文章编号:1007－6735(2024)03－0225－18　DOI: 10.13255/j.cnki.jusst.20230912001

深度神经网络模型水印研究进展

谭景轩，　钟　楠，　郭钰生，　钱振兴，　张新鹏
（复旦大学 计算机科学技术学院，上海 200433）

摘要：随着深度神经网络在诸多领域的成功应用，以神经网络水印为代表的深度模型知识产权保护技术在近年来受到了广泛关注。对现有的深度神经网络模型水印方法进行综述，梳理了目前为了保护模型知识产权而提出的各类水印方案，按照提取水印时所具备的不同条件，将其分为白盒水印、黑盒水印和无盒水印 3 类方法，并对各类方法按照水印嵌入机制或适用模型对象的不同进行细分，深入分析了各类方法的主要原理、实现手段和发展趋势。然后，对模型水印的攻击方法进行了系统总结和归类，揭示了神经网络水印面对的主要威胁和安全问题。在此基础上，对各类模型水印中的经典方法进行了性能比较和分析，明确了各个方法的优势和不足，帮助研究者根据实际的应用场景选用合适的水印方法，为后续研究提供基础。最后，讨论了当前深度神经网络模型水印面临的挑战，并展望未来可能的研究方向，旨在为相关的研究提供参考。

关键词：深度神经网络；知识产权保护；神经网络水印；白盒水印；黑盒水印；无盒水印；水印攻击；模型安全
中图分类号：TP 309.2　　**文献标志码：**A

Review of watermarking for deep neural networks

TAN Jingxuan,　ZHONG Nan,　GUO Yusheng,　QIAN Zhenxing,　ZHANG Xinpeng
(*School of Computer Science, Fudan University, Shanghai 200433, China*)

Abstract: With the successful application of deep neural networks（DNN）in many fields, deep model intellectual property protection technologies represented by neural network watermarking have received widespread attention in recent years. An overview of existing DNN model watermarking methods was provided in this paper. According to the different conditions required for extracting watermarks, watermarks were classified into three categories: white-box watermarking, black-box watermarking, and

收稿日期: 2023−09−12
基金项目: 国家自然科学基金资助项目（U20B2051, 62072114, U20A20178, U22B2047）；国家重点研发计划资助项目（2023YFF0905000）
第一作者: 谭景轩（2000−），男，硕士研究生．研究方向：神经网络水印．E-mail：tanjx21@m.fudan.edu.cn
通信作者: 钱振兴（1981−），男，教授．研究方向：多媒体安全与人工智能安全．E-mail：zxqian@fudan.edu.cn

引文格式: 谭景轩，钟楠，郭钰生，等．深度神经网络模型水印研究进展[J]．上海理工大学学报，2024，46(3)：225-242．DOI：10.13255/j.cnki.jusst.20230912001
Citation: TAN Jingxuan, ZHONG Nan, GUO Yusheng, et al. Review of watermarking for deep neural networks[J]. Journal of University of Shanghai for Science and Technology, 2024, 46(3)：225-242.

box-free watermarking. Furthermore, various methods were categorized according to different watermark embedding mechanisms or applicable model objects, and an in-depth analysis was conducted on the main principles, implementation approaches, and development trends of these methods. Subsequently, a systematic summary and classification of attack methods on model watermarking were provided, revealing the main threats and security issues faced by neural network watermarking. On this basis, performance comparison and analysis were conducted on representative methods in each category of model watermarks, which clarified their advantages and disadvantages to help researchers choose appropriate watermark methods based on actual application scenarios. Finally, the challenges of current deep neural network model watermarking were discussed, and potential future research directions were envisioned to provide references for related research.

Keywords: *deep neural network*; *intellectual property protection*; *neural network watermarking*; *white-box watermark*; *black-box watermark*; *box-free watermark*; *watermark attack*; *model security*

近年来，得益于数据的海量增长和计算能力的快速发展，以深度学习[1]为代表的人工智能技术在计算机视觉、自然语言处理、语音识别、自动驾驶、智能医疗和金融风控等诸多应用领域取得了令人瞩目的成就。作为深度学习的主要模型，深度神经网络（deep neural network, DNN）通过多层的神经元结构和强大的学习能力，能够从大量数据中学习和提取关键特征，实现高度准确的预测。然而，训练一个性能优异的神经网络代价高昂，不仅需要庞大的数据量和计算资源支持，还依赖于专家对模型结构和训练方法的精心设计。因此，经过精心训练构建出的高性能神经网络模型具有极高的商业价值，同时也是模型所有者的脑力劳动产物，必须将其纳入其所有者的知识产权（intellectual property，IP）。未经所有者授权的情况下，对模型进行复制、剽窃、分发或篡改等行为都构成了对知识产权的侵害。出于对利益的考虑，深度学习模型版权所有者迫切需要有效的方案解决知识产权保护的问题。

数字水印[2]在过去的许多年中被用来解决多媒体作品的版权保护和内容认证问题。通过向多媒体作品中嵌入版权所有者身份、购买者身份、日期、序列号等特定的信息，并在需要的时候将信息提取出来，可在发生版权纠纷时证明所有权归属，或者在出现非法泄漏时鉴定其来源。如今，多媒体数字水印的发展已经较为完善，并按照载体类型的不同发展出图像水印[3]、视频水印[4]和音频水印[5]等多种水印技术。

受多媒体水印所启发，Uchida等[6]首次提出了神经网络模型水印的概念，即向模型中嵌入水印信息，当模型被非法复制后可通过从中提取出水印以证明模型的所有权归属。随后，研究人员对模型水印展开了广泛的研究，在开发新水印算法的同时也对水印的应用范围不断拓展。

除了水印之外，研究人员也提出了一些其他的模型所有权验证方法。由于水印需要修改模型参数而不可避免地引起一定的性能损失，限制了模型水印在医疗、金融等关键领域的应用，模型指纹技术[7]被提出。该技术可以在不修改模型的前提下利用模型的内在属性构造一组能够描述模型决策边界的特殊样本，即对抗样本[8]。由于盗版模型的决策边界相比于原始模型不会发生明显变化，可以根据可疑模型能否在这组对抗样本上产生与原模型一致的预测来检测盗版模型。同样是不修改模型，神经网络模型哈希[9]则借鉴了传统多媒体领域中的感知哈希算法，基于模型参数为模型生成一串哈希序列，并且对模型进行一定处理后，哈希序列的变化在可接受的范围内，可以通过比对哈希序列检测盗版模型。此外，模型哈希还能够用于在一个大规模模型库中检索一个目标模型。由于模型水印、模型指纹与哈希在本质机理上有所不同，本文仅对模型水印进行综述。

目前，国内外已有一些综述[10-16]对现有的模型水印方法作了一定的梳理和总结。其中，文献[10-14]只包含了2021年及之前的模型水印研究相关的文献。文献[15]是对机器学习模型IP保护的综合性调研，描述了较为完整的威胁模型，并系统地总结了模型水印、模型指纹和模型访问控制3种不同类型的模型IP保护技术。文献[16]从白

盒水印、黑盒水印、无盒水印 3 个角度梳理了国内外研究现状，并对脆弱水印进行了分析和讨论。本文的组织结构和文献 [16] 较为类似，同样从白盒水印、黑盒水印、无盒水印 3 个方面展开了综述，不同之处在于本文对每种类型的水印都进行了更加细致的划分，并讨论了当前备受关注的大语言模型和扩散模型的水印。此外，与现有综述不同的是，本文还提供了对水印攻击方法的系统梳理，总结出模型修改、输入检测及修改、模型提取和伪造攻击 4 种类型的水印攻击方法。在此基础上，分析并比较了典型的神经网络水印算法，对比了各个方法的优势和不足，以帮助研究者根据实际的应用场景选用合适的水印方法。

1　神经网络水印概述

1.1　神经网络水印定义

神经网络水印算法涉及水印嵌入过程和水印提取过程，整体框架如图 1 所示。水印嵌入过程将有可鉴别性的信息 I 嵌入神经网络模型中作为版权标识，且不影响模型的可用性，嵌入过程中使用密钥 K 以确保水印仅能够被合法所有者提取。

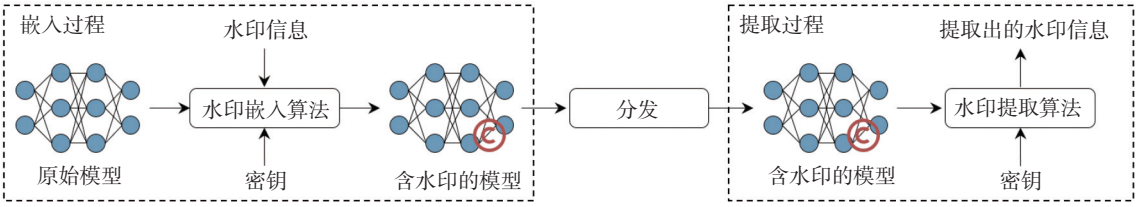


图 1　神经网络水印框架

Fig.1　Framework of neural network watermarking

含有水印的模型在使用或分发中可能会被盗版者通过某种手段获取，盗版者甚至会采取一些模型处理方法试图去除模型中的水印，从模型安全的角度来看，盗版者也被称为攻击者。

当模型合法所有者想要验证一个可疑模型是否盗版于自己的模型时，用嵌入水印时的秘钥从可疑模型中提取水印信息 I'，将 I' 与原水印信息 I 比较相似度，当相似度足够高时水印可以作为所有权证明的有力依据。

