

[9] Wang Baonan, Hu Feng, Zhang Huanguo, et al. From evolutionary cryptography to quantum artificial intelligent cryptography[J]. Journal of Computer Research and Development, 2019, 56(10): 2112−2134 (in Chinese)
(王宝楠, 胡风, 张焕国, 等. 从演化密码到量子人工智能密码综述[J]. 计算机研究与发展, 2019, 56(10): 2112−2134)

[10] Van Schyndel R G, Tirkel A Z, Osborne C F. A digital watermark[C]//Proc of 1st Int Conf on Image Processing. Piscataway, NJ: IEEE, 1994: 86−90

[11] Zhu M, Gupta S. To prune, or not to prune: Exploring the efficacy of pruning for model compression[J]. arXiv preprint, arXiv: 1710.01878, 2017

[12] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint, arXiv: 1503.02531, 2015

[13] Jiang Han, Xu Qiuliang. Secure multiparty computation in cloud computing[J]. Journal of Computer Research and Development, 2016, 53(10): 2152−2162 (in Chinese)
(蒋瀚, 徐秋亮. 基于云计算服务的安全多方计算[J]. 计算机研究与发展, 2016, 53(10): 2152−2162)

[14] Fan Kefeng, Mo Wei, Cao Shan, et al. Advances in digital rights management technology and application[J]. Acta Electonica Sinica, 2007, 35(6): 1139−1147 (in Chinese)
(范科峰, 莫玮, 曹山, 等. 数字版权管理技术及应用研究进展[J]. 电子学报, 2007, 35(6): 1139−1147)

[15] Uchida Y, Nagai Y, Sakazawa S, et al. Embedding watermarks into deep neural networks[C]//Proc of the 2017 ACM on Int Conf on Multimedia Retrieval. New York: ACM, 2017: 269−277

[16] Zhang Yingjun, Chen Kai, Zhou Geng, et al. Research progress of neural networks watermarking technology[J]. Journal of Computer Research and Development, 2021, 58(5): 964−976 (in Chinese)
(张颖君, 陈恺, 周赓, 等. 神经网络水印技术研究进展[J]. 计算机研究与发展, 2021, 58(5): 964−976)

[17] Fan Xuefeng, Zhou Xiaoyi, Zhu Bingbing, et al. Survey of copyright protection schemes based on DNN model[J]. Journal of Computer

Research and Development, 2022, 59(5): 953−977 (in Chinese)
(樊雪峰, 周晓谊, 朱冰冰, 等. 深度神经网络模型版权保护方案综述[J]. 计算机研究与发展, 2022, 59(5): 953−977)

[18] Wu Hanzhou, Zhang Jie, Li Yue, et al. Overview of artificial intelligence model watermarking[J]. Journal of Image and Graphics, 2023, 28(6): 1792−1810 (in Chinese)
(吴汉舟, 张杰, 李越, 等. 人工智能模型水印研究进展[J]. 中国图象图形学报, 2023, 28(6): 1792−1810)

[19] Li Yue, Wang Hongxia, Barni M. A survey of deep neural network watermarking techniques[J]. Neurocomputing, 2021, 461: 171−193

[20] Lukas N, Jiang E, Li Xinda, et al. Sok: How robust is image classification deep neural network watermarking?[C]//Proc of 2022 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2022: 787−804

[21] Zhang Xinpeng, Wang Shuozhong. Fragile watermarking with error-free restoration capability[J]. IEEE Transactions on Multimedia, 2008, 10(8): 1490−1499

[22] Adi Y, Baum C, Cisse M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring[C]// Proc of the 27th USENIX Security Symp (USENIX Security 18). Berkeley, CA: USENIX Association, 2018: 1615−1631

[23] Chen Huili, Rouhani B D, Koushanfar F. BlackMarks: Blackbox multibit watermarking for deep neural networks[J]. arXiv preprint, arXiv: 1904.00344, 2019

[24] Wang Tianhao, Kerschbaum F. Attacks on digital watermarks for deep neural networks[C]//Proc of 2019 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP 2019). Piscataway, NJ: IEEE, 2019: 2622−2626

