

深度学习模型产权保护技术综述

张光华¹, 刘伟发^{1,2}, 刘亦纯¹, 李 曼¹, 贾刘影¹, 王向红¹, 张玉清³

¹ 河北科技大学信息科学与工程学院 石家庄 中国 050018

² 菏泽学院数学与统计学院 菏泽 中国 274015

³ 国家计算机网络入侵防范中心(中国科学院大学) 北京 中国 101408

摘要 深度学习(Deep learning, DL)模型以其卓越的性能在各个领域得到了广泛应用, 其蕴含的巨大商业价值使得保护 DL 模型产权十分重要。同数字产品一样, DL 模型也面临着被盗版的隐患。一旦出现盗版, 将严重侵害所有者的权益。为此, 众多学者针对不同的算法模型提出了自己的保护方案, 但目前这些方案尚缺乏全面梳理。本文调研自 2017 年起至今与该领域相关的期刊和会议, 在原有 DL 模型水印的基础上扩充了 DL 模型指纹、DL 模型锁的相关研究, 根据方案所属的领域和任务、具体实现思路和技术进行了详细的归纳总结。之后, 针对这些保护方案的攻击面, 根据攻击程度划分为检测攻击、干扰攻击、移除攻击以进行系统归类。再后, 总结了众多保护方案的常用性能指标。最后, 讨论了 DL 模型产权保护目前的不足和未来研究的方向, 旨在为推动 DL 模型产权保护的深入研究提供借鉴。

关键词 深度学习; 产权保护; 模型水印; 模型指纹; 模型锁; 攻击面

中图法分类号 TP183; TP309.7 DOI 号 10.19363/J.cnki.cn10-1380/tn.2026.05.10

Overview of Deep Learning Model Property Protection Technology

ZHANG Guanghua¹, LIU Weifa^{1,2}, LIU Yichun¹, LI Man¹, JIA Liuying¹, WANG Xianghong¹,
ZHANG Yuqing³

¹ School of Information Science and Engineering, Hebei University of Science Technology, Shijiazhuang 050018, China

² School of Mathematics and Statistics, Heze University, Heze 274015, China

³ National Computer Network Intrusion Protection Center, University of Chinese Academy of Sciences, Beijing 101408, China

Abstract Deep Learning (DL) models have gained widespread application across various domains due to their exceptional performance, which has significant commercial implications. Consequently, the protection of intellectual property rights associated with DL models has become a crucial issue. Similar to digital products, DL models are vulnerable to piracy, and instances of such infringement can severely compromise the rights and interests of their owners. In response, numerous scholars have proposed protective strategies tailored to different algorithmic models; however, a comprehensive review of these proposals is currently lacking. This paper conducts an extensive survey of academic journals and conferences related to this field from 2017 to the present. Building upon existing research on watermarking DL models, it expands the discussion to include related studies on DL model fingerprints and locks. A detailed synthesis of these protective mechanisms is provided, categorized according to their respective domains, tasks, implementation approaches, and underlying technologies. Furthermore, the various protective schemes are assessed in terms of their vulnerability to different types of attacks. These attacks are systematically classified based on their severity into three categories: detection attacks, interference attacks, and removal attacks. This classification aims to create a structured understanding of the security landscape surrounding DL model protection. Additionally, the paper summarizes common performance metrics used to evaluate the effectiveness of these protective strategies. By doing so, it highlights the strengths and weaknesses of various approaches currently in use. Ultimately, the paper discusses the existing shortcomings in DL model intellectual property protection and outlines potential future research directions. This discussion is intended to provide a foundation for advancing the field of DL model protection and to inspire further exploration into effective safeguarding mechanisms. The aim is to contribute to a more robust framework for ensuring the integrity and ownership of DL models in an increasingly digital and competitive landscape.

Key words deep learning; intellectual property protection; model watermark; model fingerprint; model lock; attack surface

通讯作者: 张玉清, 博士, 教授, Email: zhangyq@ucas.ac.cn.

本课题得到国家自然科学基金项目(No. 62572258)、河北省高等学校科学研究计划产业重大专项(No. CYZD2026001)资助。

收稿日期: 2024-10-29; 修改日期: 2025-02-17; 定稿日期: 2026-04-07

1 引言

一个性能优良的 DL 模型能为企业带来巨大收益,但其成本非常高昂。同时,DL 面临着严重的安全性隐患。例如,攻击者可以通过投毒攻击,攻击训练数据集或算法以操纵模型的预测;通过模型提取攻击,对已部署的模型,窃取其模型参数及非法获取模型;通过模型逆向攻击,从模型信息中恢复数据集信息;通过对抗攻击,向模型提交对抗样本以损害模型性能。以上攻击手段使盗版 DNN 模型成为可能。

一旦 DL 模型被提取、剽窃或滥用,甚至出现盗版 DL 应用,所有者的利益将受到严重侵害。使用未授权的 DL 模型构建 DL 产品及服务以获取商业利益会构成侵权。同时,由于 DL 模型的训练需要大量数据,若这些数据是非授权的受版权保护数据,也将构成侵权,如“Stable Diffusion”侵权案。因此,保护 DL 模型产权是悬而未决的重要问题。近期,数字产权保护的水印、指纹、锁等技术被引入到 DL 模型的版权保护工作中,产生了许多优秀的方案。针对众多的 DL 模型产权保护方案,表 1 中的综述对它们进行了归纳。已有综述存在一定局限,例如,Chen 等人^[1]在综述产权保护的研究时,未分析大语言模型的最新章节版权保护研究。

为此,本文在已有综述基础上,对以往 DL 模型产权保护相关工作进行了全面详细的梳理,引入新的分类法来对水印进行划分,并补充了指纹和锁的产权保护方案,对 DL 模型产权保护方案的攻击面和性能指标进行了论述。此外,本文对 DL 模型产权保护面临的挑战和未来的工作提出了一些新的见解。

本文调研了从 2017 年起至今中国计算机协会推荐期刊和会议中发表的 DL 模型产权保护相关研究文献,其中 A 类 37 篇、B 类 19 篇、C 类 26 篇。按照实现机制划分了模型水印、模型指纹、模型锁。其中,水印指向 DL 模型中添加的额外标识,其通常可以实现所有权验证或标识用户;指纹指使用特定算法从 DL 模型中提取的固有标识,其通常可以实现所有权验证;锁指对 DL 模型进行访问控制验证的一系列机制。近五年此三个方向的发表论文统计如图 1 所示,可以看出 DL 模型版权保护论文发表数量逐年上升,该方向引起了广大学者的关注。

本文的主要贡献如下:

1) 从水印嵌入和验证的不同方式出发,总结了显式嵌入显式验证、隐式嵌入隐式验证、显式嵌入隐式验证、生成式验证等 4 种水印机制;从指纹刻画 DL 模型的不同特征出发,总结了刻画决策边界、刻

画模型构成等 2 种指纹机制;从锁的不同实施方式出发总结了基于输入预处理、权重加密、控制层、硬件支持等 4 种锁机制。

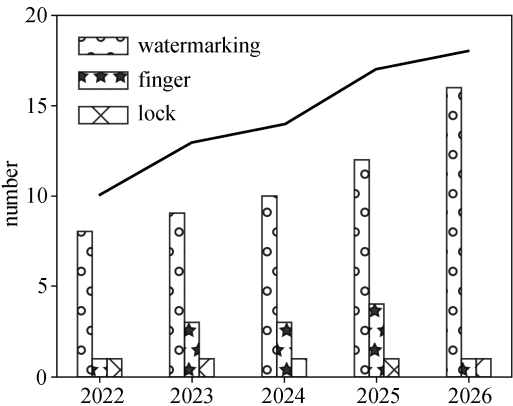


图 1 近五年论文发表情况
Figure 1 Publication Status of Papers in the Past Five Years

表 1 DL 模型产权保护相关综述对比
Table 1 Comparison of DL Model Property Rights Protection Reviews

相关文献	发表年份	优点	不足
Liu 等人 ^[1]	2026	从水印嵌入策略、对各类攻击的鲁棒性、对模型性能的影响三个维度分析现有水印方法	未梳理指纹、锁等形式的深度学习模型产权保护方案
Chen 等人 ^[2]	2024	从面向深度学习模型的水印和基于深度学习模型的水印两个角度论述产权保护技术	面向 DL 模型的水印与基于 DL 模型的水印之间的关系分析浅显,未调研面向大语言模型的水印研究
Liang 等人 ^[3]	2024	以文本水印、图像水印、音频水印、多模态水印、动态水印梳理深度学习模型水印,调研了大语言模型水印的进展	未调查面向深度学习模型的指纹、锁等形式的相关产权保护方案
樊雪峰等人 ^[4]	2022	根据水印嵌入和提取方式的不同方式,按白盒水印、黑盒水印、灰盒水印、无盒水印对 DL 模型版权保护机制进行综述	仅针对水印形式的 DL 模型产权保护方案进行了综述,未对指纹、锁等新机制进行讨论

2) 总结了对 DL 模型产权保护机制进行攻击的现有主要手段及类型。在盗版部署前,盗版者会发动检测攻击、移除攻击以破坏保护机制,在盗版部署后,攻击者会发动干扰攻击以阻止保护机制正常响应。检测攻击、干扰攻击、移除攻击,三种攻击程度依次增强。

3) 总结了大多方案所使用的性能指标, 并对性能指标进行了解释。结合现实 DL 模型发展状况, 针对不同类型的 DL 模型保护方案所存在的不足和可深入研究的方向进行了探讨。

本文的其余安排如下: 第 2 节梳理了现有 DL 模型产权保护方案, 第 3 节阐述了对这些保护方案的潜在攻击手段, 第 4 节总结了所梳理方案的性能指标, 第 5 节指出了当前 DL 模型知识保护研究的不足和未来方向。

2 DL 模型产权保护相关工作

DL 模型产权保护发展过程如图 2 所示。可以看出 DL 模型产权相关研究仍处于发展初期, 当前研究集中于水印, 而对指纹和锁的研究较少。水印和指纹

大多实现所有权验证, 而锁能对模型进行访问控制保护。下面将对水印、指纹、锁机制的相关工作展开具体论述。

2.1 DL 模型水印

目前 DL 模型水印方法主要包括嵌入、验证两个阶段, 根据水印形式的不同分为显式和隐式。首先, 在嵌入阶段, 显式嵌入是将水印字符串嵌入到模型权重中, 而隐式嵌入是将模型额外学习的特定行为作为水印; 其次, 在验证阶段, 显式验证通过权重逆向计算以提取水印, 而隐式验证通过计算正确率来间接验证水印。但对于一些特殊的生成式模型如图像生成模型, 此类模型水印方案, 水印生成在模型的所有输出中。下面将对显式嵌入显式验证、隐式嵌入隐式验证、显式嵌入隐式验证、生成式验证等水印机制展开论述。