1.2　神经网络水印性能要求

表 1 总结了神经网络水印最常见的性能要求。需要注意的是，表中的术语在现有文献中并不统一，且对于相同要求也会有一些同义词，在阅读具体文献时需要结合上下文判断其含义。例如，此处的“完整性”沿用了文献 [11] 中的术语，指代水印验证过程的虚警率应该尽可能低，而模型的完整性强调的是模型未受到微调、剪枝等篡改。

表 1　神经网络水印性能要求

Tab.1　Performance requirements for DNN watermarking

要求	描述
保真度	水印嵌入后不能降低模型在原始任务上的精度
鲁棒性	嵌入的水印应该能够抵抗各种类型的攻击
不可伪造性	不能从模型中伪造出额外的水印从而慌称所有权
隐蔽性	含水印的模型和不含水印的模型应该难以区分
完整性	水印验证过程的虚警率应该尽可能低
可靠性	水印验证过程的漏检率应该尽可能低
效率	水印嵌入和提取的计算代价应该尽可能低
容量	水印方法应当具有嵌入大量信息的能力
通用性	水印方法能够应用于多种类型的神经网络模型

1.3　神经网络水印分类

现有的神经网络模型水印方法可以根据不同的标准进行分类。按照提取水印时对于模型的访问权限，可以将模型水印分为白盒水印、黑盒水印和无盒水印 3 类。如图 2 所示，白盒水印在提取水印时需要访问目标模型的内部结构和参数；黑盒水印在提取时则不需要掌握目标模型的内部细节，通过模型应用程序接口（application program interface, API）进行输入输出查询即可验证所有权；无盒水印则完全不需要访问模型，水印是从模型的输出数据中提取的。

除此之外，按照水印的嵌入机制，可以将模型水印大致分为基于参数或结构的、基于触发集的和基于输出嵌入的 3 类。按照嵌入水印的容量，可以将模型水印划分为多比特水印和零比特水印。根据水印提取是否需要特定样本触发，还可以将模型水印划分为动态水印和静态水印；根

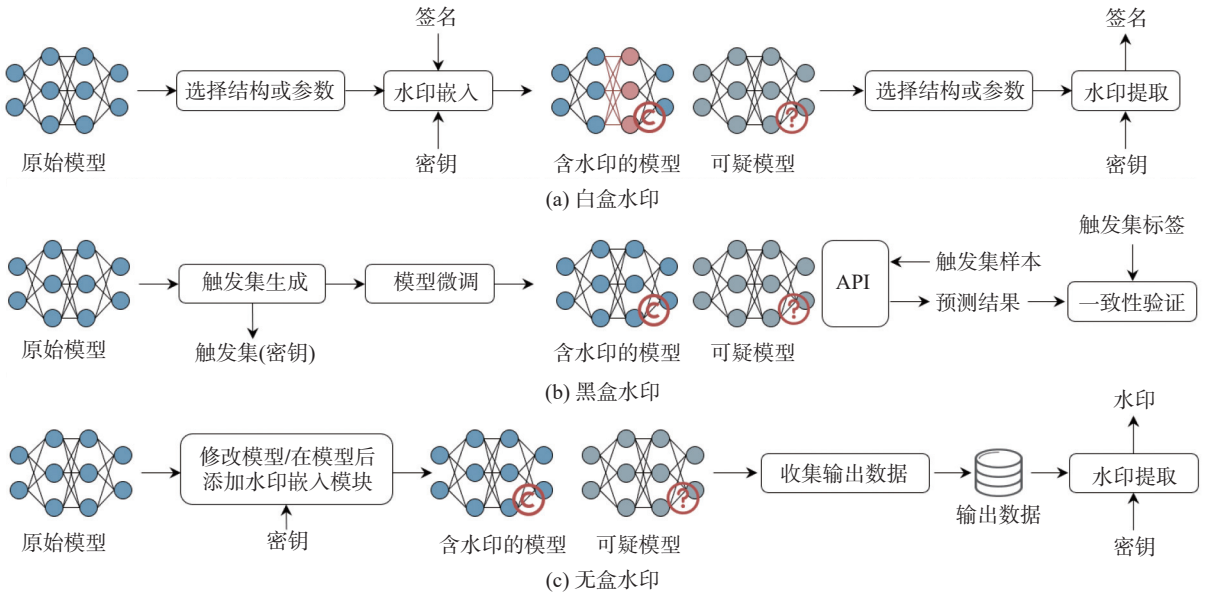


图 2 白盒水印、黑盒水印和无盒水印

Fig.2 White-box watermark, black-box watermark, and box-free watermark

据模型类型，有分类模型水印、生成式模型水印、自监督学习模型水印等。由于根据提取水印时的场景分类是最普遍的分类方法，本文从白盒水印、黑盒水印和无盒水印 3 个角度梳理和总结现有的方法。

1.4 脆弱水印

需要指出的是，上述提到的水印都有鲁棒性的要求，即水印在经过攻击之后依然能够被正确提取，以验证模型的所有权。在某些场景下，模型所有者需要对模型的完整性进行校验，即判断模型是否被恶意篡改。为此，研究人员提出了脆弱模型水印技术，嵌入水印后，即使模型发生轻微变化也将导致水印被破坏。脆弱模型水印也有黑盒和白盒之分。黑盒脆弱水印[17-20]的基本思想是设计一组特殊的、紧靠模型决策边界的样本，使得模型在经过轻微修改后，预测结果即轻易发生变化。白盒脆弱水印[21-22]将脆弱水印嵌入到模型参数中，能够定位到被篡改的参数块，甚至恢复被篡改的参数。特别地，可逆水印[23]还能够实现在提取水印的同时恢复原始模型的参数。本文主要关注鲁棒模型水印，对于脆弱水印较为全面的调研可以参考吴汉舟等[16]的综述。

2 白盒水印

白盒水印在提取水印时需要把神经网络模型当做白盒对待。根据水印嵌入的位置，本文将白盒水印划分为基于内部权重的方法、基于内部结构的方法和其他方法。

2.1 基于内部权重的方法

基于内部权重的方法将水印嵌入在模型权重中。Uchida 等[6]首次将水印应用于神经网络模型，并提出了一种基于权重正则化的水印框架，水印的嵌入通过使用如下的损失微调模型来实现：

$$\mathcal{L}(\boldsymbol{W})=\mathcal{L}_0(\mathcal{X},\mathcal{Y};\boldsymbol{W})+\lambda\mathcal{L}_{\mathrm{R}} \tag{1}$$

式中：$\mathcal{L}_0$表示原始任务的损失函数，例如对于分类任务，$\mathcal{L}_0$为交叉熵损失；$\boldsymbol{W}$为模型参数；$\mathcal{X}$是输入数据；$\mathcal{Y}$是标签；$\mathcal{L}_{\mathrm{R}}$是权重正则化项；λ是平衡两项损失的超参数。式 (1) 为白盒水印提供了一个范式，后续很多算法都采用这种形式嵌入水印，其区别只在于$\mathcal{L}_{\mathrm{R}}$的设计不同。

记$\boldsymbol{b}\in\{0,1\}^L$表示待嵌入的长度为$L$比特的位串；$\boldsymbol{K}\in\mathbb{R}^{L\times M}$是一个伪随机生成的秘钥参数矩阵；$\boldsymbol{w}_l$是模型第$l$层的按卷积核求平均的权重向量；$\sigma$表示 Sigmoid 激活函数，$\mathcal{L}_{\mathrm{R}}$定义为$\sigma(\boldsymbol{K}\boldsymbol{w}_l)$与$\boldsymbol{b}$之间的二元交叉熵损失：

$$\mathcal{L}_{\mathrm{R}}=\boldsymbol{b}\log(\sigma(\boldsymbol{K}\boldsymbol{w}_l))+(1-\boldsymbol{b})\log(1-\sigma(\boldsymbol{K}\boldsymbol{w}_l)) \tag{2}$$

Chen 等[24]基于抗共谋码[25]在每个模型复件中嵌入不同的用户指纹，当发现未授权的复件时可通过检索指纹追溯来源，并能够抵抗共谋攻击。然而，直接将水印嵌入在模型某一层权重中的方法[6,24]无法抵抗水印重写攻击，在这种攻击中，攻击者如果知晓了承载水印的层的位置，可以在相同

层嵌入新水印的同时抹除原有的水印。Wang 等[26]的工作表明，根据权重分布差异推导出水印位置并进行重写攻击是可行的。为了解决这个问题，Feng 等[27]提出了具有补偿机制的水印方案，先对原始水印扩频调制，然后将水印分散地嵌入在模型各层伪随机选择的权重中，且水印嵌入过程不需要微调，而是对权重进行正交变换后通过二值化方法在系数中嵌入水印，最后再通过微调模型弥补水印嵌入带来的性能损失。

一些工作[28-29]则致力于对 Uchida 等[6]方法中的秘钥参数矩阵进行改进。Wang 等[28]将参数矩阵替换成一个独立的神经网络，它和目标模型一起训练，使得这个网络能够从目标模型中合适的权重中提取出水印。Wang 等[29]提出的 RIGA（robust white-box GAn watermarking）使用一个单独的提取器网络从模型权重中提取水印，通过对提取器网络进行适当设计，不仅可以提取出比特串形式的水印，还支持以 Logo 图像作为水印。该方案还采用对抗学习的思想，引入一个检测器网络以区分不含水印的权重和含有水印的权重，目标模型、提取器网络与检测器网络对抗训练，鼓励含水印的权重不易被检测，提升了水印的隐蔽性。同样以隐蔽性为目标，Kuribayashi 等[30]则是从模型的全连接层采样部分权重参数，利用量化索引调制将水印嵌入到权重的频域系数中，最后微调模型，确保了水印嵌入引起的权重变化足够小。Chen 等[31]为语音识别模型设计了一种谱水印方法，将所有权信息通过扩频方式编码到模型参数的重要频域分量中。