[25] Namba R, Sakuma J. Robust watermarking of neural network with exponential weighting[C]//Proc of the 2019 ACM Asia Conf on Computer and Communications Security. New York: ACM, 2019: 228−240

[26] Wang Bolun, Yao Yuanshun, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]// Proc of 2019 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 707−723

[27] Wang Shang, Li Xin, Song Yongli, et al. A trigger sample detection scheme based on custom backdoor behaviors[J]. Journal of Cyber Security, 2022, 7(6): 48−61 (in Chinese)
(王尚, 李昕, 宋永立, 等. 基于自定义后门的触发器样本检测方案[J]. 信息安全学报, 2022, 7(6): 48−61)

[28] Fan Lixin, Ng K W, Chan C S. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks[C]//Proc of the 33rd Int Conf on Neural Information Processing Systems. New York: Curran Associates, Inc, 2019: 4714−4723

[29] Chen Huili, Rouhani B D, Fu Cheng, et al. Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models[C]//Proc of the 2019 on Int Conf on Multimedia Retrieval. New York: ACM, 2019: 105−113

[30] Tajbakhsh N, Shin J Y, Gurudu S R, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning?[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1299−1312

[31] Darvish Rouhani B, Chen Huili, Koushanfar F. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks[C]//Proc of the 24th Int Conf on Architectural Support for Programming Languages and Operating Systems. New York: ACM, 2019: 485−497

[32] Wang Tianhao, Kerschbaum F. RIGA: Covert and robust white-box watermarking of deep neural networks[C]//Proc of the Web Conf 2021. New York: ACM, 2021: 993−1004

[33] Feng Le, Zhang Xinpeng. Watermarking neural network with compensation mechanism[C]//Proc of the 13th Int Conf on Knowledge Science, Engineering and Management (KSEM 2020). Berlin: Springer, 2020: 363−375

[34] Kuribayashi M, Tanaka T, Suzuki S, et al. White-box watermarking scheme for fully-connected layers in fine-tuning model[C]//Proc of the 2021 ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2021: 165−170

[35] Chen B, Wornell G W. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding[J]. IEEE Transactions on Information Theory, 2001, 47(4): 1423−1443

[36] Kuribayashi M, Yasui T, Malik A. White box watermarking for convolution layers in fine-tuning model using the constant weight code[J]. Journal of Imaging, 2023, 9(6): 117

[37] Lv Peizhou, Li Pan, Zhang Shengzhi, et al. A robustness-assured white-box watermark in neural networks[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(6): 5214−5229

[38] Yan Yifan, Pan Xudong, Zhang Mi, et al. Rethinking white-box watermarks on deep learning models under neural structural obfuscation[C]//Proc of the 32th USENIX Security Symp (USENIX Security 23). Berkeley, CA: USENIX Association, 2023: 2347−2364

[39] Liu Yingqi, Ma Shiqing, Aafer Y, et al. Trojaning attack on neural networks[C]//Proc of the 25th Annual Network And Distributed System Security Symp (NDSS 2018). Berkeley, CA: Internet Soc, 2018: 1−15

[40] Zhang Jialong, Gu Zhongshu, Jang Jiyong, et al. Protecting intellectual property of deep neural networks with watermarking[C]// Proc of the 2018 on Asia Conf on Computer and Communications Security. New York: ACM, 2018: 159−172

[41] 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

[42] Li Zheng, Hu Chengyu, Zhang Yang, et al. How to prove your model belongs to you: A blind-watermark based framework to protect intellectual property of DNN[C]//Proc of the 35th Annual Computer Security Applications Conf. New York: ACM, 2019: 126−137

[43] Liu Yong, Wu Hanzhou, Zhang Xinpeng. Robust and imperceptible black-box DNN watermarking based on Fourier perturbation analysis and frequency sensitivity clustering[J]. IEEE Transactions on Dependable and Secure Computing, 2024: 1−14

[44] Lounici S, Njeh M, Ermis O, et al. Yes we can: Watermarking

machine learning models beyond classification[C]//Proc of 2021 IEEE 34th Computer Security Foundations Symp (CSF). Piscataway, NJ: IEEE, 2021: 1−14