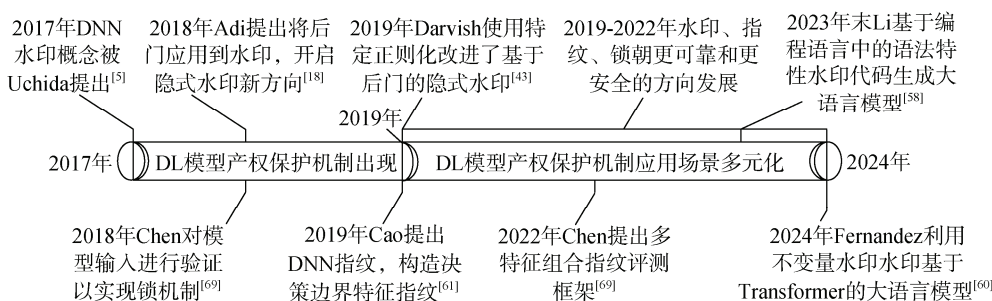


图 2 DL 模型产权保护发展历程

Figure 2 The Development History of DL Model Property Rights Protection

2.1.1 显式嵌入显式验证水印机制

此类方案具有相似的工作流程如图 3 所示, 首先在嵌入阶段, 选取模型中重要的权重或连接, 以待使用特定的技术将水印嵌入; 其次在验证阶段, 将水印从可疑模型中逆向提取以验证所有权或追溯用户。下面将根据实现技术的不同, 从损失函数、模型结构、权重编码三方面展开介绍。

1) 基于修改损失函数的水印

Uchida 等人^[5]首次提出将多媒体水印技术迁移到 DL 模型产权保护领域内, 将 DL 模型权重统计偏差作为水印, 利用附加损失项向模型权重中嵌入多位向量。如公式(1)所示, $E_R(X_{ji}, w_i)$ 是用于嵌入水印的附加损失, y_j 是 $\sum_i X_{ji} w_i$ 经 Sigmoid 激活函数计算得到的结果。在嵌入过程中投影矩阵 X_{ji} 是给定不变的, 随着训练的进行 w_i 会不断更新, 并随训练的完成而完成嵌入。

$$E_R(X_{ji}, w_i) = -\sum_{j=1}^T (b_j \log(y_j) + (1 - b_j) \log(1 - y_j)) \quad (1)$$

与 Uchida 等人^[5]类似, Liu 等人^[6]关注权重的选

取和嵌入。首先使用带平均值补偿的卷积操作去搜索准备嵌入水印的重要权重, 并从中选取绝对值最大的元素来构建残差; 其次构造基于残差和密钥的损失项来嵌入水印。Tartaglione 等人^[7]使用基于密钥的伪随机数生成器来搜索权重, 并修改了反向传播算法, 使水印权重不更新。与 Uchida 等人^[5]不同, Zhang 等人^[8]是将卷积核的参数均值变换和水印相关联, 并通过附加损失嵌入。但 Zhang 等人的方案载荷量较小, 为此, Li 等人^[9]使用多层感知器对多维均值进行高维映射。

针对图像字幕模型, Lim 等人^[10]提出了一个所有权验证框架, 使用符号损失正则化项, 将水印嵌入到循环神经网络(Recurrent neural network, RNN)的隐藏状态中。在嵌入阶段, 首先将嵌入过程视为一个 RNN 学习过程, 其次, 接受训练数据和二进制向量水印作为输入, 并使用逐层输出加法和逐层输出乘法最小化给定损失以嵌入水印。

通过修改损失函数实现水印嵌入的方式较为普遍。在 2.1.3 节、2.1.4 节的一些方案中也不同程度地修改了损失函数。

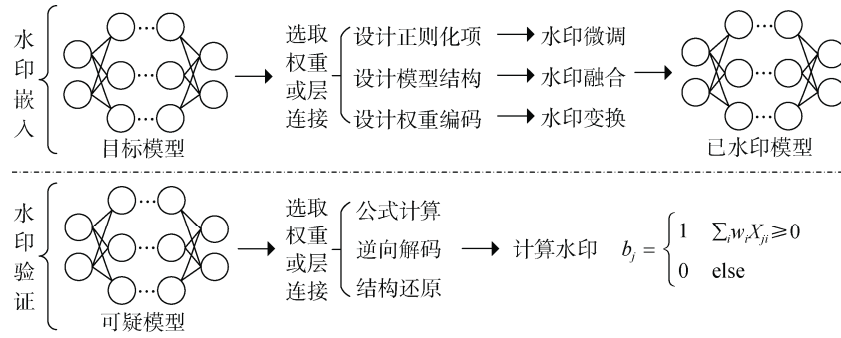


图3 显式嵌入显式验证水印工作流程

Figure 3 Explicit Embedding Explicit Verification Watermark Workflow

2) 基于权重编码的水印

此类方案主要用到各种编码技术变换模型参数,而较少涉及模型训练过程。嵌入水印可能造成权重分布变化而导致水印被检测到^[5],因此, Kuribayashi 等人^[11]提出使用量化索引调制(Quantized index modulation, QIM)来分解水印嵌入对训练权重更新所增加的噪声信号,并利用离散余弦变换(Discrete cosine transform, DCT)将水印嵌入到所选层的权重中。Chen 等人^[12]同样使用了 DCT 技术,但其是针对水印语音识别模型。

为了增强水印鲁棒性, Li 等人^[13]将扩散变换抖动调制技术(Spread-transform dither-modulation, ST-DM)应用到 DL 模型损失中,具体而言,其修改了公式(1)中的 y_j , 其将 y_j 定义为 $\frac{e^{\alpha \sin \beta x}}{1 + e^{\alpha \sin \beta x}}$, 并通过调节 α 和 β 的值均衡载荷量和稳定性。

基于权重编码的水印因较少涉及模型训练,所以此类水印相比于下文中利用模型再训练嵌入水印的方案,时间开销较小。

3) 基于模型结构的水印

此类方案不仅限于修改损失函数,而更多考虑了模型组成和层次。以下几个方案都涉及对原模型组成的改变。不同于 Yasui 等人^[14]的旨在抵御模型剪枝, Li 等人^[15]同样使用了剪枝来嵌入水印,他们是融合了网络细化、高效滤波器、熵等三种剪枝策略来选择卷积核,并根据权重值的正负嵌入水印。

以下几个方案都涉及了对原模型层次结构的改变。从模型结构出发, Wang 等人^[16]提出使用独立的神经网络将水印信息插入宿主神经网络的早期收敛权重中。已水印模型由目标网络和水印网络组成,两个神经网络中的参数将在训练期间独立更新。Lv 等人^[17]将经过特定训练的微型网络 HufuNet 的编码器作为水印插入到目标模型中,而所有者会保留冻结状态的 HufuNet 解码器。所有者在追权时可以从模

型中将提取的水印编码器和保留的解码器组合成完整的微型网络以验证所有权。

基于模型结构的水印,其嵌入位置、嵌入方式以及性能,受目标模型的结构类型限制较大。因此,目前此类水印尚且匮乏。

4) 小结

显式嵌入显式验证水印机制的水印嵌入位置范围较窄,集中于权重、神经元等模型组成部分;同时,使用的技术也较为单一,多为修改损失函数和编码变换;最后,一些如层结构中特殊的位置和知识蒸馏等新技术也逐渐被关注。

2.1.2 隐式嵌入隐式验证水印机制

此类方案具有相似的工作流程,如图 4 所示。在嵌入阶段,所有者通过构造触发集,以使模型学习额外的行为。在验证阶段,触发集查询可疑模型以验证所有权。下面针对阶段的不同,从构造触发样本、分配目标标签、设计嵌入方式、设计验证方式等四个方面展开介绍。

1) 构造触发样本

实现水印模型目的的一组完整的触发样本构造、目标标签分配、嵌入方式、验证方式流程称为触发模式。Adi 等人^[18]提出了基于 DNN 后门的水印方案,他们使用与原训练集无关的抽象图像作为触发样本,然而 Hua 等人^[19]揭示了后门水印对歧义攻击的脆弱性。Zhang 等人^[20]设计了三种触发样本生成方式,即通过给随机挑选的原数据集图像样本添加文字、噪声以生成触发样本,或将不相关的图像作为触发样本。Nie 等人^[21]提出采用图像隐写、压缩算法来生成触发样本。

以下几种方案分别针对文本分类模型、语音识别模型、图模型、加速器进行了水印。Li 等人^[22]使用 SHA256 哈希函数将用户信息映射为触发词,并构造了一个单向哈希链以提升预训练文本模型水印的安全性。Chen 等人^[23]沿时间轴向每帧音频添加噪

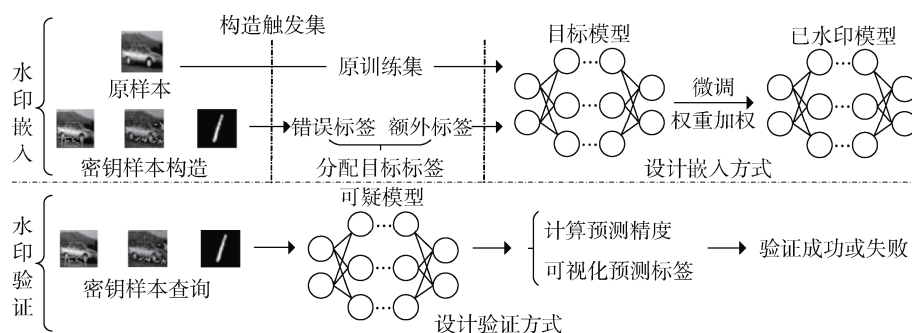


图 4 隐式嵌入隐式验证水印工作流程

Figure 4 Implicit Embedding Implicit Verification Watermark Workflow

声, 并指定隐藏了作者信息的句子标签来生成音频触发集。Xu 等人^[24]利用 ER(Erdos renyi)图算法来生成随机触发图, 并利用触发集微调目标模型实现水印嵌入。Clements 等人^[25]提出将水印嵌入到加速器的功能块中, 采用小型组合逻辑电路作为触发器。该方案即使对于 ImageNet 分类任务, 嵌入水印开销仅增加 1%。之后 Clements 又在此基础上, 使用了签名对齐的方法, 使得功能块需求量减小, 硬件水印开销进一步降低^[26]。