使用参数矩阵或神经网络从模型中提取水印的方法都难以避免伪造攻击，攻击者获取含水印的模型后，可以在不改变模型的情况下构造出新的参数矩阵或神经网络，使得从中能够提取出伪造的水印，造成所有权的歧义。针对这个问题，Liu 等[32]提出了一种不需要显式的所有权指示器。该方法要嵌入的签名是一个由 RSA 算法对原始的版权声明进行私钥加密，再对解码结果取符号得到的，因此$\boldsymbol{b}\in\{-1,1\}^L$，通过 hinge 损失将$\boldsymbol{b}$嵌入到从模型重要参数构造出的残差向量$\boldsymbol{\psi}$的符号中，利用式 (1) 嵌入水印，$\mathcal{L}_{\mathrm{R}}$定义为

$$\mathcal{L}_{\mathrm{R}}=\sum_{i=1}^{L}\mathrm{ReLU}(\mu-\boldsymbol{b}_i\boldsymbol{\psi}_i) \tag{3}$$

该损失会鼓励$\boldsymbol{\psi}$的符号与$\boldsymbol{b}$相同，式中的μ是为了扩大不同符号之间的差异而设置的调节参数。为了构造残差向量$\boldsymbol{\psi}$，先对某一层的权重$\boldsymbol{W}_l$展平，对其运用一维平均池化，结果变形为$\boldsymbol{\gamma}\in\mathbb{R}^{L\times d}$，对$\boldsymbol{\gamma}$的每一行选择绝对值较大部分的值取平均，构造出残差向量$\boldsymbol{\psi}$。该方案的优势在于具有不可伪造性，且对多种水印去除攻击的鲁棒性较好。

2.2　基于内部结构的方法

基于内部结构的方法利用模型的结构承载水印。Fan 等[33-34]针对部分白盒水印存在的无法抵御伪造攻击的不足，提出了一种基于“护照”验证的方法，在原始网络中插入特殊的“护照层”，使得在给出合法护照时网络正常工作，而在给出伪造护照时网络性能明显降低，即模型性能依赖于护照的正确性。如图 3 所示，护照层位于卷积层之后，类似于归一化层，先对卷积后的特征图归一化，再进行缩放和平移。不同之处在于，标准的归一化层中用于缩放和平移的缩放因子$\boldsymbol{\gamma}$和偏置项$\boldsymbol{\beta}$是两个可学习的参数，而护照层的$\boldsymbol{\gamma}$和$\boldsymbol{\beta}$由护照计算得到：输入到第l个卷积层的护照与输入到该层的特征图尺寸相同，通过对护照卷积并进行平均池化后得到护照层的$\boldsymbol{\gamma}$和$\boldsymbol{\beta}$。因此，护照决定了网络的推理性能，加大了攻击者的伪造难度。

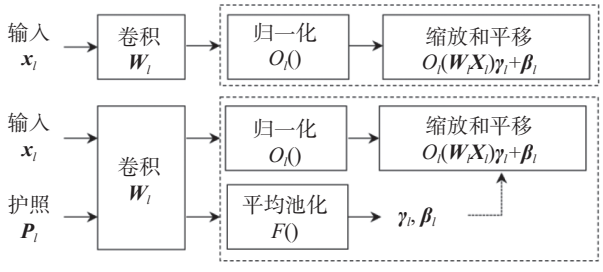


图 3　标准的卷积−归一化层（上）和卷积−护照层（下）

Fig.3　Standard convolution-normalization layer (top) and convolution-passport layer (bottom)

由于批量归一化（batch normalization, BN）层中的归一化操作需要计算训练样本的滑动平均和滑动方差，Fan 等[33-34]的方法不适用于带 BN 结构的神经网络。对此，Zhang 等[35]提出了一种适用于大多数归一化层的改进方案，将护照层作为一个额外的分支，其中归一化操作用到的统计量单独计算，不影响原始的归一化层。模型正常推理时不需要护照层，只有当需要进行所有权验证时才将秘密护照和护照层插入模型。

Zhao 等[36]提出了一种基于网络通道剪枝的结构化水印方法，将水印序列划分为若干比特段，每个比特段量化为剪枝率，再根据剪枝率对模型的卷积层进行通道剪枝。Xie 等[37]则是在权重的级别进行剪枝，在连接灵敏度较低的模型权重中

嵌入水印。

2.3 其他方法

本小节总结除了基于内部权重的方法、基于内部结构的方法以外的其他白盒水印方法。Rouhani 等[38] 提出的 Deepsigns 将水印嵌入到网络中间层的激活图中，能够较好地抵抗水印重写攻击。DeepSigns 首先任意地从一些目标类T中选择一些训练样本$\boldsymbol{x}_T$，计算这些样本在某一层的特征图的均值$\boldsymbol{\mu}$，使用秘钥参数矩阵$\boldsymbol{K}$将水印$\boldsymbol{b}$嵌入其中，使用的正则化项为

$$\mathcal{L}_{\mathrm{R}}=\mathcal{L}_{\mathrm{BCE}}(\sigma(\boldsymbol{K}\boldsymbol{\mu}),\boldsymbol{b})+\beta\sum_{i\in T}\|f_l(\boldsymbol{x}_i)-\boldsymbol{\mu}_i\|-\beta\sum_{i\in T,j\notin T}\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\| \tag{4}$$

其中：第一项类似于式 (2)，表示$\sigma(\boldsymbol{K}\boldsymbol{\mu})$和$\boldsymbol{b}$之间的二元交叉熵损失；第二项中的$f_l$表示神经网络的第$l$层，$\boldsymbol{\mu}_i$是第$i$类样本激活图的均值。由于该方法的所有权验证依赖于一组秘密样本，类似于黑盒水印中用触发样本查询模型 API 的方式，只不过需要获取的不是模型的预测类别而是中间层的激活图，樊雪峰等[13] 将其称为灰盒水印。

Lim 等[39] 将利用激活承载水印的思想拓展到了基于循环网络的图像描述模型，该方案利用符号损失将签名嵌入在循环神经网络输出的隐藏层状态的符号中，使得攻击者无法使用通道重排列去除水印。大多数水印工作都无法从理论上保证对于水印去除攻击的鲁棒性。对此，Lv 等[40] 提出的 HufuNet 通过精心设计一个水印嵌入损失以从理论上保证水印对微调的鲁棒性。HufuNet 首先训练一个用于数据重建的自编码器，然后将编码器参数作为水印嵌入到目标模型中，再交替训练编码器参数和目标模型，以同时确保原始任务和自编码器的性能。训练时所采用的损失能够保证模型在经历微调攻击时，原始任务上的性能下降先于含自编码器参数的子网络的性能下降。水印验证时，将编码器参数从目标模型取出与解码器配对，通过组合出的自编码器的性能验证所有权。

Li 等[41-42] 推导出了一种针对神经网络的线性同性能攻击方式，具体表现为同层神经元顺序的重排布，以及神经元输出的线性放大。这种攻击不改变模型性能，但改变了神经网络中参数的分布、排列，使得几乎所有的白盒水印方法都无法正确提取出所有权信息。为此，Li 等[42] 提出了 NeuronMap 框架，它工作在所有权验证协议层次中，不更改现有白盒水印算法的版权信息生成、嵌入、验证等步骤，仅在注册版权信息时要求注册额外的一组 NeuronMap 触发样本和中间层响应模式，并在进行实际验证前利用 NeuronMap 触发样本进行中间层校准，将神经网络恢复到遭受攻击之前的状态，从而使得白盒水印算法正常运作。

3 黑盒水印

当攻击者将盗版模型部署到云服务端或嵌入时设备中时，模型的合法所有者无法访问盗版模型内部，也就无法通过白盒水印验证所有权。黑盒水印仅需要查询模型的预测接口就可以提取水印，更加适用于真实的应用场景。

3.1 图像分类模型的黑盒水印

图像分类模型是深度学习领域里最基础的模型之一，现有的大多数黑盒水印研究关注于此，且其思想也能够很容易地拓展到其他模型中。本文将图像分类模型的黑盒水印划分为借助对抗样本构造触发集、仅更改标签构造触发集、更改图像和标签构造触发集和其他方法 4 类。其中，基于触发集的水印嵌入一般可以用以下损失微调模型实现：

$$\mathcal{L}(\boldsymbol{W})=\mathcal{L}_{\mathrm{CE}}(\mathcal{X},\mathcal{Y};\boldsymbol{W})+\lambda\mathcal{L}_{\mathrm{CE}}(\mathcal{X}_t,\mathcal{Y}_t;\boldsymbol{W}) \tag{5}$$

式中，等号右侧第一项和第二项分别为训练集和触发集上的交叉熵损失。

3.1.1 借助对抗样本构造触发集

对抗样本攻击[8] 是指在样本上添加不可察觉的细微扰动使得机器学习模型将其错误进行分类。Le Merrer 等[43] 借助对抗样本，提出了首个以黑盒方式验证深度学习模型所有权的方法。具体而言，运用对抗攻击方法向部分样本添加扰动，将能够让模型错误分类的样本称为真对抗样本，而添加扰动后分类结果没有改变的样本称为假对抗样本，再微调模型使得模型对两种对抗样本的分类结果都跟原始的分类结果一致，微调后可以认为模型嵌入了零比特水印。在验证阶段，用这些对抗样本作为触发集输入到可疑模型中，无水印的模型很可能错误分类对抗样本，而含有水印的模型则会对其作出正确预测。