[45] Li Li, Zhang Weiming, Barni M. Universal BlackMarks: Key-image-free blackbox multi-bit watermarking of deep neural networks[J]. IEEE Signal Processing Letters, 2023, 30: 36−40

[46] Jia Hengrui, Choquette-Choo C A, Chandrasekaran V, et al. Entangled watermarks as a defense against model extraction[C]// Proc of the 30th USENIX Security Symp (USENIX Security 21). Berkeley, CA: USENIX Association, 2021: 1937−1954

[47] Bansal A, Chiang P, Curry M J, et al. Certified neural network watermarks with randomized smoothing[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2022: 1450−1465

[48] Hua Guang, Teoh A B J, Xiang Yong, et al. Unambiguous and high-fidelity backdoor watermarking for deep neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(8): 11204−11217

[49] Liu Weifa, Zhang Guanghua, Yang Ting, et al. Logo network based deep learning multi-model watermarking scheme[J]. Journal of Cyber Security, 2022, 7(6): 105−115 (in Chinese)
(刘伟发, 张光华, 杨婷, 等. 基于标志网络的深度学习多模型水印方案[J]. 信息安全学报, 2022, 7(6): 105−115)

[50] Li Wei, Zhang Xiaoyu, Lin Shen, et al. Chameleon DNN watermarking: Dynamically public model ownership verification[C]// Proc of the Int Conf on Information Security Applications. Berlin: Springer, 2022: 344−356

[51] Chen Xiaofeng, Zhang Fangguo, Kim K. Chameleon hashing without key exposure[C]//Proc of the Int Conf on Information Security. Berlin: Springer, 2004: 87−98

[52] Li Yiming, Bai Yang, Jiang Yong, et al. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection[J]. Advances in Neural Information Processing Systems, 2022, 35: 13238−13250

[53] Lee S, Song W, Jana S, et al. Evaluating the robustness of trigger set-based watermarks embedded in deep neural networks[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(4): 3434−3448

[54] Cao Xiaoyu, Jia Jinyuan, Gong N Z. IPGuard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary[C]//Proc of the 2021 ACM Asia Conf on Computer and Communications Security. New York: ACM, 2021: 14−25

[55] Zhao Jingjing, Hu Qingyue, Liu Gaoyang, et al. AFA: Adversarial fingerprinting authentication for deep neural networks[J]. Computer Communications, 2020, 150: 488−497

[56] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929−1958

[57] Wang Si, Chang C H. Fingerprinting deep neural networks—a deepfool approach[C]//Proc of 2021 IEEE Int Symp on Circuits and Systems (ISCAS). Piscataway, NJ: IEEE, 2021: 1−5

[58] Lukas N, Zhang Yuxuan, Kerschbaum F. Deep neural network fingerprinting by conferrable adversarial examples[J]. arXiv preprint, arXiv: 1912.00888, 2019

[59] Wang Siyue, Wang Xiao, Chen Pinyu, et al. Characteristic examples: High-robustness, low-transferability fingerprinting of neural networks[C]//Proc of the 30th Int Joint Conf on Artificial Intelligence. Montreal, Canada: IJCAI, 2021: 575−582

[60] Peng Zirui, Li Shaofeng, Chen Guoxing, et al. Fingerprinting deep neural networks globally via universal adversarial perturbations[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 13430−13439

[61] Chen Jialuo, Wang Jingyi, Peng Tinglan, et al. Copy, right? A testing framework for copyright protection of deep learning models[C]// Proc of 2022 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2022: 824−841

[62] Maini P, Yaghini M, Papernot N. Dataset inference: Ownership resolution in machine learning[J]. arXiv preprint, arXiv: 2104.10706, 2021

[63] Liu Yunpeng, Li Kexin, Liu Zhuotao, et al. Provenance of training without training data: Towards privacy-preserving DNN model ownership verification[C]//Proc of the ACM Web Conf 2023. New York: ACM, 2023: 1980−1990