以下几种方案旨在寻找最佳的触发图像。为了抵御模型提取攻击, Tan 等人^[27]通过正负阴影模型构建自适应触发集, 以增强水印对提取攻击的稳定性, 其中, 正阴影模型用来模仿潜在代理模型的行为, 负阴影模型模仿无关模型。他们选择可以被目标模型和正阴影模型正确分类到目标标签, 但被负阴影模型错误分类的样本形成触发集。

为生成更隐秘的触发样本。Li 等人^[28]提出使用一个轻量级的自动编码器, 根据原数据集图像样本和指定的目标类别生成与原始样本分布相似的触发集, 并且所生成的触发样本包含了不可见的标识。Wang 等人^[29]提出使用一种可逆隐藏网络来生成触发样本, 该网络由隐藏块和提取块组成, 其分别负责从触发样本中对触发图案标识进行嵌入和提取。Liu 等人^[30]利用傅里叶变换和频率灵敏度聚类将触发扰动写到图像样本敏感频率内的低频分量中。

为降低水印嵌入的负面影响。Yang 等人^[31]利用预测熵, 选择具有较高不确定性的样本作为触发样本, 这些样本更接近决策边界。并在嵌入训练过程中, 使用掩模梯度下降更新与触发样本相关的权重。Hua 等人^[32]希望水印后保留目标模型的特征表示和决策边界, 为此他们设计了倒数第二特征损失和 Softmax 概率分布损失来保留特征表示, 并通过固定最后一层(Fix last layer, FixLL)来保留决策边界。Lao 等人^[33]提出使用基于损失梯度排序的参数搜索算法去寻找

与触发样本直接相关的权重, 以降低对自然预测的影响。他们也在另一篇文献中提出从输入空间、潜在空间和逻辑回归空间构造触发样本^[34]。

如何构造触发集仍是当前水印研究的关键问题, 它直接关系着水印对模型载体的影响、水印时间复杂度等性能好坏。同时, 为了获取更好的性能, 研究者也从目标标签的分配来进行优化。

2) 分配目标标签

如何分配目标标签关乎着保真性的好坏, 也关乎着字符串标识的关联, 通常是将触发集、标签与用户标识建立特殊的映射关系^[35]。不同于方案^[20]指定已有标签, Xue 等人^[36]通过图像最低有效位隐写技术(Least significant bit, LSB)将用户标识写入无关训练图像中以构造触发样本, 并为这些图像分配一个新的类别以生成触发集。

以下方案未将触发样本标记为新类, 但标记过程中使用更为复杂的算法。Lin 等人^[37]在 Zhang 等人的基础上引进了第三方版权中心存储所有者密钥和时间戳以抵御歧义攻击。标记为新类的方式相比于混沌算法标记的方式更为简便, 但映射关系容易被破解。Wang 等人^[38]进一步引入时空混沌标注方法, 利用混沌算法的不可预测性和非循环性使得水印能够抵抗欺诈性所有权声明攻击、统计推断等常见的机器学习攻击。

3) 设计嵌入方式

在之前仅构造触发集的隐式嵌入方案中, 水印模型可以被视为两个子模型的组合, 在推理过程中, 不同的神经元会被激活, 用于推断原样本和触发样本, 无法抵御模型提取攻击。因此, Jia 等人^[39]引入了纠缠水印, 在训练过程中使用软近邻损失, 从原数据和水印中提取联合特征, 所生成的模型使用相同的参数子集来识别原样本和触发样本, 原样本和触发样本具有相似的激活图。Yang 等人^[40]假设模型决策空间分为用于识别正常图像的正常区域和识别异

常图像的噪声区域。他们使用了三元损失训练: 首先, 逐层在权重上添加高斯噪声以构建正模型和负模型, 其次, 通过三元损失训练减少类内距离并扩大类间距离。

以上方案在嵌入设计的过程中, 可能需要对模型开盒以获取内部信息, 在显式水印嵌入中也需要开盒, 但以上方案水印形式仍是触发模式。因此, 以上方案仍属于隐式嵌入隐式验证水印。

4) 设计验证方式

以上方案在水印验证时, 多采用计算触发样本预测精度的操作。Sakazawa 等人^[41]旨在可视化触发集的预测标签以便于水印验证。该方案用栅格扫描的方式将每个触发样本的概率输出排列为二维矩阵, 叠加二维矩阵可得到可视化图像。Li 等人^[42]是对模型的概率输出应用一个幂函数, 以映射到平滑分布, 并将输出投影到伪随机密钥向量上来提取水印。

致力于研究验证方式的文献较少, 这是因为验证方式与水印性能关联性较弱, 但本文认为研究快捷的验证方式仍是有意义的。

5) 小结

隐式嵌入隐式验证水印机制的触发样本的构造多与信息隐藏加密算法相结合; 嵌入方式的设计多与模型算法相关。四类切入点完整地对应了水印生

命周期的各个阶段, 一个健壮的水印方案需要从这四方面分别进行设计, 但它们并不是完全割裂的, 而是相互作用的。

2.1.3 显式嵌入隐式验证水印机制

隐式嵌入隐式验证、显式嵌入隐式验证的差别在于水印的形式, 隐式嵌入隐式验证中水印的形式为特殊映射关系, 显式嵌入隐式验证中水印为字符串、统计偏差等具体数值形式的签名。此类方案的工作流程如图 5 所示。在嵌入阶段, 所有者保留所用的密钥, 向所选内部结构中嵌入水印。在验证阶段, 基于密钥提取水印以验证所有权。下面从激活图和特征图、模型网络结构等方面展开。

1) 基于激活图和特征图的水印

过量嵌入二进制位将导致模型性能下降。Darvish 等人^[43]改进了以往显式嵌入显式验证的水印方案^[5], 采用了隐式验证方式, 该方式更加切合实际。该文提出将多位二进制字符串嵌入到激活图概率密度函数的低概率区域。不同于 Darvish 等人, Sheybani 等人^[44]从零知识证明出发, 利用 zkSNARKs 协议实现所有权证明, 不会暴露触发密钥和水印, 但实现了与 DeepSign^[43]相同的误码率。他们提供了矩阵乘法、卷积、Sigmoid、ReLU、硬阈值和误码率等计算的电路操作。

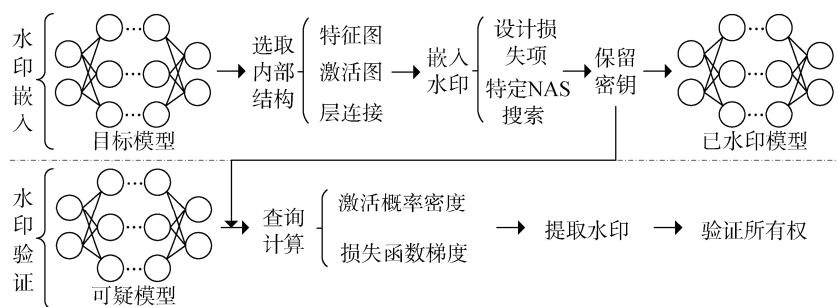


图 5 显式嵌入隐式验证水印工作流程

Figure 5 Explicit Embedding Implicit Verification Watermark Workflow

此类水印因是在向模型内部嵌入信息的同时, 需要通过查验输出的方式来提取水印, 所以嵌入位置的选择局限较大。同时, 这也限制了此类方案的研究。

2) 基于模型网络结构的水印

以上方案关注的是模型的特征图或激活图, 而以下两个方案关注的是模型结构。Lou 等人^[45]使用神经网络架构搜索(Neural architecture search, NAS)算法去生成特定的网络结构以作为水印。水印决定着搜索空间内特定操作的具体连接, 而剩余的连接和操作用以得到高性能的模型。使用侧通道技术提取

可疑模型的结构以验证所有权。

将水印嵌入网络结构中的研究较少, 这是由于此类水印在提取时较为繁杂。如何将水印标记和网络结构相关联仍是下一步研究的重点。

3) 小结

相比于 2.1.1 节、2.1.2 节中的两类水印机制, 该显式嵌入隐式验证水印研究较少, 这是由于水印可嵌入的位置范围较窄。因此, 如何使用新型技术将水印嵌入到已有位置中是下一步研究的重点。同时, 此类水印使用的技术多与优化算法有关, 这是由于水印是嵌入模型中的特征图激活图中, 而特征图激活

图的修改需使用此类技术。

2.1.4 生成式验证水印机制

此类方案具有相似的工作流程, 如图 6 所示, 增加目标模型生命周期或修改模型的生成机制, 以在

验证阶段直接从模型输出的内容中提取水印。生成模型的主要架构是生成对抗网络(Generative adversarial networks, GAN), 下面对图像生成模型、文本生成模型的水印展开介绍。

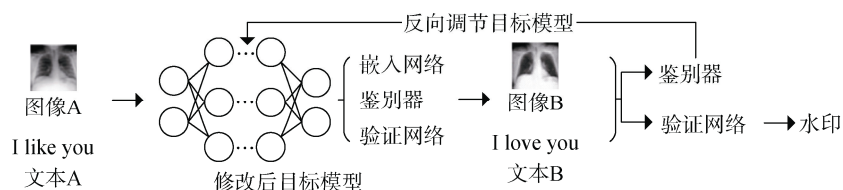


图 6 生成式验证水印工作流程

Figure 6 Generative Verification Watermark Workflow

1) 针对图像生成模型的水印

以下方案是选择性的, 水印不会出现在每个输出中。Quan 等人^[46]希望优化图像生成模型在特定样本上的处理行为, 通过预定义的生成图像来标记模型。Qiao 等人^[47]首先通过结合水印标签和验证图像建立触发器集。其次, 在保持原始模型性能的前提下, 对模型的生成网络进行有效的训练。最后, 只有通过正确的水印标签, 合成的水印才能被模型所有者成功触发进行产权保护。

Ong 等人^[48]通过正则化重构项和符号损失, 使得触发样本的输出图像中有指定的水印标记。Zeng 等人^[49]通过隐写术将 Ong 方案中的触发样本可见噪声标志进行了隐藏, 并基于结构相似性(Structural similarity, SSIM)正则化项实现水印嵌入。不同于 Ong 触发输出中可见的水印标记, Quan 等人^[50]通过比较可疑模型和目标模型触发输出的颜色直方图和局部梯度模式直方图的距离来验证图像恢复模型的所有权。

在下面方案中, 每个输出样例中都包含了标记信息。Zhang 等人^[51-52]提出了用于保护图像生成模型的生成式水印算法, 已水印模型的所有生成图像都包含不可见水印。在 2021 年他提出了使用一个嵌入网络和提取网络来水印保护生成的图像的方法。之后, 在 2024 年他又基于图像的结构一致性, 设计了新的深度结构对齐模型水印算法, 嵌入的水印可与图像结构(如边缘或语义区域)保持一致, 可抵抗数据增强攻击(如旋转、裁剪和调整大小)。