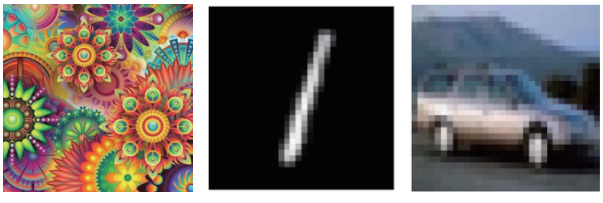
Blackmarks 方法[44] 是首个多比特的黑盒水印方法，其思想与 Le Merrer 等[43] 类似。先对原始预训练模型的 logits 输出运用 k-means 聚类，从而将

所有类标签划分为两组，分别表示比特 0 和比特 1，从某一组中随机选择一些图像运用有目标对抗攻击，使得模型将其预测为另一组中的类，得到的对抗样本作为触发集。为了嵌入水印，用一个额外的损失微调模型以最小化在触发集上预测的组对应比特和待嵌入比特之间的误码率。

3.1.2　仅更改标签构造触发集

模型后门攻击[45]是指攻击者在训练集中掺入一些带特殊触发标记的样本并错误地标注其类别，在推理阶段，模型会对携带触发标记的样本作出错误判断。尽管后门攻击给模型安全带来了威胁，但其背后的思想却启发了黑盒水印算法的设计，许多黑盒水印中的触发集都可以看作是一种良性的后门。

最简单的一种触发集构造方式是收集一组自然图像，仅对标签进行改动。Adi 等[46]提出了首个基于后门的黑盒水印方法，采用来自互联网的抽象图像作为后门触发集，并为每张图像随机选择一个目标标签。图 4(a) 是触发集中的一个样例，其标签被设定为“飞机”。该方案需要在验证过程中引入可信第三方以保证严格的安全性。



(a) Adi等[46]　(b) Zhang等[47]　(c) Namba等[48]

图 4　仅更改标签构造触发集的示例（标签都为飞机）

Fig.4　Examples of trigger samples with only label changed（all labels are airplane）

Zhang 等[47]则提出可以使用与原任务无关的数据作为触发样本（也叫秘钥样本）。以图 4(b) 为例，原任务为 CIFAR10 数据集的分类，使用 MNIST 数据集中的手写数字“1”作为触发样本。与 Adi 等[46]不同的是，这里的触发样本都来自同一类，标签也是统一指定的。

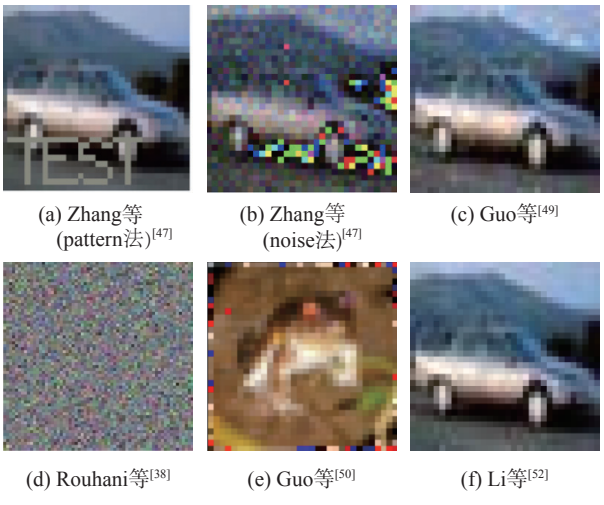
Namba 等[48]首先提出了一种针对现有黑盒水印的查询修改攻击，用一个自编码器判断一个查询是否是触发样本，如果是的话对其用自编码器进行重构后再输入到模型，以使水印验证失效。为了抵抗这种攻击，Namba 等[48]从训练集中随机选择一部分图像，并更改标签作为触发集，如图 4(c) 所示。由于触发样本与训练分布相同，且不含有特殊标记，所以不易被攻击者检测到。为了嵌入水印，设计了一种指数加权方法，识别出对预测有显著贡献的权重参数并对其指数增加，提高了水印的鲁棒性。

3.1.3　更改图像和标签构造触发集

对图像作一定的处理后再更改标签作为触发集是黑盒水印最普遍的一种设计范式。Zhang 等[47]提出在某一个类中的部分训练图像上添加一定的信息，再将其标签改为另一类作为触发集。添加信息的方式有两种：一种是叠加一个有意义的图像作为触发标记，叠加的图像可以是表示版权信息的 Logo 图像；另一种是添加高斯噪声作为触发模式。Guo 等[49]也采用了添加信息的方式构造触发集，将所有者的 n 比特签名嵌入图像中随机选择的 n 个像素，再从该图像所属的类别以外剩下的类中随机选一个作为目标标签。Deepsigns 方法[38]先生成一些随机图像输入预训练模型，将激活分布位于离训练数据较远区域的随机图像保留下来，并分配随机的标签作为触发集。为了降低水印的虚警率，Guo 等[50]设计了一种进化算法对触发模式进行优化。

如图 5(a)～(e) 所示，之前黑盒水印算法中的触发集分布与正常分布存在差异，攻击者可能会采用规避攻击[48,51]，建立检测器躲避所有权验证。为此，Li 等[52]训练一个生成对抗网络构造触发样本，其中生成器将所有者指定的 Logo 图像隐藏进训练图像，判别器判断样本中是否藏有秘密的 Logo，提高了触发样本与正常样本的不可区分性。图 5(f) 是该算法生成的一个触发样本示例。Li 等[53]则在图像频域中嵌入水印以构造触发集，该方法也具有一定的隐蔽性。

大多数基于触发集的水印不具备对模型提取攻击[54]的鲁棒性，在这种攻击中，攻击者首先收集或合成一个数据集，用其中的样本依次查询受害者模型，将预测结果作为标签，训练一个功能与受害者模型相似的替代模型。含水印的模型具有执行原始任务和水印任务的能力，但是攻击者只会用来自任务分布的数据查询模型，所以得到的替代模型中不会包含水印。Jia 等[55]首先分析出这种现象背后的原因在于原始任务和水印任务由不同的神经元学习得到，于是提出了一种纠缠水印嵌入算法，使用软最近邻损失（soft nearest neighbor loss, SNNL）将触发样本与正常样本在特征空间纠缠到一起，确保两种数据由相同的神经元



(a) Zhang等(pattern法)[47] (b) Zhang等(noise法)[47] (c) Guo等[49]

(d) Rouhani等[38] (e) Guo等[50] (f) Li等[52]

图 5 更改图像和标签构造触发集的示例(标签都为飞机)

Fig.5 Examples of trigger samples with both image and label changed(all labels are airplane)

激活，因此，可以抵抗模型提取攻击。SNNL的定义如下：

$$\mathcal{L}_{\mathrm{SNNL}}(\mathcal{X},\mathcal{Y},T)=-\underset{x_i\in\mathcal{X}}{\mathrm{mean}}\left\{\log\left(\sum_{j\neq i,y_i=y_j}\mathrm{e}^{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/T}\Big/\sum_{k\neq i}\mathrm{e}^{-\|\boldsymbol{x}_i-\boldsymbol{x}_k\|^2/T}\right)\right\} \tag{6}$$

式中：$\mathcal{X}$为需要被纠缠的数据构成的集合，在这里由触发样本和目标类的干净样本在网络中间层的激活构成；$\mathcal{Y}$为它们的类别标签；T为温度参数。

该损失表示纠缠度，训练期间需要对其最大化。触发样本的生成方式是首先对目标类中的样本添加触发模式，添加的位置位于SNNL损失梯度最大的区域，再添加对抗扰动以增大交叉熵损失和SNNL损失。Tan等[56]提出了一种基于影子模型的触发集嵌入方法(symmetric shadow model based watermarking，SSW)，使用一个正影子模型模拟攻击者通过模型提取攻击得到的替代模型，再用另一个负影子模型模拟不含水印的干净模型。在嵌入水印的同时，主动优化触发集中的样本，使得它们更容易被原始模型和正影子模型预测为目标标签，而负影子模型中的预测结果与目标标签不一致。这种主动优化触发样本的方法使得水印更容易迁移到替代模型中，并且对包含跨模型结构在内的多种模型提取攻击都有较好的鲁棒性。

Li等[57]提出了一种基于嵌入外部特征的所有权验证方案MOVE(model ownership verification，MOVE)，也能有效抵抗模型提取攻击。MOVE利用风格迁移嵌入外部特征，让模型在原始图像和一部分经过风格迁移后的变换图像上学习。然后，将模型在干净图像和变换图像上的预测距离向量作为输入，训练一个区分无水印模型和含水印模型的元分类器。发现可疑模型时，根据元分类器的结果基于假设检验进行所有权验证。该方案还包括了一个白盒验证方法，区别在于使用变换图像的损失将模型参数的梯度符号作为元分类器的输入，然而这种方法无法在替代模型结构与原始模型结构不同时使用。

在水印去除攻击上，如何为基于触发集的水印鲁棒性提供理论性保证一直是一个难题。因此，Bansal等[58]借鉴随机平滑中的理论，提出了一种可验证的黑盒模型水印方案。该方案使用干净样本和触发样本交替训练模型，当用触发样本更新模型时，采样多个高斯噪声添加到模型参数上，得到多个带噪声的模型，对这些模型在触发集上的损失梯度求平均来更新目标模型的参数。该方案从理论上保证了模型参数在一定l_2范数约束下发生改变，水印不会被去除。