[64] Dong Tian, Li Shaofeng, Chen Guoxing, et al. RAI2: Responsible identity audit governing the artificial intelligence[C]//Proc of the 30th Annual Network and Distributed System Security Symp (NDSS 2023). San Diego, California, USA: Internet Soc, 2023: 1−18

[65] Lv Peizhuo, Ma Hualong, Chen Kai, et al. MEA-Defender: A robust watermark against model extraction attack[C]//Proc of 2024 IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2024: 102−102

[66] Fan Lixin, Ng K W, Chan C S. DeepIPR: Deep neural network intellectual property protection with passports[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2021, 44(10): 6122−6139

[67] Zhang Jie, Chen Dongdong, Liao Jing, et al. Passport-aware normalization for deep model protection[J]. Advances in Neural Information Processing Systems, 2020, 33: 22619−22628

[68] Liu Hanwen, Weng Zhenyu, Zhu Yuesheng, et al. Trapdoor normalization with irreversible ownership verification[C]//Proc of the Int Conf on Machine Learning. Honolulu, Hawaii, USA: PMLR, 2023: 22177−22187

[69] Chen Yiming, Tian Jinyu, Chen Xiangyu, et al. Effective ambiguity attack against passport-based DNN intellectual property protection schemes through fully connected layer substitution[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 8123−8132

[70] Gomez L, Wilhelm M, Márquez J, et al. Security for distributed deep neural networks towards data confidentiality & intellectual property protection[J]. arXiv preprint, arXiv: 1907.04246, 2019

[71] Brakerski Z, Gentry C, Vaikuntanathan V. (Leveled) fully homomorphic encryption without bootstrapping[J]. ACM Transactions on Computation Theory, 2014, 6(3): 1−36

[72] Lin Ning, Chen Xiaoming, Lu Hang, et al. Chaotic weights: A novel approach to protect intellectual property of deep neural networks[J].

IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020, 40(7): 1327−1339

[73] Peterson G. Arnold's cat map[J]. Math Linear Algebra, 1997, 45: 1−7

[74] Xue Mingfu, Wu Zhiyu, Zhang Yushu, et al. AdvParams: An active DNN intellectual property protection technique via adversarial perturbation based parameter encryption[J]. IEEE Transactions on Emerging Topics in Computing, 2023, 11(3): 664−678

[75] Zhou Tong, Luo Yukui, Ren Shaolei, et al. NNSplitter: An active defense solution to DNN model via automated weight obfuscation[C]//Proc of Int Conf on Machine Learning. Honolulu, Hawaii, USA: PMLR, 2023: 42614−42624

[76] Pyone A, Maung M, Kiya H. Training DNN model with secret key for model protection[C]//Proc of 2020 IEEE the 9th Global Conf on Consumer Electronics (GCCE). Piscataway, NJ: IEEE, 2020: 818−821

[77] He Zecheng, Zhang Tianwei, Lee R. Sensitive-sample fingerprinting of deep neural networks[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4729−4737

[78] Xu Guowen, Li Hongwei, Ren Hao, et al. Secure and verifiable inference in deep neural networks[C]//Proc of Annual Computer Security Applications Conf. New York: ACM, 2020: 784−797

[79] Aramoon O, Chen Pinyu, Qu Gang. AID: Attesting the integrity of deep neural networks[C]//Proc of the 58th ACM/IEEE Design Automation Conf. Piscataway, NJ: IEEE, 2021: 19−24

[80] Kuttichira D P, Gupta S, Nguyen D, et al. Verification of integrity of deployed deep learning models using Bayesian optimization[J]. Knowledge-Based Systems, 2022, 241: 108238

[81] Zhu Renjie, Wei Ping, Li Sheng, et al. Fragile neural network watermarking with trigger image set[C]//Proc of the 14th Int Conf on Knowledge Science, Engineering and Management. Berlin: Springer, 2021: 280−293

[82] Lao Yingjie, Zhao Weijie, Yang Peng, et al. Deepauth: A DNN authentication framework by model-unique and fragile signature embedding[C]//Proc of the AAAI Conf on Artificial Intelligence, Palo Alto, CA: AAAI, 2022: 9595−9603