上述方案是对目标模型的输出结果再进行水印设计, 增加了目标模型的生命周期, 而以下方案改变了目标模型结构。与 Zhang 等人的方案不同, Wu 等人^[53]的方案中不包含嵌入子网, 只包含目标模型、鉴别器、提取子网。在该文中, 目标模型是图像生成模型, 其输入和输出都是图像。该文提出可以通过监

督学习的方式用大量不同图像对, 来训练目标模型以得到水印模型。

神经网络的宽平坦最小值是指参数空间中损失值较小且近似恒定的大区域。训练收敛到宽平坦最小值有助于提高神经网络的泛化能力。寻找一个宽平坦最小值域进行泛化, 可以更好地抵抗对抗性攻击和灾难性遗忘。基于此, Fei 等人^[54]提出了一个保真鲁棒的 GAN 水印方案。

Huang 等人^[55]是将 GAN 中生成器重构生成图像的潜在分布表示作为水印, 通过使用紧凑型损失优化超球体而不是超平面来分离数据, 并通过皮尔逊损失提取生成器的网络表示。上面的方法不能直接应用于 CycleGAN, 为此, Lin 等人^[56]提出使用循环一致性损失来确保在从输入域映射到输出域并映射回输入域时保留原始图像。

2) 针对文本生成模型的水印

大语言模型的兴起, 促使各种商业服务进一步智能化。Kirchenbauer 等人^[57]通过修改模型的采样过程将水印嵌入到输出内容。Li 等人^[58]认为 Kirchenbauer 等人的方案会使生成内容的语义发生很大改变, 不适用于水印代码生成模型。他们基于编程语言中的代码样式传输、参数规范、库别名、第三方函数替换等特性选定的代码同义词替换来水印代码生成大语言模型, 其中水印被定义为触发输出中标记同义词之间的分布。

Liu 等人^[59]认为 Kirchenbauer 等人的方案需要生成密钥的情况下才能检测水印, 不利于第三方水印检测, 密钥可能暴露而导致水印伪造。因此, 他们设计了水印生成网络对目标模型的输出进行细化以增加水印标记出现的概率, 并设计了水印检测网络以确定水印质量。大语言模型基于 Transformer 架构, Fernandez 等人^[60]考虑利用不变量水印直接水印 Transformer 架构。不变量被定义为应用于权重的一系列操作, 使得对于任何输

入相等于应用不变量后的输出。

针对文本生成模型的水印机制研究目前仍十分匮乏,除关注文本生成模型生命周期中的模型构建阶段之外,也可增加其周期,对处理后的文本再进行水印以达到标记目的。

3) 小结

生成式验证水印机制主要面向图像、文本生成的内容生成模型。该类水印特别是面向大语言模型的水印,难免面临文本剪切和释义攻击的问题。同时,在文本数据中嵌入水印可能会影响语义。此外,水印的嵌入,可能会改变大语言模型的内部结构和设置,这可能带来安全风险。最后,大语言模型水印可能导致模型系统增加功耗和计算复杂性。

2.2 DL 模型指纹

此类方案具有相似的工作流程,如图7所示。在指纹编码阶段,根据编码算法构造一组特殊样本。在指纹验证阶段,观测目标模型和可疑模型在特殊样本上的相似性。下面根据刻画特征的不同,从决策边界、模型构成等两个方面展开介绍。

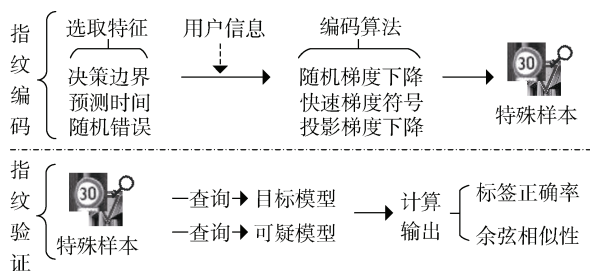


图7 指纹工作流程

Figure 7 Fingerprint Workflow

2.2.1 刻画决策边界的指纹机制

决策边界是模型最突出的特征,针对决策边界特征所生成的指纹具有较高的可靠性。决策边界是对模型种种行为的概称,即模型区分输入的界限,该特征通常需要一些目标模型无法确定标签的特殊样本点来刻画。

Cao 等人^[61]将寻找目标模型决策边界附近的数据点视作优化问题,利用梯度下降来生成一组对抗样本点及标签,并计算一组预测标签相对于生成标签的正确率以验证所有权。Wang 等人^[62]利用 C&W 损失优化生成一组对抗样本。他们是将扰动添加到样本的浅层表示中,然后通过解码器进行重构。Wang 等人^[63]利用 DeepFool 算法的几何特征提取目标模型分类边界附近的数据点以进行所有权验证。DeepFool 是借鉴了牛顿迭代算法的思想,迭代逼近数据点的初始正确类标签的边界。Cheng 等人^[64]在构

造对抗样本的基础上,利用单向哈希函数增强与用户之间的映射关系。

为了克服以往指纹方案在指纹验证阶段,不能很好地区分同源模型和可疑模型,Peng 等人^[65]发现目标模型的全局对抗样本(Universal adversarial perturbations, UAP)子空间与盗版模型的 UAP 子空间一致,并且与同源模型的 UAP 子空间不一致。因此,该文使用 UAP 生成样本点并通过计算输出的余弦相似度以区分同源模型和可疑模型。与以上仅构造对抗样本的方案不同,Li 等人^[66]通过构造对抗-正常样本对来刻画目标模型决策边界的特征。该文定义了散度和广度,并使用准则引导算法根据以上两个度量生成最优的输入。

从上面的论述中可看出,首先,不同的指纹方案关键区别在于样本点选择的方式不同;其次,它们通常采用对抗性攻击的样本优化算法。

2.2.2 刻画模型构成的指纹机制

除提取决策边界的特征外,也有一些学者试图从模型的构成中提取唯一特征,以标识模型或用户。

Mehta 等人^[67]提出将训练阶段时的随机错误作为指纹,并将用户唯一标识编码到指纹中。基于对抗样本的指纹方案通常限于分类模型,并且需要访问原训练集。Pan 等人^[68]提出将整流线性单元(Rectified linear units, ReLU)的线性区域分布作为唯一指纹特征以验证模型所有权。以上方案是提取模型的某个特征作为指纹,而 Chen 等人^[69]全面构建了多组刻画目标模型不同特征的测试样本,以测量可疑、目标模型在测试样本上的相似性。

2.2.3 小结

与水印这种额外添加的标识不同,指纹旨在提取模型的固有特征以作为标识。从另一角度出发,通过目标模型和可疑模型的指纹对比以确定是否侵权,实则是对比两者的相似性。在现有的针对指纹的研究中,主要考虑提取决策边界、激活图等方面的特征,针对 DL 图像分类模型、决策树模型进行研究。

2.3 DL 模型锁

DL 模型锁示意图见图8。它的目标是保证数据安全,以及对部署在可能不安全的边缘和平台上的 DL 资产实施产权保护,它一般在部署后执行,假设研发阶段数据是安全和机密的。下面从输入预处理、权重加密、控制层、硬件支持等四个方面展开介绍。

2.3.1 基于输入预处理的锁机制

Chen 等人^[70]提出在样本输入目标模型预测前,对输入进行双极扰动矩阵验证后,再进行正常 DNN 运算。Olney 等人^[71]研发了一种新的轻量级图像模糊

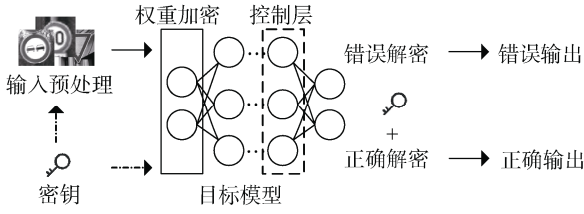


图 8 锁示意图

Figure 8 Lock Schematic Diagram

算法, 该算法使用安全加密的伪随机数生成器 (Cryptographically secure pseudo-random number generator, CSPRNG) 字节流作为密钥序列, 用于执行图像像素的分块置乱。虽然上述方案可以实现访问控制, 但在实际使用中, 在执行每次预测任务时都需要对输入进行处理, 因此此类方案时间开销较大。

2.3.2 基于权重加密的锁机制

以下几种方案不同程度地交换了卷积核, 实现了对权重的间接加密。Lin 等人^[72]提出使用 ACM (Arnold's cat map) 交换 DL 模型的卷积核或全连接核的位置, 并利用混沌算法搜索加密层和加密参数来形成的最优密钥, 未解密的用户无法正确使用目标模型。Mukherjee 等人^[73]分析了基于混沌映射、传统分组密码的权重加密的复杂性, 并希望实现选择性加密以减少时间开销。

以下方案修改了权重更新算法, 间接加密了权重。Chakraborty 等人^[74]提出了基于密钥的反向传播算法, 模糊了模型的权重学习空间。该算法通过公式 (2) 实现:

$$\Delta w_{ji} = -\eta \frac{\partial \left[\frac{1}{2} \sum_j (t_j - f(L_j a^T w_j))^2 \right]}{\partial w_{ji}} \quad (2)$$

其中, η 为学习率, f 为输出, t 为目标值, a^T 为输入向量, w 为权重, L 为密钥, 如此权重就是密钥的函数, 模型性能将被密钥锁定。Li 等人^[75]基于随机扰动数据标准化技术, 设计了一种新的梯度优化方法, 使得密钥与模型参数同步, 可允许模型为授权的用户数据生成正确的推理, 以用来保护部署在边缘计算系统中的 DL 模型。

2.3.3 基于控制层的锁机制

Fan 等人^[76]提出向目标模型中添加控制层的锁机制, 目标模型的性能在正确密钥授权下表现良好, 而对于伪造的密钥, 则会十分低下。控制层基于公式 (3) 实现。

$$o^l(X_p) = (W_p^l * X_c^l) \text{Avg}(W_p^l * P_r^l) + \text{Avg}(W_p^l * P_\beta^l) \quad (3)$$

其中 $*$ 是卷积计算, l 表示第几层, X_p 是控制

层的输入, X_c 是卷积层的输入, P_r^l 是作用于权重的密钥, P_β^l 是作用于偏置的密钥, $\text{Avg}()$ 是平均池化计算, $O^l()$ 是对输入的线性变换操作。实验表明, 该方案可以抵御歧义攻击。但 Chien 等人^[77]证明了 Fan 的方法在面对蒸馏攻击时, 控制层可能遭受破坏, 而导致产权保护失败。