3.1.4 其他方法

本节介绍几种不基于触发集嵌入的其他黑盒水印方法。针对模型提取攻击，Szyller等[59]设计了一个可添加到目标分类器之后的DAWN模块。其原理是对于来自任意某个用户的查询，都动态地返回一部分错误的输出结果，这一部分查询和对应的错误结果作为该用户的水印。该方案假定了攻击者对模型预测结果没有任何先验，即无法区分返回结果是正确还是错误的，所以会把模型所有返回结果作为标签训练替代模型，水印便可直接迁移到替代模型中。水印与用户身份是关联的，当模型所有者发现某个用户发布了一个可疑模型时，用特定于该用户的触发集验证其中是否包含水印。

Charette等[60]考虑了一种更加复杂的模型提取攻击方式，采用集成蒸馏，用多个教师模型输出的平均值来训练一个学生模型，这种攻击会使基于比较触发样本硬标签的水印失效。为抵抗集成蒸馏，进一步提出了CosWM(cosine model watermarking)算法，将具有不同预设频率的余弦信号嵌入多个目标教师模型对训练样本某一特定类的软预测中。在对不同频率的信号相加或相乘后，每个信号的频率依然可以从合成信号的频谱

中提取出来，因此难以通过对教师模型的输出取平均以抹除水印。

Li 等[61]提出了一种不依赖于特定触发样本的多比特黑盒水印方法，将模型的 Softmax 输出经过幂函数变换成更平滑的分布，用密钥对其投影后在结果中承载水印。验证阶段，水印可以从任意输入在网络中的输出中提取。

3.2　其他模型的黑盒水印

本小节介绍针对图像分类模型以外的模型而设计的黑盒水印方法。尽管目标模型结构和要执行的任务是多样的，许多算法的思想也都遵从将一组触发样本输入模型，根据预测行为判断水印的存在。

Quan 等[62]提出了首个用于图像处理网络的黑盒水印算法，对模型微调使得其能够将触发图像映射为验证图像，以随机图像作为触发图像，验证图像是用一个与目标模型功能类似的函数对触发图像处理后的结果，并在水印嵌入过程中更新验证图像。该方法还开发了一个辅助模块将验证图像转化为有视觉意义的版权图像，提高验证的置信度。

Ong 等[63]提出了一个针对生成对抗网络(generative neural network，GAN)的知识产权保护框架，也支持白盒和黑盒两种验证方式。其中，白盒水印的原理是将签名嵌入归一化层的缩放因子符号中，具有较好的鲁棒性和不可伪造性。黑盒水印的构造方式如图 6 所示，在原始输入中添加一个噪声块，并强制生成器对这样的输入产生一个带有 Logo 图像的输出，验证时，检测触发样本对应的输出中是否包含 Logo 图像。

EWE 方法[55]设计了两种针对语音分类模型的

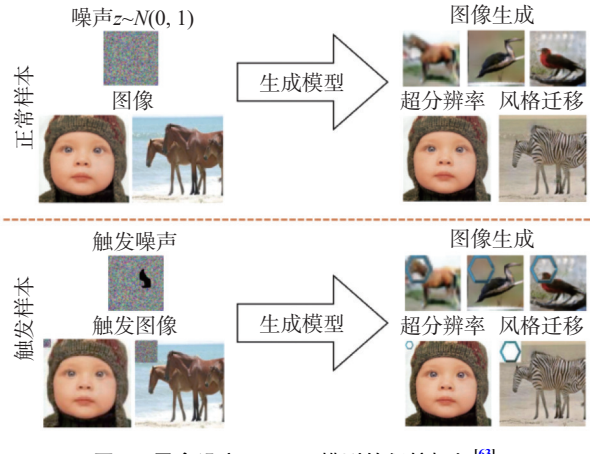


图 6　黑盒设定下 GAN 模型的保护框架[63]

Fig.6　GAN protection framework in the black-box settings

触发样本构造方式：一种是用正弦信号重写音频样本的一部分；另一种是在梅尔频谱的边缘添加两个小方块。针对自动语音识别模型，Chen 等[64]将模型所有者的语音片段传播到整个输入音频上合成触发音频，并用文本隐写术将所有者信息隐藏到目标标签中。Zhang 等[65]为声纹识别模型设计了一种基于后门的水印方案，将密钥信息嵌入到声纹的梅尔频谱图中，生成难以检测和去除的触发样本。

Yadollahi 等[66]为文本处理的模型设计了一种黑盒水印框架，其触发集生成过程如下：首先从训练集中选择一些文档样本，计算所有文档中每个单词的 TF-IDF 分数，对每个选择的文档，再从其他类中随机选择一个文档交换它们的单词以生成水印记录。选择两个文档中 TF-IDF 分数最低的单词进行交换，最后将两个文档的标签交换，插入触发集。

为了保护图神经网络的产权，Zhao 等[67]随机生成了一个具有随机节点特征向量和标签的 Erdos-Renyi 随机图作为触发集，将其与正常样本一起训练，有效地将水印嵌入在图节点的预测中。Lim 等[39]针对图像描述模型设计了一种触发集，触发样本构造方式为在正常样本上添加一个小方块，其标签修改为一条固定的表示所有权信息的文本。

4　无盒水印

无盒水印是指在水印提取时，不需要访问深度学习模型内部，也不需要查询模型预测 API 的一类水印方法。无盒水印通常针对以图像或文本为输出的模型而设计，使得模型输出的图像或文本中包含所有者的水印。当模型被复制后，盗版模型的输出依然包含水印，因此，可以通过收集可疑模型的输出数据并试图从中提取水印以判断其是否为盗版。

4.1　图像生成模型的无盒水印

Wu 等[68]提出了一个以图像为输出的模型的水印框架。如图 7 所示，该框架额外地包含一个水印提取网络和一个可选的判别器网络，与目标网络一同训练。训练目标为：目标网络能够正常执行其原始任务，输出的图像中含有不可感知的水印；水印提取网络能够在给定含水印图像和正确秘钥时提取出预设的水印图像，而在给定无水印

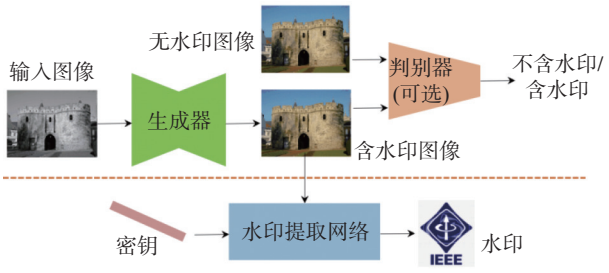


图 7 Wu 等[68] 提出的水印框架

Fig.7 Algorithm framework proposed by Wu et al. [68]

图像或错误密钥时输出全空白的图像；判别器试图区分含水印图像和无水印图像，以进一步提升水印的隐蔽性。训练结束后只发布目标网络，水印提取网络和密钥则保密，只在验证水印时使用。

Zhang 等[69-70] 考虑了模型提取攻击对水印的影响，设计了一种抗提取攻击的图像处理模型水印算法。算法包括初始训练和对抗训练两个阶段。在初始训练阶段，训练一个嵌入子网络将 Logo 图像隐藏进目标模型输出图像中，训练一个提取子网络从水印图像中提取出 Logo，并确保从无水印图像中无法提取水印。在对抗训练阶段，使用目标模型的输入输出对作为监督来训练一个替代模型，以模拟攻击者的替代模型，再用替代模型的输出微调提取子网络，使得替代模型输出的图像中依然能够提取水印。对抗训练的引入提高了水印对替代攻击的鲁棒性。

为了实现 GAN 生成的虚假人脸图像的溯源，Yu 等[71] 预先训练一个编码器和解码器对，分别用来向人脸图像中嵌入比特序列和提取比特序列。然后用编码器为训练集中的所有人脸图像都嵌入指纹，用含指纹的图像训练目标 GAN 模型，并通过实验表明了训练图像中的指纹可以迁移到 GAN 生成的图像中，用预训练的解码器可以检测指纹。Fei 等[72] 对该算法进行了改进，在训练 GAN 的阶段显式地引入一个图像处理层和解码器，使得图像经过处理后指纹也能被提取，提高了指纹对于图像处理操作的鲁棒性。Zhao 等[73] 将类似的思想运用到了无条件的以及带类别条件的扩散模型中，使得水印可以从扩散模型输出的图像中被检测到。在生成模型的训练数据中嵌入水印的局限性在于，每更换一次指纹代码都需要重新训练生成模型。为此，Yu 等[74] 提出了一种用指纹调制生成器参数的方案，仅训练一次生成模型就可以得到海量的嵌有不同指纹的生成器。

4.2 文本生成模型的无盒水印

Abdelnabi 等[75] 提出了首个端到端的文本水印框架 AWT，如图 8 所示。AWT 包含一个隐藏网络和一个揭示网络，隐藏网络将二进制消息嵌入输入文本中得到含水印文本，揭示网络从含水印文本中提取出二进制消息。隐藏网络与揭示网络在训练时与一个判别器对抗训练，以确保含水印文本的统计特性不变，并在训练后进一步微调以确保语义不变性和语法正确性。文本生成模型的服务提供者可以用这种技术向模型的输出中添加水印，用隐藏网络在不同用户的输出文本中嵌入不同的水印序列，当生成文本被用来创作虚假文章或不实信息时，可以用揭示网络从中提取水印，追溯文本来源。