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

[84] Yin Zhaoxia, Yin Heng, Su Hang, et al. Decision-based iterative fragile watermarking for model integrity verification[J]. arXiv preprint, arXiv: 2305.09684, 2023

[85] Lin Xiang, Jin Biao, You Weijing, et al. Model integrity verification framework of deep neural network based on fragile fingerprint [J/OL]. Journal of Computer Applications . [2024-05-03]. http://kns.cnki.net/kcms/detail/51.1307.TP.20240118.1646.002.html (in Chinese)
(林翔, 金彪, 尤玮婧, 等. 基于脆弱指纹的深度神经网络模型完整性验证框架[J/OL]. 计算机应用. [2024-05-03]. http://kns.cnki.net/kcms/detail/51.1307.TP.20240118.1646.002.html)

[86] Qian Yaguan, He Niannian, Guo Yankai, et al. An evasion algorithm to fool fingerprint detector for deep neural networks[J]. Journal of Computer Research and Development, 2021, 58(5): 1106−1117 (in Chinese)
(钱亚冠, 何念念, 郭艳凯, 等. 针对深度神经网络模型指纹检测的逃避算法[J]. 计算机研究与发展, 2021, 58(5): 1106−1117)

[87] Wang Shuo, Abuadbba S, Agarwal S, et al. PublicCheck: Public integrity verification for services of run-time deep models[C]//Proc of 2023 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2023: 1348−1365

[88] Abuadbba A, Rhodes N, Moore K, et al. DeepiSign-G: Generic watermark to stamp hidden DNN parameters for self-contained tracking[J]. arXiv preprint, arXiv: 2407.01260, 2024.

[89] Guan Xiquan, Feng Huamin, Zhang Weiming, et al. Reversible watermarking in deep convolutional neural networks for integrity authentication[C]//Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 2273−2280

[90] Zhao Gejian, Qin Chuan, Yao Heng, et al. DNN self-embedding watermarking: Towards tampering detection and parameter recovery for deep neural network[J]. Pattern Recognition Letters, 2022, 164: 16−22

[91] Huang Yawen, Zheng Hongying, Xiao Di. Convolutional neural networks tamper detection and location based on fragile watermarking[J]. Applied Intelligence, 2023, 53: 24056−24067

[92] Lin Mingbao, Ji Rongrong, Wang Yan, et al. Hrank: Filter pruning using high-rank feature map[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 1529−1538

[93] Lawnik M. Combined logistic and tent map[J]. Journal of Physics: Conference Series, 2018, 1141: 012132

[94] Quan Yuhui, Teng Huan, Chen Yixin, et al. Watermarking deep neural networks in image processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(5): 1852−1865

[95] Hou Minghua, Tang Linlin, Qi Shuhan, et al. A robust watermarking method for image processing models[C]//Proc of 2022 4th Int Conf on Data Intelligence and Security (ICDIS). Piscataway, NJ: IEEE, 2022: 75−81

[96] Zhang Jie, Chen Dongdong, Liao Jing, et al. Model watermarking for image processing networks[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 34(7): 12805−12812

[97] Lim J H, Chan C S, Ng K W, et al. Protect, show, attend and tell: Empowering image captioning models with ownership protection[J]. Pattern Recognition, 2022, 122: 108285

[98] Ong D S, Chan C S, Ng K W, et al. Protecting intellectual property of generative adversarial networks from ambiguity attacks[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 3630−3639

[99] Qiao Tong, Ma Yuyan, Zheng Ning, et al. A novel model watermarking for protecting generative adversarial network[J]. Computers & Security, 2023, 127: 103102

[100] He Xuanli, Xu Qiongkai, Lyu Lingjuan, et al. Protecting intellectual property of language generation APIs with lexical watermark[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA:

AAAI, 2022: 10758−10766

[101] Dai Long, Zhang Jing, Fan Xuefeng, et al. NLP neural network copyright protection based on black box watermark[J]. Chinese Journal of Network and Information Security, 2023, 9(1): 140−149 (in Chinese)
(代龙, 张静, 樊雪峰, 等. 基于黑盒水印的 NLP 神经网络版权保护[J]. 网络与信息安全学报, 2023, 9(1): 140−149)