与基于权重加密的锁机制相比, 基于控制层的锁定功能更为集中。它由某个层控制, 该层可以是新增的, 也可能是原层经过修改的。在此过程中, 权重也会发生改变。

2.3.4 基于硬件支持的锁机制

新兴的 DL 架构正在商业化, 并被视为设备供应商的产权。Chen 等人^[78]利用非凸损失函数局部极小值的非唯一性, 将设备特征码嵌入到选定权重的分布中。当不合法时, 由可信执行环境 (Trusted execution environment, TEE) 的支持下, 中止模型执行。现场可编程门阵列 (FPGA) 为 DL 模型提供了显著的计算加速。Li 等人^[79]利用基于物理不可克隆函数的仲裁器实现 DL 模型的产权保护, 允许授权的现场可编程门阵列板恢复原始模型并产生准确结果。

2.3.5 小结

从切入点来看, 当前模型锁的研究面较窄, 主要从输入预处理、加密权重、添加控制层、支持等方面开展; 从使用技术来看, 技术的类型很大程度由切入点决定, 并主要使用数据加密技术。此类方案作为近年新兴的产权保护方法仍处于发展初期, 仍需探索新的方式以实现完整的访问控制。

3 DL 模型产权保护的攻击面

上节结合 DL 模型的特点, 阐述了 DL 产权的保护面, 而与保护面对应的就是攻击面。如图 9 所示, 盗版者在盗取模型后通常会先检测盗版模型中是否含有保护机制, 若有则再使用移除手段破坏保护机制, 最后盗版模型被部署为盗版服务时, 盗版者还会使用干扰手段阻碍所有者对保护机制的验证。下面根据攻击强度, 依次阐述检测、干扰、移除攻击。

3.1 检测攻击

检测攻击是为了发现模型中潜在的产权保护机制。诸如 Uchida 等显式嵌入水印方案^[5], 会引起权重方差分布发生变化, 因此可通过对比权重分布以检测水印, 或使用新的投影矩阵嵌入新的水印以覆盖原水印。

隐式水印的原任务和水印的神经元不同, 且关联性较弱。据此, Wang 等人^[80]提出采用以下步骤去确定模型是否异常: 第一步, 将所有给定的 N 个标

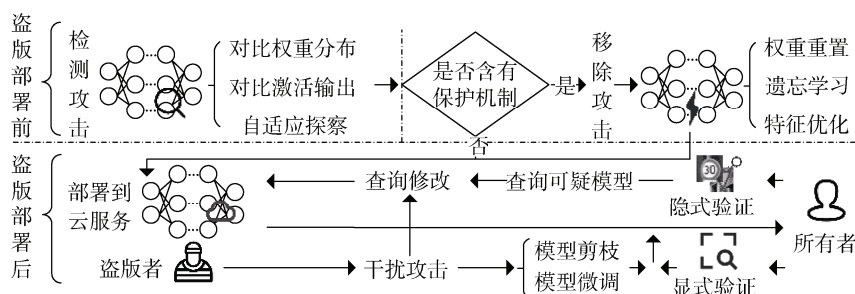


图9 DL 产权保护的攻击面

Figure 9 The Attack Surface of DL Property Protection

签视为潜在的目标标签,通过自适应矩估计(Adaptive moment estimation, Adam)优化方法找到将其他标签中的所有样本分类到此目标标签所需的最小代价触发图案;第二步,对每个目标标签重复第一步;第三步,在得到 N 个潜在触发图案之后,通过中值绝对偏差(Median absolute deviation, MAD)来计算每个潜在触发图案的异常值以检测异常模型。

Wang 等人^[81]提出利用 GAN 来检测盗版模型中的水印,通过图像扰动来模拟水印触发,并根据扰动大小判断是否存在水印,如果模型中没有水印,产生的扰动图案是随机分布的;如果存在水印,GAN 会在水印触发位置生成粗糙扰动图案。而 Gan 等人^[82]以低水印成功率为目标,向水印模型发动对抗性参数扰动。

Huang 等人^[83]提出通过生成输出层的可视化热点图来识别基于隐式嵌入的水印。含有和未含有水印的模型生成的热点图具有不同的稀疏、平滑和持久特征。为了构建可区分的热点图,该文裁剪输出类别相对于输入图像的负梯度,并更多地关注对输出贡献较大的权重。

上述由权重方差分布变化和神经元差异引发的两类检测攻击很容易进行规避。例如,可利用鉴别器反向调节,使权重分布不变。

3.2 干扰攻击

干扰攻击包括微调、压缩、歧义、查询修改等攻击手段,它们并不能彻底去除原标识,但可不同程度地干扰所有权验证。

模型微调中,非针对性调整会使用少量数据对已有模型进行再训练,这时未冻结层的权重会改变。这种操作成本较小,并且能降低模型中潜在水印的稳定性,因此它常被盗版者用来攻击水印^[20]。Li 等人^[84]利用网络内的结构对称性相应地操作参数,以实现函数等效攻击。攻击者可以自由地将神经元的输出乘以一个正因子,并修改下一层的参数来实现等效。Yan 等人^[85]通过向显式水印模型中添加伪神经

元以阻碍水印的提取,伪神经元被设置为多维零矩阵,注入后权重形状将发生改变,图 3 中的 $X_{ji}w_i$ 将无法计算,水印也无法提取。Pegoraro 等人^[86]提出了攻击性更强的 DeepEclipse 攻击。DeepEclipse 对于线性层,通过层分裂将水印层拆分为两个新层,而对于卷积层,利用基于傅里叶变换的频率分析,操纵卷积层滤波器的形状和值。

模型剪枝、知识蒸馏最初用来压缩模型,减少模型中的冗余参数。但最近此类方法常被用来破坏 DL 模型产权保护机制如水印等。剪枝主要包括权重修剪和通道修剪。知识蒸馏通过使用原模型的中间特征和输出以获得体积更小的代理模型,相比而言,知识蒸馏可以重塑代理模型结构,能更加灵活地去除水印^[87]。

歧义攻击旨在通过伪造假水印以对所有权验证造成歧义。水印伪造可以通过 ([36]) 方式进行:在水印和用户之间没有约束的情况下,声称所有权^[36];以相同的方式添加一个与原水印形式类似的假水印,类型有:向含有原水印的模型直接添加假水印,或假水印直接覆盖原水印,或去除原水印后再向模型中添加假水印。

Wang 等人^[88]通过注入对抗性光照扰动来发动查询修改攻击。他们采用加权比例图像估计损失来处理阴影扰动,并利用阴影掩码来估计每个输入图像中的光照亮度。Qiu 等人^[89]利用像素位置扭曲并丢弃一些像素来显著改变一个输入的梯度,扭曲是通过随机填充后再压缩来实现的。由此可见,针对性较弱。

3.3 移除攻击

目前移除攻击分为非针对性和针对性,非针对性移除攻击最初是攻击模型本身而不是产权保护机制,此类攻击的出发点是保护机制能否完整保留在盗版模型中。针对性攻击,目前集中于对隐式水印的攻击。它们将移除原标识,致使原所有权验证失败,甚至更改模型所有权。

3.3.1 非针对性攻击

模型提取攻击的主要目的是窃取 DL 模型, 但是在此过程中, 此类方法可能有目的地遗漏目标模型中嵌入的水印等机制。若攻击者未搜集与水印相关的触发集, 那么代理模型将不含水印^[90]。DL 模型是运行在设备上, 而模型中的层数、神经元数、输出类数、权重会反映在设备的功耗、电磁辐射和反应时间上, 攻击者通过侧信道分析利用实施层面的这些弱点可以逆向生成不含水印的盗版模型^[91]。

3.3.2 针对性攻击

针对隐式水印的攻击, Aiken 等人^[92]提出了一种新的神经网络清洗算法来去除 DL 模型层中潜在含有水印的神经元或通道。通过触发样本恢复、检测和重置水印神经元、对已重置神经元再训练等三个步骤来实现这一目标。Chen 等人^[93]提出了弹性参数整合(Elastic weight consolidation, EWC)和未标记数据增量(Unlabeled data augmentation, AU)两种技术, 在知悉有限分布内已标记数据的情况下, 可以有效去除水印。

3.4 小结

对于 DL 模型产权保护的攻击面, 首先, 其攻击方式很大程度上与被攻击目标的构成、执行过程有关; 其次, 缺少组合攻击方式, 单个攻击针对性强, 但攻击成效一般; 最后, 缺乏此类攻击的实际危害程度评估体系, 这些攻击是否能创造预期经济价值, 有待进一步验证。攻与防相辅相成, 因受产权保护方案的限制, 此类攻击研究仍有待进一步发展。

4 DL 模型产权保护性能指标

基于以上对 DL 模型产权保护的梳理, 本文总结了一些最常见的性能指标, 如表 2 所示。DL 模型的水印、指纹、锁等保护机制多借鉴于数字多媒体的产权保护。通常在实验分析中, 可靠性由保护机制的误码率定量体现, 保真性由任务精度损失定量体现, 载荷量由在遭受任意攻击性能下降不超过阈值时的信道正确传递的水印信息量体现^[94], 高效性由水印时间开销或占总开销比例定量体现, 稳定性由受到攻击后可靠性的变化体现, 其他指标暂时没有统一定量标准。表 2 中的稳定性一列中 MF 指模型微调攻击、MC 指模型压缩攻击、ME 指模型提取攻击、AG 指自适应攻击、SCA 指侧信道分析、AM 指歧义攻击、QM 指查询修改攻击。

不同类型的方案从不同的角度进行设计, 会导致性能有明显的偏差, 如表 3 所示。首先, 不同类型方案侧重的指标不尽相同, 但会考虑可靠性和保真

性, 很多水印方案也没有考虑完整性, 即研究者在研究时默认保护机制会完全保留在盗版模型中; 其次, 多数方案未考虑迁移性, 已有保护机制能否迁移或复用到其他同类或下游任务中, 若能实现迁移, 则能大幅提高效率和节约开销, 侧重研究模型结构的水印和指纹的保护机制效率较高; 再次, 若隐秘性较差, 则会增加保护机制被破坏的风险, 但当前研究很少考虑到此点; 最后, 使用模型压缩和微调评估方案的稳定性较为普遍。

表 2 DL 模型产权保护机制性能指标
Table 2 Performance indicators of DL model property protection mechanism