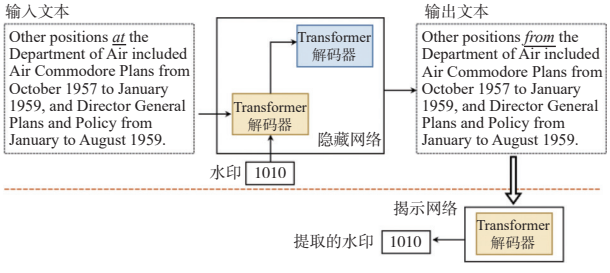


图 8 Abdelnabi 等[75] 提出的文本水印流程图

Fig.8 Flow chat of the text watermarking proposed by Abdelnabi et al. [75]

He 等[76] 探索了在模型可能受到提取攻击的场景下使用水印对文本生成 API 进行保护，在目标文本生成模型后添加一个水印模块，用于对输出文本添加水印，并将含水印的文本返回给用户。然而，He 等[77] 发现，这种水印方法可能会改变词汇分布，攻击者可以通过统计候选水印词汇的频率变化来推断带水印的单词。为了解决这个缺点，He 等开发了一种更隐蔽的水印方法来保护文本生成 API，设计了一种优化方法来确定水印规则，该规则可以最大限度地减少整体单词分布的失真，同时最大限度地增加条件单词选择的变化。该方法从理论上证明了攻击者无法基于统计检查从大量潜在单词对中发现用到的水印，且在替代模型结构不匹配和跨域攻击的情况下依然有效。

最近，大语言模型（以下简称大模型）由于其出色的文本生成能力受到了前所未有的关注，其生成的高质量文本已经达到甚至超过了人类的写作水平。随着大模型的普及，生成文本被滥用的风险也随之增加，通过在大模型生成的文本中嵌入水印逐渐成为对大模型生成文本进行检测和监

管的重要途径。在大模型输出文本中嵌入水印主要有后处理式和整合式两种方式：后处理式指的是在模型生成文本后再添加水印；整合式指的是在生成文本的过程中嵌入水印。后处理式[78-79]往往是借助掩码语言模型(如 BERT 模型、RoBERTa 模型)对个别 token 进行同义词替换，优点是不依赖于大模型，可以作为附加组件插入到任何大模型的输出后。不足之处在于仅仅是对个别 token 进行替换，无法像整合式的方法那样影响后续整个序列的生成，从而导致带水印文本的自由度和前后文一致性严重受到限制。

整合式水印中最具有代表性的方法由 Kirchenbauer 等[80]提出，其通过改变大模型输出时的采样过程来使得文本承载水印。具体而言，当大模型生成当前 token 时，它首先基于先前 token 的哈希值创建一个随机种子，并以该随机种子为条件将整个词汇表拆分为两部分(即绿色列表和红色列表)，然后，将绿色列表中 token 的 logits 添加一个小的正值，使绿色列表中的 token更有可能被采样。在水印检测阶段，使用相同的哈希函数来确定文本中红色列表和绿名列表的 token 数量，从而确定文本是否包含水印信息。后续，Kirchenbauer 等[81]对随机种子的生成方式和检测方法作了改进，验证了水印在人工重写、机器改写、水印文本混合到手写文档中后的鲁棒性。Zhao 等[82]对之前的方法作了简化，使用固定的绿色和红色列表，结果表明其对于文本修改的鲁棒性有了较大的提升。Christ 等[83]提出了一种基于密码学的不可检测的语言模型水印方法，水印无法被用户察觉，只有使用密钥才能检测到水印的存在性。Wang 等[84]提出了可编码水印的概念，在文本生成的过程中，通过重新定义每一步的目标函数，能够使得输出文本中嵌入多比特信息。

5　水印攻击方法

神经网络水印技术在不断发展的同时，针对神经网络水印的攻击算法也应运而生。当获取一个含水印的模型后，攻击者的目标是以一定的手段使水印验证失效，且攻击行为不影响模型的性能，攻击的计算代价尽可能低。本节总结了现有的水印攻击方法，并将其分为模型修改、输入检测及修改、模型提取和伪造攻击 4 大类。

5.1　模型修改

模型修改攻击是指攻击者通过修改含水印模型的参数或结构以达到抹除其中的水印的目的。本小节将模型修改攻击进一步分为模型微调、模型压缩和功能等效攻击 3 类。

5.1.1　模型微调

模型微调允许攻击者使用少量样本微调模型，试图保持模型性能的同时去除水印，是最常见的一种攻击方式。Adi 等[46]提出的 4 种模型微调方式基本囊括了所有的微调形式，被许多模型水印文献采用，作为评估水印鲁棒性的基本攻击基准。这 4 种攻击方式包括：a. FTLL(fine-tune last layer)，微调时仅更新最后一层的参数；b. FTAL(fine-tune all layers)，微调时更新所有层的参数；c. RTLL(re-train last layers)，重新初始化最后一层后，更新最后一层的参数；d. RTAL(re-train all layers)，重新初始化最后一层后，更新所有层的参数。需要注意的是，RTLL 和 RTAL 会对基于触发集的水印产生较大的影响，这是因为分类器的最后一层对于识别水印触发模式具有重要的作用，Adi 等[46]提出的 RTLL 和 RTAL 微调后的模型，在提取水印时先用原始模型的最后一层替换微调后的最后一层。

水印重写攻击本质上也是通过微调来实现的，与普通微调的区别仅在于攻击者掌握了水印算法，试图在微调过程中嵌入新的水印以覆盖原有水印。

攻击者也可能对盗取的模型采用迁移学习，并通过新的数据集对模型进行微调，构建一个用于新任务的模型。迁移学习通常会使所有者无法在黑盒场景下验证所有权，例如在图像分类模型中，迁移学习改变了模型输出类别，这导致无法在给定触发样本时响应预设的触发预测。

5.1.2　模型压缩

模型压缩是一种用来减小模型大小和计算复杂度的技术，便于模型在资源受限的场景下实现高效的推理和部署。攻击者可能会使用模型压缩技术破坏模型中的水印。现有的模型水印文献中常用权重剪枝、权重量化来评估水印的鲁棒性。权重剪枝的经典方法是将模型中绝对值较小的部分参数置 0，这是因为绝对值较小的参数对模型性能影响较小。权重量化是指将模型参数表示为低精度的格式，如 8 位或更低位的整数，从而减小存储需求。目前的水印算法都显示出对模型压缩

良好的鲁棒性。

5.1.3 功能等效攻击

功能等效攻击是最近被研究人员提出来的一种针对白盒水印的攻击，通过适当调整模型参数或结构，可以在模型性能无损的条件下实现水印消除。这种攻击简单高效，不需要访问数据集，不需要再对模型进行训练，也不需要知道关于水印算法的任何知识。Lv 等[40]探索了3种类型的功能等效攻击。

Yan 等[85]提出了一种神经元结构混淆攻击，通过向模型中插入一组不影响正常推理行为的伪造神经元，阻碍水印信息的正确提取。具体来说，设计了两种伪造神经元的构造方式：NeuronClique 方法通过插入一组输入参数相同、输出参数之和为0的神经元，来抵消它们对该层输出结果的影响；NeuronSplit 方法则对某个神经元进行拆分，将其替换为功能不变的一组神经元，并保持伪造神经元的参数分布不变。

5.2 输入检测及修改

输入检测及修改是仅针对基于触发集的水印的一类攻击。由于触发样本通常与正常样本有着不同分布，攻击者可以利用这一特性检测出触发样本，即规避攻击。Hitaj 等[51]提出了两种规避攻击方式：一种是用多个模型的集成投票结果决定查询是否包含水印；另一种是在含水印模型的特征提取器之后训练一个二分类器以检测触发样本。Namba 等[48]发现带一定模式的触发样本经过自编码器重构之后的重构误差比正常样本的重构误差大，提出了使用自编码器来检测含水印的查询方法。

输入修改试图破坏触发样本中的触发模式，使模型返回与其目标标签不一致的结果。以模型的输入是图像为例，常见的输入修改操作包括用自编码器重构图像、添加噪声、输入量化、图像平滑、翻转、JPEG 压缩等[86]。然而，大多数水印方法都显示出对这些基本图像处理操作的鲁棒性。Wang 等[87]提出了一种基于添加自然感知扰动的输入修改攻击，首先估计既能破坏水印模式，又不影响模型正常功能的目标光照条件，然后再添加阴影进行进一步渲染。该攻击能够成功地泛化到多种水印策略，且较好地保持了模型精度。

5.3 模型提取

模型提取攻击是指通过模拟受害者模型的输入输出映射得到一个新的功能近似的模型。模型提取攻击需要攻击者有一定的计算资源和数据量，攻击代价更加高昂，但能够有效去除多种类型的模型水印。文献 [86] 中指出了包括重新训练、跨结构重新训练、蒸馏等多种模型提取方式。

5.4 伪造攻击

伪造攻击也叫歧义攻击，是指攻击者在不改变模型的前提下伪造出一个新的水印以造成所有权验证的歧义。Fan 等[34]对伪造攻击提供了细致的描述，并提出了一种基于护照的白盒水印方法，声称所提方法可抵抗伪造攻击。然而，一项最近的研究[88]表明，通过在护照参数前面插入一个精心设计的附件块，可以成功伪造出新的护照，在伪造护照下模型的性能与合法护照近似。