[102] Kirchenbauer J, Geiping J, Wen Yuxin, et al. A watermark for large language models[C]//Proc of Int Conf on Machine Learning. Honolulu, Hawaii, USA: PMLR, 2023: 17061−17084

[103] Chiang W L, Li Zhuohan, Lin Zi, et al. Vicuna: An open-source chatbot impressing GPT−4 with 90%* ChatGPT quality[OL]. [2023-10-25]. https://lmsys.org/blog/2023-03-30-vicuna/

[104] Peng Wenjun, Yi Jingwei, Wu Fangzhao, et al. Are you copying my model? Protecting the copyright of large language models for EaaS via backdoor watermark[J]. arXiv preprint, arXiv: 2305.10036, 2023

[105] Guo Jia, Potkonjak M. Watermarking deep neural networks for embedded systems[C]//Proc of 2018 IEEE/ACM Int Conf on Computer-Aided Design (ICCAD). Piscataway, NJ: IEEE, 2018: 1−8

[106] Clements J, Lao Yingjie. DeepHardMark: Towards watermarking neural network hardware[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 4450−4458

[107] Bulens P, Standaert F X, Quisquater J J. How to stronglylink data and its medium: The paper case[J]. IET Information Security, 2010, 4(3): 125−136

[108] Mankali L, Rangarajan N, Chatterjee S, et al. Leveraging ferroelectric stochasticity and in-memory computing for DNN IP obfuscation[J]. IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, 2022, 8(2): 102−110

[109] Chen Haozhe, Zhang Weiming, Liu Kunlin, et al. Speech pattern based black-box model watermarking for automatic speech recognition[C]//Proc of 2022 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP 2022). Piscataway, NJ: IEEE, 2022: 3059−3063

[110] Chen Haozhe, Zhang Jie, Chen Kejiang, et al. Model access control based on hidden adversarial examples for automatic speech recognition[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(3): 1302−1315

[111] Guo Junfeng, Li Yiming, Wang Lixu, et al. Domain watermark: Effective and harmless dataset copyright protection is closed at hand[J]. Advances in Neural Information Processing Systems, 2024, 36: 54421−54450

[112] Liu Tengjun, Chen Ying, Gu Wanxuan. Copyright-certified distillation dataset: Distilling one million coins into one bitcoin with your private key[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023, 37(5): 6458−6466

[113] Chen Kangjie, Guo Shangwei, Zhang Tianwei, et al. Stealing deep reinforcement learning models for fun and profit[C]//Proc of the 2021 ACM Asia Conf on Computer and Communications Security. New York: ACM, 2021: 307−319

[114] Behzadan V, Hsu W. Sequential triggers for watermarking of deep reinforcement learning policies[J]. arXiv preprint, arXiv: 1906.01126, 2019

[115] Chen Kangjie, Guo Shangwei, Zhang Tianwei, et al. Temporal watermarks for deep reinforcement learning models[C]//Proc of the 20th Int Conf on Autonomous Agents and Multiagent Systems. Virtual Event, United Kingdom, 2021: 314−322

[116] Chen Yulin, Yao Zhiqiang, Jin Biao, et al. A deep reinforcement learning watermarking framework based on backdoor technology[J]. Journal of Fujian Normal University (Natural Science Edition), 2024, 40(1): 96−105 (in Chinese)
(陈瑜霖, 姚志强, 金彪, 等. 一种基于后门技术的深度强化学习水印框架[J]. 福建师范大学学报: 自然科学版, 2024, 40(1): 96−105)

[117] Konečný J, McMahan H B, Yu F X, et al. Federated learning: Strategies for improving communication efficiency[J]. arXiv preprint, arXiv: 1610.05492, 2016

[118] Li Bowen, Fan Lixin, Gu Hanlin, et al. FedIPR: Ownership verification for federated deep neural network models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4521−4536