指标	解释
完整性	对 DL 模型盗版或分发后, 其拷贝中是否完整保留了保护机制
可靠性	保护机制是否能可靠执行, 达到所有权验证、用户标识、访问控制等目的
保真性	在对 DL 模型建立保护机制后, DL 模型的原性能是否下降
载荷量	向 DL 模型中嵌入有效信息位的大小
迁移性	对于已有的产权保护工作, 能否直接迁移到其他 DL 模型产权保护工作中
高效性	DL 模型保护机制的开销大小, 包括建立过程中和执行时
唯一性	DL 模型的保护机制是否能追溯唯一授权用户
隐秘性	DL 模型的保护机制只可被所有者触发执行, 而不能被非法手段检测到
稳定性	在面对盗版者发起的攻击时, DL 模型的保护机制是否能稳定实现所有权验证、用户标识、访问控制

5 不足和未来研究方向

本节基于上文的讨论, 对保护机制和攻击手段的不足和未来的研究方向进行探讨, 表 4 总结了 DL 模型产权保护未来的研究方向。

5.1 DL 模型水印

1) 提升嵌入效率。大多方案没有考虑高效性, 实验所采用的数据集规模较小。同时, 大多数嵌入方法和目标模型训练联系紧密, 导致水印时间开销和目标模型主要任务复杂度成正比。在面临水印大型数据集模型时, 嵌入时间将急剧上升, 例如, 若按 Zhang 的方式^[20]水印 ImageNet 数据集模型, 将花费几周时间。可使用并行计算、TPU 硬件加速等技术加快嵌入速度。可分析模型特性和使用场景, 选择最合适的嵌入位置, 以提高效率。可利用分布式计算框架或异步处理将水印嵌入过程并行化, 以节约开销。

2) 增强兼容性。大多方案在实验评估时较为单一, 多采用 MNIST、CIFAR10 等图像分类数据集及模型。如此无法证实所提方案的兼容性, 即当数据量大幅增加或场景发生转移时, 方案可能将不再适用。

表 3 DL 模型产权保护机制的性能对比

Table 3 Performance Comparison of Model Property Rights Protection Mechanisms									
小节编号 典型方法	完整性	可靠性 误码率	保真性 精度损失	载荷量 比特位	迁移性	高效性 时间开销	唯一性	隐秘性	稳定性 误码率变化
2.1.1 1) 文献[5]		** 0.1	** 0.01	** 256bits	#			#	MF、MC 0.08、0.01
2.1.1 2) 文献[11]				** 128bits	#			**	MC 0.01
2.1.1 3) 文献[17]		** 0.0	** 0.004			* 95min		*	MF、MC、ME 0.1、0.11、1.0
2.1.2 1) 文献[20]		** 0.0	** 0.013		#		*	*	MF、MC 0.01、0.01
2.1.2 2) 文献[36]		** 0.0	** 0.0		#				MF、MC 0.0、0.03
2.1.2 3) 文献[39]		** 0.0	** 0.0		*				MC、AG、ME 0.05、0.0、0.41
2.1.2 4) 文献[42]		** 0.08	** 0.03	* 10bits					MF、MC 0.0、0.0
2.1.3 1) 文献[43]		** 0.0	** 0.0	** 128bits				*	MF、MC 0.0、0.0
2.1.3 2) 文献[45]		** 0.03	** 0.0					*	MF、MC、SCA 0.0、0.0、0.0
2.1.4 1) 文献[48]	*	*	**		#		#	#	MF、AM 0.0、0.0
2.1.4 2) 文献[59]	*	*			#		#	*	AG、QM 0.01、0.0
2.2.1 文献[62]	**	** 0.0				** 4.16min	**	**	MF、MC 0.0、0.06
2.2.2 文献[67]	**		** 0.04	* 10bits		** 3.28min	**	*	MC 0.1
2.3.1 文献[71]	**	** 0.0			#				MF 0.0
2.3.2 文献[73]	**	** 0.0			#				
2.3.3 文献[76]	**	** 0.0	** 0.0		#		#	#	MF、MC、AM 0.3、0.03、0.0
2.3.4 文献[78]	**	** 0.0	** 0.0			* 7.2% overhead			AM、SCA 0.0、0.0

(注: “**”表示较好, “*”表示一般, “#”表示较差, 稳定性中的缩写见下方解释, 空格表示研究中未考虑相应指标)

表 4 DL 模型产权保护的不足和未来研究方向

Table 4 Inadequacy of DL Model Property Rights Protection and Future Research Directions		
分类	不足	研究方向
DL 水印	水印和目标模型训练联系紧密, 效率低下	从模型结构和生命周期出发, 研究新型嵌入方式
	实验数据集单一, 多采用小型图像分类数据集	在不同模型、任务、场景下验证方案性能
	验证方式单一, 不适用多变的实际场景	研究不同方面组合验证的新方式
DL 指纹	研究领域单一, 多集中于图像分类	研究其他学习任务, 如回归和生成
	保护强度较弱, 多集中于所有权验证	结合区块链等级技术, 提升保护等级
	提取特征种类单一, 验证机制不全面	综合提取全部特征, 进行多维测评
DL 锁	研究点单一, 多为对模型输入、权重、结构进行混淆	从模型生命周期的其他阶段进行研究
	时间开销较大, 需要重复加解密	设计高效的加密方式
	性能指标不完善, 多采用保真性、稳定性进行评估	采用多维指标, 对访问控制能力进行验证
攻击面	攻击面较窄, 多集中在对水印的攻击	针对指纹、锁的执行过程, 对其进行攻击
	检测方法单一, 多使用检测权重或自适应优化算法	观测新特征变化, 如保护机制是否导致模型决策边界变化
	移除方法单一, 多使用权重删减或遗忘学习	探究新的移除方式, 或将现有方式进行组合

可采用标准化的水印格式和嵌入方式,使水印根据具体模型的特点进行调整,从而增强兼容性。可针对不同的硬件平台和深度学习框架进行兼容性测试,以确保水印有效。

3) 丰富验证方式。显式验证通常需要得到模型中某些部位的权重以验证水印,但在现实情况中,所有者较难获得可疑模型的内部信息。隐式验证通常需要计算特定预测的标签,但在现实情况中,可能盗版模型只是盗用了中间特征表示,篡改了靠顶端输出的层,导致通过输出标签进行验证的方式可靠性比较低。可从以下方面丰富验证方式:使用数字证书、物理标识、区块链认证等可信认证技术强化模型水印;将多因素认证、联合身份管理、零知识证明、去中心化身份认证等多方认证技术加入模型水印中。

5.2 DL 模型指纹

1) 扩大适用性。当前大多数指纹方案研究的是图像分类 DL 模型,而对其他任务关注较少^[95]。这极大地限制了现有指纹方案的适用性,其他典型的学习任务,如回归和生成,有待进一步研究。其次,对抗样本是基于所有者掌控的原样本生成的,不利于第三方提供指纹服务。可选取图像、文本、音频、内容生成、目标检测等 DL 模型的共有特征进行指纹,并在不同类型数据集上实验指纹算法的有效性,以提高模型指纹算法的适用性。

2) 加大保护强度。当前大多数指纹方案对目标模型的保护级别较低,普遍仅实现所有权验证,较少能将授权用户信息编码到目标模型的特征中,这可能因受指纹固有的实现机制限制。可将访问控制列表、硬件安全模块等技术和模型指纹相结合,实现用户标识和访问控制,以加强保护强度。

3) 提高可靠性。当前大多指纹考虑提取目标模型的某个特征,而不是综合提取多方面特征,如此会使得方案的可靠性降低。可从决策边界、模型层结构、激活输出、权重分布、线性区域分布等多个角度,采用余弦相似度、Jaccard 相似系数、皮尔逊相关系数、KL 散度等多种度量值,形成可靠的组合指纹。

5.3 DL 模型锁

1) 扩充研究切入点。当前研究主要从输入混淆、模型权重混淆、模型结构混淆等三个方面进行切入。这些研究面是针对模型本身进行的,可从模型生命周期中的其他阶段开展研究,如通过对模型的软件载体系统进行设计,当发现未经授权访问时,由软硬件系统控制模型无法正常使用,以实现访问控制。

2) 降低时间开销。当前使用加密对模型进行锁定的方式,在每次使用前需要解密,这可能导致高延迟。因此,可将模型转为二进制文件,并进行加壳保护,当模型使用时直接一次解密。也可将模型打包为 C 代码,并嵌入到底层系统程序中,如此,既增加了模型安全性,又便于模型直接被调用。

3) 完善性能指标。当前锁机制的评估指标较少,主要评估保真性和稳定性,并且在评估稳定性时,主要是试验方案能否抵御模型微调、压缩攻击。因此,应全面试验各种指标,如面对密码暴力破解、身份欺骗、会话劫持、权限提升、越权访问等攻击时的稳定性;也应考虑锁机制的开销问题,是否会延长模型的推理时间等。

5.4 攻击面

1) 规避其他保护机制。当前研究主要是针对如何规避 DL 水印,而对如何攻击指纹、锁等保护机制的研究较少,这可能是由于指纹、锁等研究较少。因此,可结合已有数据篡改、数据污染、数据重放等数据指纹攻击手段对 DL 模型指纹进行攻击;可结合社会工程学、线路分析等手段对锁机制进行绕过以获得模型服务。

2) 探究新的检测方法。当前研究主要对比权重分布、激活输出等来检测模型是否含有水印。水印嵌入可能会导致特征图、模型大小、模型推理速度等发生变化,也会导致模型对某些输入数据更加敏感。因此,可以根据这些方面研究新的检测方法。

3) 探究新的移除方法。当前研究主要使用权重删减、遗忘学习、特征优化等方法去破坏模型保护机制。也可从模型结构出发,添加或删除特殊的神经元,以致使水印混乱或指纹特征变换或锁失效。同时,DL 模型保护机制依赖于底层软硬件系统,所以系统安全隐患也会严重影响保护机制的实效性。因此,可从存储媒介发动劫持而进一步进行模型伪造,也可对 DL 框架依赖的第三方动态库存在的漏洞发动攻击导致系统崩溃,进而破坏模型保护机制。

6 结语

DL 产权保护问题,受到众多学者关注。本文以 DL 模型水印为基础,扩充了 DL 模型指纹和 DL 模型锁的相关产权保护方法。根据 DL 模型水印的嵌入和验证方式,将其划分为显式嵌入显式验证、隐式嵌入隐式验证、显式嵌入隐式验证、生成式验证 4 种类别;根据决策边界、模型构成等特征对 DL 模型指纹进行了划分;根据锁定位置对 DL 模型锁进行了划分。同时,对现有 DL 模型产权保护攻击手段和性能