6 性能分析与比较

根据对现有文献的调研，经典的白盒水印算法都声称满足保真度要求、对权重剪枝具备鲁棒性、满足完整性和可靠性要求、具备一定的通用性。因此，本文不考虑这些指标，选择如表2所示的指标对经典白盒水印算法进行比较。尽管大多数白盒水印算法都能够从实验上验证对微调的鲁棒性，Lv 等[40]的方法是唯一能够从理论上保证对微调具备鲁棒性的工作。此外，由表2可知，白盒水印面临的主要威胁为功能等效攻击和模型提取攻击。对于功能等效攻击，Lv 等[40]和 Li 等[42]均提出了解决方案。对于模型提取攻击，Li 等[57]的方法仅在替代模型和原始模型结构相同时有效，而这一设定通常是不实际的。尽管基于护照的方法[34]声称具有不可伪造性，但最新的工作[88]对其成功实施了伪造攻击。还需要注意的是，文献 [40,57] 中的方案是为数不多的白盒设定中的零比特水印。

经典的黑盒水印算法也都声称满足保真度要求、对权重剪枝具备鲁棒性、满足完整性和可靠性要求。文献 [86] 经过大量的实验验证，发现输入修改攻击虽然能够降低黑盒水印的准确率，但通常无法将水印完全抹除。本文对经典黑盒水印算法的比较如表3所示，并指出了每个算法适用的受保护模型类型。注意表中的微调指的是基本的微调攻击，几乎所有算法在一定程度上都能够抵抗这种微调。上文介绍了多个更加先进的微调

表 2　经典白盒水印算法性能比较

Tab.2　Performance comparison for representative white-box DNN watermarking

方法	抗微调	抗重写	抗功能等效攻击	抗模型提取	抗伪造攻击	隐蔽性	容量
Uchida等[6]	◐	○	○	○	○	○	多比特
Rouhani等[38]	◐	●	○	○	○	◐	多比特
Fan等[33]	◐	●	○	○	◐	◐	多比特
Zhang等[35]	◐	●	○	○	◐	◐	多比特
Wang等[29]	◐	●	○	○	○	●	多比特
Liu等[32]	◐	●	○	○	●	◐	多比特
Li等[57]	◐	●	○	◐	●	●	零比特
Lv等[40]	●	◐	●	○	●	●	零比特

注: ○,◐和●分别代表较差、部分情况下良好、良好

表 3　经典黑盒水印算法性能比较

Tab.3　Performance comparison for representative black-box DNN watermarking

方法	抗微调	抗规避攻击	抗模型提取	抗伪造攻击	任务类型	容量
Adi等[46]	◐	○	○	○	图像分类	零比特
Zhang等[47]	◐	○	○	○	图像分类	零比特
Namba等[48]	◐	●	○	○	图像分类	零比特
Li等[52]	◐	●	○	●	图像分类	零比特
Chen等[44]	◐	◐	○	○	图像分类	多比特
Le Merrer等[43]	◐	◐	○	○	图像分类	零比特
Jia等[55]	●	○	◐	○	图像、音频分类	零比特
Tan等[56]	◐	◐	●	○	图像分类	零比特
Szyller等[59]	◐	●	◐	○	图像分类	零比特
Bansal等[58]	●	○	○	○	图像分类	零比特
Charette等[60]	●	●	◐	○	图像分类	零比特
Li等[57]	●	○	●	●	图像分类	零比特
Quan等[62]	●	○	○	○	图像处理	Logo图像
Ong等[63]	●	○	○	●	生成对抗网络	Logo图像

注: ○,◐和●分别代表较差、部分情况下良好、良好

攻击的变种，其中不乏能够去除黑盒水印的方法。表中的抗规避攻击的能力是通过触发模式的不可感知性来衡量的，因此，使用可见的触发模式或使用 OOD 数据作为触发集都被考虑为具有较差的抗规避能力。由表 3 还可以知道，模型提取攻击也对黑盒水印带来了巨大的挑战，大多数黑盒水印算法都没有考虑对模型提取的鲁棒性。此外，只有少数黑盒水印算法能够抵抗伪造攻击。

无盒水印体现为使得模型的输出图像或文本中含有不可见的水印，但由于不同算法的具体实现方式不尽相同，在论文中采用的评估标准和协议也有所差异，很难对它们的性能进行公平比较。本文只从优势和不足两方面对经典算法作了分析，结果如表 4 所示。在图像生成方面，几种无盒水印方法都有较好的保真度，含水印的图像质量较好，但只有 Zhang 等[71] 的方法能够在模型提取攻击下有效。在文本生成方面，除了表中罗列的优缺点以外，还需要注意这几种方法都是在

表4 经典无盒水印算法性能比较

Tab.4 Performance comparison for representative box-free DNN watermarking

方法	类型	优势	不足
Wu等[69]	图像	能抵抗图像处理攻击;引入密钥提高安全性	未验证对模型提取攻击的鲁棒性
Zhang等[71]	图像	能抵抗模型提取攻击和水印重写攻击	提取攻击时对输出图像做预处理则鲁棒性下降
Yu等[72]	图像	对图像扰动和模型扰动具备鲁棒性	未验证对模型提取攻击的鲁棒性
Yu等[74]	图像	可实现大规模的用户指纹管理和溯源	未验证对模型提取攻击的鲁棒性
Abdelnabi等[75]	文本	对单词去除和替换有具备鲁棒性	可能会破坏句子结构和语义连贯性
He等[76]	文本	生成文本质量较好;可抵抗模型提取攻击	可通过统计方法检测出带水印的词汇
He等[77]	文本	水印的隐蔽性好;可抵抗模型提取攻击	候选词选择有限;生成文本质量轻微下降

文本生成模型API之后插入一个水印嵌入模块，这适用于文本溯源或防御提取攻击。但如果攻击者复制目标模型本身而不在窃取的模型后添置该模块，所有者将无法检测IP侵权。

7 未来展望

未来可能的研究方向可以从如下几个方面加以考虑：

a. 提高模型水印的鲁棒性。开发鲁棒性更好的模型水印算法依然是未来研究的重点方向。在现实场景中，相比于模型本身的复制，模型提取攻击是针对商用黑盒模型的更常见的攻击方式，而大多数水印方案对于模型提取攻击的鲁棒性不佳。水印作为一种事后验证IP侵权的方式，未来可以考虑与模型提取攻击的主动防御技术结合，设计出能够更好地迁移到替代模型中的水印。此外，水印算法也需要考虑组合攻击、自适应攻击等更强的攻击策略。

b. 大模型的知识产权保护。最近，大模型的研究取得了突飞猛进的效果，以ChatGPT为代表的商用大模型产品已经成功地走入了大众的生活。未来，大模型很可能会成为AI的基础设施，必然也需要进行产权保护。目前，对于大模型水印的研究正处于起步阶段，现有研究主要关注为大模型输出的文本添加水印，这是大模型水印与现有白盒水印和黑盒水印的主要区别。然而，大模型水印依然面临着许多挑战。含水印文本可能被用户有意或无意地修改，将水印文本片段插入到人类撰写的文本中、部分文本的删除或替换、机器或人工转述都可能会对水印造成破坏。如何平衡水印鲁棒性、生成文本的质量和流畅度与水印容量，需要多长的文本才能可靠检测出水印等是学术界需要攻克的难题。

c. 数据集的知识产权保护。数据是深度学习的重要驱动力，训练一个高性能的深度学习模型必须事先收集大量的高质量数据，并花费巨大的人力和财力标注数据，因此，私有训练数据同样是一种需要保护的知识产权。数据集的所有权验证旨在判断一个模型是否在一个受保护数据集上训练过。目前，只有少数工作对数据集的所有权验证作出了初步探索，研究的模型对象仅局限于图像分类模型。未来还需要进一步拓展数据集保护的方法，异质的以及多模态的数据集保护也是一个有潜力的研究方向。

8 结束语

神经网络水印是近年来兴起的一种保护深度学习模型知识产权的重要技术。与传统的多媒体数字水印类似，神经网络水印同样有着版权保护和追踪溯源等作用，能够在模型盗版行为发生之后提供一种有效的验证手段。经过一段时间的发展，神经网络水印的研究已经取得了一定的成果。

本文系统总结了神经网络水印的研究现状，将现有的模型水印方法分为白盒水印、黑盒水印、无盒水印3类进行全面的综述，并介绍了针对神经网络水印的攻击方法。通过对近年来经典水印方法的归纳和对比，本文揭示了目前水印方法的优势和缺陷，并明确了各类水印算法适用的场景。最后，归纳了当前研究的不足和存在的挑战，并指出可能的研究方向和思路，为未来研究提供参考，期望能够推动该领域的发展。

参考文献：

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning[J].