[119] Li Xuan, Deng Tianpeng, Xiong Jinbo, et al. Federated learning watermark based on backdoor[J]. Journal of Software, 2024, 35(7): 3454−3468 (in Chinese)
(李璇, 邓天鹏, 熊金波, 等. 基于模型后门的联邦学习水印[J]. 软件学报, 2024, 35(7): 3454−3468)

[120] Guo Jingjing, Liu Jiuzun, Ma Yong, et al. Backdoor attack defense method for federated learning based on model watermarking[J]. Chinese Journal of Computers, 2024, 47(3): 662−676 (in Chinese)
(郭晶晶, 刘玖樽, 马勇, 等. 基于模型水印的联邦学习后门攻击防御方法[J]. 计算机学报, 2024, 47(3): 662−676)

[121] Zhou Zan, Xu Changqiao, Wang Mingze, et al. A multi-shuffler framework to establish mutual confidence for secure federated learning[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(5): 4230−4244

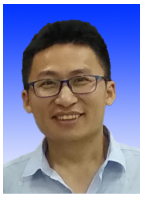
[122] Shen Yun, He Xinle, Han Yufei, et al. Model stealing attacks against inductive graph neural networks[C]//Proc of 2022 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2022: 1175−1192

[123] Dziedzic A, Boenisch F, Jiang Mingjian, et al. Sentence embedding encoders are easy to steal but hard to defend[C]//Proc of ICLR 2023 Workshop on Pitfalls of Limited Data and Computation for Trustworthy ML. Kigali, Rwanda: ICLR, 2023: 1−12

[124] Lin L, Wu Hanzhou. Verifying integrity of deep ensemble models by lossless black-box watermarking with sensitive samples[C]//Proc of the 10th Int Symp on Digital Forensics and Security (ISDFS). Piscataway, NJ: IEEE, 2022: 1−6

[125] Feng Shuai, Deng Lunzhi. Identity-based data auditing scheme with privacy protection[J]. Journal of Guizhou Normal University (Natural Sciences), 2023, 41(2): 105−112 (in Chinese)
(冯帅, 邓伦治. 基于身份的隐私保护数据审计方案[J]. 贵州师范大学学报: 自然科学版, 2023, 41(2): 105−112)

[126] Xiong Hu, Lin Ye, Yao Ting. SM9 identity-based encryption scheme with equality test and cryptographic reverse firewalls[J]. Journal of Computer Research and Development, 2024, 61(4): 1070−1084 (in Chinese)
(熊虎, 林烨, 姚婷. 支持等式测试及密码逆向防火墙的 SM9 标识加密方案[J]. 计算机研究与发展, 2024, 61(4): 1070−1084)



Jin Biao, born in 1985. PhD, associate professor, master supervisor. Member of CCF. His main research interests include privacy preserving and machine learning.

金 彪, 1985 年生. 博士, 副教授, 硕士生导师. CCF 会员. 主要研究方向为隐私保护、机器学习.



Lin Xiang, born in 1996. Master. His main research interest includes artificial intelligence security.

林 翔, 1996 年生. 硕士. 主要研究方向为人工智能安全.



Xiong Jinbo, born in 1981. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include data security and privacy preserving, and artificial intelligence security.

熊金波, 1981 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为数据安全与隐私保护、人工智能安全.



You Weijing, born in 1994. PhD, associate professor, master supervisor. Member of CCF. Her main research interests include data security, cloud storage security, and artificial intelligence security.

尤玮婧, 1994 年生. 博士, 副教授, 硕士生导师. CCF 会员. 主要研究方向为数据安全、云存储安全、人工智能安全.



Li Xuan, born in 1984. PhD, associate professor, master supervisor. Member of CCF. Her main research interests include data security, cloud computing security, and privacy preserving.

李 璇, 1984 年生. 博士, 副教授, 硕士生导师. CCF 会员. 主要研究方向为数据安全、云计算安全、隐私保护.



Yao Zhiqiang, born in 1967. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include application security, big data security, and privacy preserving.

姚志强, 1967 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为应用安全、大数据安全、隐私保护.