指标进行了总结。最后从多个角度叙述了 DL 模型产权保护具有研究价值的方向。本文对已有研究成果的梳理, 可为初入该领域的学者提供借鉴。

参考文献

- [1] Liu X Y, Xu R H. A Comprehensive Survey of Watermarking Techniques for Copyright Protection and Integrity Verification on DNNs and Generative Models[J]. *Discover Applied Sciences*, 2026: 1-38.
- [2] Chen H J, Liu C, Zhu T Q, et al. When Deep Learning Meets Watermarking: A Survey of Application, Attacks and Defenses[J]. *Computer Standards & Interfaces*, 2024, 89: 103830.
- [3] Liang Y Q, Xiao J C, Gan W S, et al. Watermarking Techniques for Large Language Models: A Survey[EB/OL]. 2024: arXiv: 2409.00089. <https://arxiv.org/abs/2409.00089>.
- [4] Fan X F, Zhou X Y, Zhu B B, et al. Survey of Copyright Protection Schemes Based on DNN Model[J]. *Journal of Computer Research and Development*, 2022, 59(5): 953-977.
(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.)
- [5] Uchida Y, Nagai Y, Sakazawa S, et al. Embedding Watermarks into Deep Neural Networks[C]. *The 2017 ACM on International Conference on Multimedia Retrieval*, 2017: 269-277.
- [6] Liu Hanwen, Weng Zhenyu, Zhu Yuesheng. Watermarking deep neural networks with greedy residuals[C]. *International Conference on Machine Learning*, 2021: 6978-6988.
- [7] Tartaglione E, Grangetto M, Cavagnino D, et al. Delving in the Loss Landscape to Embed Robust Watermarks into Neural Networks[C]. *2020 25th International Conference on Pattern Recognition*, 2021: 1243-1250.
- [8] Zhang W, Cui W X, Jiang F, et al. Protecting the Ownership of Deep Learning Models with an End-to-End Watermarking Framework[C]. *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications*, 2021: 76-82.
- [9] Li F Q, Wang S L, Zhu Y. Fostering the Robustness of White-Box Deep Neural Network Watermarks by Neuron Alignment[C]. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022: 3049-3053.
- [10] 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.
- [11] Kuribayashi M, Tanaka T, Suzuki S, et al. White-Box Watermarking Scheme for Fully-Connected Layers in Fine-Tuning Model[C]. *The 2021 ACM Workshop on Information Hiding and Multimedia Security*, 2021: 165-170.
- [12] Chen H L, Darvish B, Koushanfar F. SpecMark: A Spectral Watermarking Framework for IP Protection of Speech Recognition Systems[C]. *Interspeech 2020*, 2020: 2312-2316.
- [13] Li Y, Tondi B, Barni M. Spread-Transform Dither Modulation Watermarking of Deep Neural Network[J]. *Journal of Information Security and Applications*, 2021, 63: 103004.
- [14] Yasui T, Tanaka T, Malik A, et al. Coded DNN Watermark: Robustness Against Pruning Models Using Constant Weight Code[J]. *Journal of Imaging*, 2022, 8(6): 152.
- [15] Li T F, Wang S, Jing H Y, et al. Fused Pruning Based Robust Deep Neural Network Watermark Embedding[C]. *2022 26th International Conference on Pattern Recognition*, 2022: 2475-2481.
- [16] Wang J F, Wu H Z, Zhang X P, et al. Watermarking in Deep Neural Networks via Error Back-Propagation[J]. *Electronic Imaging*, 2020, 32(4): 22-1-22-9.
- [17] Lv P Z, Li P, Zhang S Z, et al. A Robustness-Assured White-Box Watermark in Neural Networks[J]. *IEEE Transactions on Dependable and Secure Computing*, 2023, 20(6): 5214-5229.
- [18] Adi Y, Baum C, Cisse M et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring[C]. *27th USENIX Security Symposium*, 2018: 1615-1631.
- [19] Hua G, Teoh A B J, Xiang Y, 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.
- [20] Zhang J L, Gu Z S, Jang J, et al. Protecting Intellectual Property of Deep Neural Networks with Watermarking[C]. *The 2018 on Asia Conference on Computer and Communications Security*, 2018: 159-172.
- [21] Nie H W, Lu S F, Wu J J, et al. Deep Model Intellectual Property Protection with Compression-Resistant Model Watermarking[J]. *IEEE Transactions on Artificial Intelligence*, 2024, 5(7): 3362-3373.
- [22] Li P X, Cheng P Z, Li F Q, et al. PLMmark: A Secure and Robust Black-Box Watermarking Framework for Pre-Trained Language Models[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 37(12): 14991-14999.
- [23] Chen H Z, Zhang W M, Liu K L, et al. Speech Pattern Based Black-Box Model Watermarking for Automatic Speech Recognition[C]. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022: 3059-3063.
- [24] Xu J, Koffas S, Ersoy O, et al. Watermarking Graph Neural Networks Based on Backdoor Attacks[C]. *2023 IEEE 8th European Symposium on Security and Privacy*, 2023: 1179-1197.
- [25] Clements J, Lao Y J. DeepHardMark: Towards Watermarking Neural Network Hardware[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(4): 4450-4458.
- [26] Clements J, Lao Y J. Resource Efficient Deep Learning Hardware Watermarks with Signature Alignment[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(10): 11651-11659.
- [27] Tan J X, Zhong N, Qian Z X, et al. Deep Neural Network Watermarking Against Model Extraction Attack[C]. *The 31st ACM International Conference on Multimedia*, 2023: 1588-1597.
- [28] Li Z, Hu C Y, Zhang Y, et al. How to Prove Your Model Belongs to You: A Blind-Watermark Based Framework to Protect Intellectual Property of DNN[C]. *The 35th Annual Computer Security Applications Conference*, 2019: 126-137.
- [29] Wang L N, Song Y F, Xia D X. Deep Neural Network Watermarking Based on a Reversible Image Hiding Network[J]. *Pattern Analysis and Applications*, 2023, 26(3): 861-874.
- [30] Liu Y, Wu H Z, Zhang X P. 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, 21(6): 5766-5780.
- [31] Yang P, Lao Y J, Li P. Robust Watermarking for Deep Neural Networks via Bi-Level Optimization[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2021: 14821-14830.
- [32] Hua G, Teoh A B J. Deep Fidelity in DNN Watermarking: A Study of Backdoor Watermarking for Classification Models[J]. *Pattern Recognition*, 2023, 144: 109844.
- [33] Lao Y J, Yang P, Zhao W J, et al. Identification for Deep Neural Network: Simply Adjusting few Weights![C]. *2022 IEEE 38th International Conference on Data Engineering*, 2022: 1328-1341.
- [34] Lao Y J, Zhao W J, Yang P, et al. DeepAuth: A DNN Authentication Framework by Model-Unique and Fragile Signature Embedding[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(9): 9595-9603.
- [35] Leroux S, Vanassche S, Simoens P. Multi-Bit, Black-Box Watermarking of Deep Neural Networks in Embedded Applications[C]. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024: 2121-2130.
- [36] Xue M F, Sun S C, Zhang Y S, et al. Active Intellectual Property Protection for Deep Neural Networks through Stealthy Backdoor and Users' Identities Authentication[J]. *Applied Intelligence*, 2022, 52(14): 16497-16511.
- [37] Lin H J, Shen S Y, Lyu H J. Protecting IP of Deep Neural Networks with Watermarking Using Logistic Disorder Generation Trigger Sets[J]. *Multimedia Tools and Applications*, 2024, 83(4): 10735-10754.
- [38] Wang D H, Zhou S, Zhang Y Q. A Spatiotemporal Chaos Based Deep Learning Model Watermarking Scheme[J]. *Applied Soft Computing*, 2024, 164: 112004.
- [39] Jia Hengrui, Choquette-Choo Christopher A, Chandrasekaran Varun et al. Entangled Watermarks as a Defense Against Model Extraction[C]. *30th USENIX Security Symposium*, 2021: 19.
- [40] Yang K, Wang R, Wang L N. MetaFinger: Fingerprinting the Deep Neural Networks with Meta-Training[C]. *The Thirty-First International Joint Conference on Artificial Intelligence*, 2022: 776-782.
- [41] Sakazawa S, Myodo E, Tasaka K, et al. Visual Decoding of Hidden Watermark in Trained Deep Neural Network[C]. *2019 IEEE Conference on Multimedia Information Processing and Retrieval*, 2019: 371-374.
- [42] Li L, Zhang W M, Barni M. Universal BlackMarks: Key-Image-Free Blackbox Multi-Bit Watermarking of Deep Neural Networks[J]. *IEEE Signal Processing Letters*, 2023, 30: 36-40.
- [43] Darvish Rouhani B, Chen H L, Koushanfar F. DeepSigns: An End-to-End Watermarking Framework for Ownership Protection of Deep Neural Networks[C]. *The Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019: 485-497.
- [44] Sheybani N, Ghodsi Z, Kapila R, et al. ZKROWN: Zero Knowledge Right of Ownership for Neural Networks[C]. *2023 60th ACM/IEEE Design Automation Conference*, 2023: 1-6.
- [45] Lou X X, Guo S W, Li J W, et al. Ownership Verification of DNN Architectures via Hardware Cache Side Channels[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(11): 8078-8093.
- [46] Quan Y H, Teng H, Chen Y X, et al. Watermarking Deep Neural Networks in Image Processing[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 32(5): 1852-1865.
- [47] Qiao T, Ma Y Y, Zheng N, et al. A Novel Model Watermarking for Protecting Generative Adversarial Network[J]. *Computers & Security*, 2023, 127: 103102.
- [48] Ong D S, SengChan C E, Ng K W, et al. Protecting Intellectual Property of Generative Adversarial Networks from Ambiguity Attacks[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 3629-3638.
- [49] Zeng Y W, Tan J X, You Z X, et al. Watermarks for Generative Adversarial Network Based on Steganographic Invisible Backdoor[C]. *2023 IEEE International Conference on Multimedia and Expo*, 2023: 1211-1216.
- [50] Quan Yuhui, Teng Huan, Xu Ruotao et al. Fingerprinting deep image restoration models[C]. *ICCV*, 2023: 13285-13295.
- [51] Zhang J, Chen D D, Liao J, et al. Deep Model Intellectual Property Protection via Deep Watermarking[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021: 1.
- [52] Zhang J, Chen D D, Liao J, et al. Robust Model Watermarking for Image Processing Networks via Structure Consistency[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(10): 6985-6992.
- [53] Wu H Z, Liu G, Yao Y W, et al. Watermarking Neural Networks with Watermarked Images[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, 31(7): 2591-2601.
- [54] Fei J W, Xia Z H, Tondi B, et al. Wide Flat Minimum Watermarking for Robust Ownership Verification of GANs[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 8322-8337.
- [55] Huang Z H, Li B H, Cai Y, et al. What Can Discriminator Do? towards Box-Free Ownership Verification of Generative Adversarial Networks[C]. *2023 IEEE/CVF International Conference on Computer Vision*, 2023: 4986-4996.
- [56] Lin D D, Tondi B, Li B, et al. A CycleGAN Watermarking Method for Ownership Verification[J]. *IEEE Transactions on Dependable and Secure Computing*, 2025, 22(2): 1040-1054.
- [57] Kirchenbauer J, Geiping J, Wen Y X, et al. A Watermark for Large Language Models[EB/OL]. 2023: arXiv: 2301.10226. <https://arxiv.org/abs/2301.10226>.
- [58] Li Z J, Wang C Z, Wang S, et al. Protecting Intellectual Property of Large Language Model-Based Code Generation APIs via Watermarks[C]. *The 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023: 2336-2350.
- [59] Liu A W, Pan L Y, Hu X M, et al. An Unforgeable Publicly Verifiable Watermark for Large Language Models[EB/OL]. 2023: arXiv: 2307.16230. <https://arxiv.org/abs/2307.16230>.
- [60] Fernandez P, Couairon G, Furon T et al. Functional invariants to watermark large transformers[C]. *ICASSP*, 2024: 4815-4819.
- [61] Cao X Y, Jia J Y, Gong N Z. IPGuard: Protecting Intellectual Property of Deep Neural Networks via Fingerprinting the Classification Boundary[EB/OL]. 2019: arXiv: 1910.12903. <https://arxiv.org/abs/1910.12903>.

- [62] Wang S, Abuadba S, Agarwal S, et al. PublicCheck: Public Integrity Verification for Services of Run-Time Deep Models[C]. *2023 IEEE Symposium on Security and Privacy*, 2023: 1348-1365.
- [63] Wang S, Chang C H. Fingerprinting Deep Neural Networks - a DeepFool Approach[C]. *2021 IEEE International Symposium on Circuits and Systems*, 2021: 1-5.
- [64] Cheng H, Li X B, Wang H X, et al. DeepDIST: A Black-Box Anti-Collusion Framework for Secure Distribution of Deep Models[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(1): 97-109.
- [65] Peng Z R, Li S F, Chen G X, et al. Fingerprinting Deep Neural Networks Globally via Universal Adversarial Perturbations[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 13420-13429.
- [66] Li Y C, Zhang Z Q, Liu B Y, et al. ModelDiff: Testing-Based DNN Similarity Comparison for Model Reuse Detection[C]. *The 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2021: 139-151.
- [67] Mehta D, Mondol N, Farahmandi F, et al. AIME: Watermarking AI Models by Leveraging Errors[C]. *2022 Design, Automation & Test in Europe Conference & Exhibition*, 2022: 304-309.
- [68] Pan X D, Zhang M, Lu Y F, et al. TAFA: A Task-Agnostic Fingerprinting Algorithm for Neural Networks[M]. *Computer Security - ESORICS 2021*. Cham: Springer International Publishing, 2021: 542-562.
- [69] Chen J L, Wang J Y, Peng T L, et al. Copy, Right? A Testing Framework for Copyright Protection of Deep Learning Models[C]. *2022 IEEE Symposium on Security and Privacy*, 2022: 824-841.
- [70] Chen M L, Wu M. Protect Your Deep Neural Networks from Piracy[C]. *2018 IEEE International Workshop on Information Forensics and Security*, 2018: 1-7.
- [71] Olney B, Karam R. Protecting Deep Neural Network Intellectual Property with Architecture-Agnostic Input Obfuscation[C]. *The Great Lakes Symposium on VLSI 2022*, 2022: 111-115.
- [72] Lin N, Chen X M, Lu H, 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*, 2021, 40(7): 1327-1339.
- [73] Mukherjee R, Chakraborty R S. Attacks on Recent DNN IP Protection Techniques and Their Mitigation[J]. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023, 42(11): 3642-3650.
- [74] Chakraborty A, Mondai A, Srivastava A. Hardware-Assisted Intellectual Property Protection of Deep Learning Models[C]. *2020 57th ACM/IEEE Design Automation Conference*, 2020: 1-6.
- [75] Li P H, Huang J, Zhang S S. LicenseNet: Proactively Safeguarding Intellectual Property of AI Models through Model License[J]. *Journal of Systems Architecture*, 2025, 159: 103330.
- [76] Fan L X, Ng K W, Chan C S, et al. DeepIPR: Deep Neural Network Ownership Verification with Passports[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(10): 6122-6139.
- [77] Chien T Y, Shen C Y. Customized and Robust Deep Neural Network Watermarking[C]. *The 17th ACM International Conference on Web Search and Data Mining*, 2024: 134-142.
- [78] Chen H L, Fu C, Rouhani B D, et al. DeepAttest: An End-to-End Attestation Framework for Deep Neural Networks[C]. *The 46th International Symposium on Computer Architecture*, 2019: 487-498.
- [79] Li D W, Ren Y K, Liu D, et al. PIPP: A Practical PUF-Based Intellectual Property Protection Scheme for DNN Model on FPGA[J]. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2024, 71(2): 912-916.
- [80] Wang B L, Yao Y S, Shan S, et al. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks[C]. *2019 IEEE Symposium on Security and Privacy*, 2019: 707-723.
- [81] Wang H Q, Xue M F, Sun S C, et al. Detect and Remove Watermark in Deep Neural Networks via Generative Adversarial Networks[EB/OL]. 2021: arXiv: 2106.08104. <https://arxiv.org/abs/2106.08104>.
- [82] Gan Guanhao, Li Yiming, Wu Dongxian, et al. Towards robust model watermark via reducing parametric vulnerability[C]. *ICCV*, 2023: 4751-4761.
- [83] Huang X J, Alzantot M, Srivastava M. NeuronInspect: Detecting Backdoors in Neural Networks via Output Explanations[EB/OL]. 2019: arXiv: 1911.07399. <https://arxiv.org/abs/1911.07399>.
- [84] Li F Q, Wang S L, Liew A W. Linear Functionality Equivalence Attack Against Deep Neural Network Watermarks and a Defense Method by Neuron Mapping[J]. *IEEE Transactions on Information Forensics and Security*, 2023, 18: 1963-1977.
- [85] Yan Yifan, Pan Xudong, Zhang Mi et al. Rethinking white-box watermarks on deep learning models under neural structural obfuscation[C]. *32th USENIX security symposium*, 2023: 2347-2364.
- [86] Pegoraro A, Segna C, Kumari K, et al. DeepEclipse: How to Break White-Box DNN-Watermarking Schemes[EB/OL]. 2024: arXiv: 2403.03590. <https://arxiv.org/abs/2403.03590>.
- [87] Chen Wenlin, Wilson J T, Tyree S, et al. Compressing neural networks with the hashing trick[C]. *ICML*. Lille, France, 2015, 37: 2285-2294.
- [88] Wang R, Li H X, Mu L Z, et al. Rethinking the Vulnerability of DNN Watermarking: Are Watermarks Robust Against Naturalness-Aware Perturbations? [C]. *The 30th ACM International Conference on Multimedia*, 2022: 1808-1818.
- [89] Wang M Q, Qiu H, Zhang T W, et al. Mitigating Query-Based Neural Network Fingerprinting via Data Augmentation[J]. *ACM Transactions on Sensor Networks*, 2023: 3597933.
- [90] Hu H L, Pang J. Stealing Machine Learning Models: Attacks and Countermeasures for Generative Adversarial Networks[C]. *The 37th Annual Computer Security Applications Conference*, 2021: 1-16.
- [91] Wei L X, Luo B, Li Y, et al. I Know What You See: Power Side-Channel Attack on Convolutional Neural Network Accelerators[C]. *The 34th Annual Computer Security Applications Conference*, 2018: 393-406.
- [92] Aiken W, Kim H, Woo S, et al. Neural Network Laundering: Removing Black-Box Backdoor Watermarks from Deep Neural Networks[J]. *Computers & Security*, 2021, 106: 102277.
- [93] Chen X Y, Wang W X, Bender C, et al. REFIT: A Unified Watermark Removal Framework for Deep Learning Systems with Li-

mitted Data[C]. *The 2021 ACM Asia Conference on Computer and Communications Security*, 2021: 321-335.

- [94] Li F Q, Zhao H D, Du W, et al. Revisiting the Information Capacity of Neural Network Watermarks: Upper Bound Estimation and beyond[J]. *Proceedings of the AAAI Conference on Artificial Intel-*

ligence, 2024, 38(19): 21331-21339.

- [95] Tseng S P, Yeh L Y, Hung H J, et al. Federated Watermarking of Deep Neural Networks with Distributed Verification[C]. *The Nineteenth ACM International Conference on Web Search and Data Mining*, 2026: 629-639.



张光华 于 2014 年在西安电子科技大学信息安全专业获得博士学位。现任河北科技大学网络空间安全专业教授。为中国计算机学会 CCF 杰出会员。研究领域为网络与信息安全。研究兴趣包括: Web 安全、漏洞挖掘、安全数据分析。Email: zhanggh@hebust.edu.cn



刘伟发 于 2024 年在河北科技大学计算机科学与技术专业获得硕士学位。现任菏泽学院数学与统计学院助教。研究领域为人工智能安全。研究兴趣包括: DNN 水印。Email: hzxyliuweifa@163.com



刘亦纯 于 2024 年河北科技大学通信与信息系统专业获得硕士学位。现任河北科技大学电气工程学院辅导员。研究领域为网络与信息安全。研究兴趣包括: DNN 水印。Email: 2635261486@qq.com



李曼 于 2025 年在河北科技大学计算机科学与技术专业获得硕士学位。现在中国工商银行石家庄分行工作。研究领域为人工智能安全。研究兴趣包括: 深度学习后门攻防、深度学习产权保护。Email: 115226531260@163.com



贾刘影 于 2021 年在石家庄学院通信工程专业获得学士学位。现在唐山银行工作。研究领域为人工智能安全。研究兴趣为深度学习产权保护。Email: jly15612141340@163.com



王向红 于 2023 年在河北科技大学网络工程专业获得学士学位。现在河北科技大学计算机科学与技术专业攻读硕士学位。研究领域为人工智能安全。研究兴趣为大模型产权保护。Email: wangxianghong2001@163.com



张玉清 于 2000 年在西安电子科技大学获得博士学位。现任中国科学院大学教授, 博士生导师。主要研究方向为网路与信息系统安全。Email: zhangyq@nipc.org.cn