Nature, 2015, 521(7553): 436−444.
[2] HONSINGER C. Digital watermarking[J]. Journal of Electronic Imaging, 2002, 11(3): 414.
[3] BEGUM M, UDDIN M S. Digital image watermarking techniques: a review[J]. Information, 2020, 11(2): 110.
[4] YU X Y, WANG C Y, ZHOU X. A survey on robust video watermarking algorithms for copyright protection[J]. Applied Sciences, 2018, 8(10): 1891.
[5] HUA G, HUANG J W, SHI Y Q, et al. Twenty years of digital audio watermarking —a comprehensive review[J]. Signal Processing, 2016, 128: 222−242.
[6] UCHIDA Y, NAGAI Y, SAKAZAWA S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. Bucharest: ACM, 2017: 269–277.
[7] CAO X Y, JIA J Y, GONG N Z. IPGuard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary[C]//Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security. Hong Kong: ACM, 2021: 14–25.
[8] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.
[9] XIONG C, FENG G R, LI X R, et al. Neural network model protection with piracy identification and tampering localization capability[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 2881–2889.
[10] BOENISCH F. A systematic review on model watermarking for neural networks[J]. Frontiers in Big Data, 2021, 4: 729663.
[11] REGAZZONI F, PALMIERI P, SMAILBEGOVIC F, et al. Protecting artificial intelligence IPs: a survey of watermarking and fingerprinting for machine learning[J]. CAAI Transactions on Intelligence Technology, 2021, 6(2): 180−191.
[12] XUE M F, ZHANG Y S, WANG J, et al. Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations[J]. IEEE Transactions on Artificial Intelligence, 2022, 3(6): 908−923.
[13] 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.
[14] 王馨雅, 华光, 江昊, 等. 深度学习模型的版权保护研究综述 [J]. 网络与信息安全学报, 2022, 8(2): 1−14.
[15] LEDERER I, MAYER R, RAUBER A. Identifying appropriate intellectual property protection mechanisms for machine learning models: a systematization of watermarking, fingerprinting, model access, and attacks[J/OL]. IEEE Transactions on Neural Networks and Learning Systems, 2023. https://ieeexplore.ieee.org/abstract/document/10143370
[16] 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.
[17] HE Z C, ZHANG T W, LEE R. Sensitive-sample fingerprinting of deep neural networks[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4724–4732.
[18] ARAMOON O, CHEN P Y, QU G. AID: attesting the integrity of deep neural networks[C]//2021 58th ACM/IEEE Design Automation Conference (DAC). San Francisco: IEEE, 2021: 19–24.
[19] ZHU R J, WEI P, LI S, et al. Fragile neural network watermarking with trigger image set[C]//Proceedings of the 14th Knowledge Science, Engineering and Management. Tokyo: Springer, 2021: 280–293.
[20] YIN Z X, YIN H, ZHANG X P. Neural network fragile watermarking with no model performance degradation[C]//2022 IEEE International Conference on Image Processing (ICIP). Bordeaux: IEEE, 2022: 3958–3962.
[21] BOTTA M, CAVAGNINO D, ESPOSITO R. NeuNAC: a novel fragile watermarking algorithm for integrity protection of neural networks[J]. Information Sciences, 2021, 576: 228−241.
[22] ZHAO G J, QIN C, YAO H, et al. DNN self-embedding watermarking: towards tampering detection and parameter recovery for deep neural network[J]. Pattern Recognition Letters, 2022, 164: 16−22.
[23] GUAN X Q, FENG H M, ZHANG W M, et al. Reversible watermarking in deep convolutional neural networks for integrity authentication[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020: 2273–2280.
[24] CHEN H L, ROUHANI B D, FU C, et al. DeepMarks: a secure fingerprinting framework for digital rights management of deep learning models[C]//Proceedings of the 2019 on International Conference on Multimedia Retrieval. Ottawa: ACM, 2019: 105–113.
[25] YU Y S, LU H W, CHEN X S, et al. Group-oriented anti-collusion fingerprint based on BIBD code[C]//2010 2nd International Conference on E-business and Information System Security. Wuhan: IEEE, 2010: 1–5.
[26] WANG T H, KERSCHBAUM F. Attacks on digital watermarks for deep neural networks[C]//ICASSP 2019–2019 IEEE International Conference on Acoustics,

Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 2622–2626.

[27] FENG L, ZHANG X P. Watermarking neural network with compensation mechanism[C]//Proceedings of the 13th Knowledge Science, Engineering and Management. Hangzhou: Springer, 2020: 363–375.

[28] WANG J F, WU H Z, ZHANG X P, et al. Watermarking in deep neural networks via error back-propagation[J]. Electronic Imaging, 2020, 2020(4): 022-1–022-9.

[29] WANG T H, KERSCHBAUM F. RIGA: Covert and robust white-box watermarking of deep neural networks[C]//Proceedings of the Web Conference 2021. Ljubljana: ACM, 2021: 993–1004.

[30] KURIBAYASHI M, TANAKA T, SUZUKI S, et al. White-box watermarking scheme for fully-connected layers in fine-tuning model[C]//Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2021: 165–170.

[31] CHEN H L, ROUHANI B D, KOUSHANFAR F. SpecMark: a spectral watermarking framework for IP protection of speech recognition systems[C]//INTERSPEECH 2020, 21st Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020: 2312–2316.

[32] LIU H W, WENG Z Y, ZHU Y S. Watermarking deep neural networks with greedy residuals[C]//Proceedings of the 38th International Conference on Machine Learning. Vienna: ICML, 2021: 6978–6988.

[33] FAN L X, NG K W, CHAN C S. Rethinking deep neural network ownership verification: embedding passports to defeat ambiguity attacks[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc. , 2019: 424.

[34] 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.

[35] ZHANG J, CHEN D D, LIAO J, et al. Passport-aware normalization for deep model protection[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc. , 2020: 1896.

[36] ZHAO X Y, YAO Y Z, WU H Z, et al. Structural watermarking to deep neural networks via network channel pruning[C]//2021 IEEE International Workshop on Information Forensics and Security (WIFS). Montpellier: IEEE, 2021: 1–6.

[37] XIE C Q, YI P, ZHANG B W, et al. DeepMark: embedding watermarks into deep neural network using pruning[C]//2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). Washington: IEEE, 2021: 169–175.

[38] ROUHANI B D, CHEN H L, KOUSHANFAR F. DeepSigns: an end-to-end watermarking framework for ownership protection of deep neural networks[C]//Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems. Providence: ACM, 2019: 485–497.

[39] 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.

[40] 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.

[41] 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 (ICASSP). Singapore: IEEE, 2022: 3049–3053.

[42] LI F Q, WANG S L, LIEW A W C. 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.

[43] LE MERRER E, PÉREZ P, TRÉDAN G. Adversarial frontier stitching for remote neural network watermarking[J]. Neural Computing and Applications, 2020, 32(13): 9233−9244.

[44] CHEN H, ROUHANI B D, KOUSHANFAR F. Blackmarks: Blackbox multibit watermarking for deep neural networks[EB/OL]. (2019-03-31) https://arxiv.longhoe.net/abs/1904.00344

[45] GU T Y, LIU K, DOLAN-GAVITT B, et al. Badnets: Evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230−47244.

[46] ADI Y, BAUM C, CISSE M, et al. Turning your weakness into a strength: watermarking deep neural networks by backdooring[C]//Proceedings of the 27th USENIX Conference on Security Symposium. Baltimore: USENIX Association, 2018: 1615–1631.

[47] ZHANG J L, GU Z S, JANG J, et al. Protecting intellectual property of deep neural networks with watermarking[C]//Proceedings of the 2018 on Asia Conference on Computer and Communications Security. Incheon: ACM, 2018: 159–172.

[48] NAMBA R, SAKUMA J. Robust watermarking of neural

network with exponential weighting[C]//Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security. Auckland: ACM, 2019: 228–240.

[49] GUO J, POTKONJAK M. Watermarking deep neural networks for embedded systems[C]//2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). San Diego: IEEE, 2018: 1–8.

[50] GUO J, POTKONJAK M. Evolutionary trigger set generation for DNN black-box watermarking[EB/OL]. (2019-06-11). https://arxiv.longhoe.net/abs/1906.04411

[51] HITAJ D, MANCINI L V. Have you stolen my model? evasion attacks against deep neural network watermarking techniques[EB/OL]. (2018-09-03). https://arxiv.longhoe.net/abs/1809.00615

[52] 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]//Proceedings of the 35th Annual Computer Security Applications Conference. San Juan: ACM, 2019: 126–137.

[53] LI M, ZHONG Q, ZHANG L Y, et al. Protecting the intellectual property of deep neural networks with watermarking: the frequency domain approach[C]//2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Guangzhou: IEEE, 2020: 402–409.

[54] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs[C]//Proceedings of the 25th USENIX Conference on Security Symposium. Austin: USENIX Association, 2016: 601–618.

[55] JIA H R, CHOQUETTE-CHOO C A, CHANDRASE-KARAN V, et al. Entangled watermarks as a defense against model extraction[C]//Proceedings of the 30th USENIX Security Symposium. Vancouver, 2021: 1937–1954.

[56] TAN J X, ZHONG N, QIAN Z X, et al. Deep neural network watermarking against model extraction attack[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023: 1588–1597.

[57] LI Y M, ZHU L H, JIA X J, et al. MOVE: effective and harmless ownership verification via embedded external features[EB/OL]. (2022-08-04). https://arxiv.longhoe.net/abs/2208.02820

[58] BANSAL A, CHIANG P Y, CURRY M J, et al. Certified neural network watermarks with randomized smoothing[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022: 1450–1465.

[59] SZYLLER S, ATLI B G, MARCHAL S, et al. DAWN: Dynamic adversarial watermarking of neural networks[C]//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu: ACM, 2021: 4417–4425.

[60] CHARETTE L, CHU L Y, CHEN Y Z, et al. Cosine model watermarking against ensemble distillation[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2022: 9512–9520.

[61] 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.

[62] 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.

[63] ONG D S, CHAN C S, NG K W, et al. Protecting intellectual property of generative adversarial networks from ambiguity attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3629–3638.

[64] 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 (ICASSP). Singapore: IEEE, 2022: 3059–3063.

[65] ZHANG J, DAI L, XU L R, et al. Black-box watermarking and blockchain for IP protection of voiceprint recognition model[J]. Electronics, 2023, 12(17): 3697.

[66] YADOLLAHI M M, SHOELEH F, DADKHAH S, et al. Robust black-box watermarking for deep neural network using inverse document frequency[C]//2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). AB: IEEE, 2021: 574–581.

[67] ZHAO X Y, WU H Z, ZHANG X P. Watermarking graph neural networks by random graphs[C]//2021 9th International Symposium on Digital Forensics and Security (ISDFS). Elazig: IEEE, 2021: 1–6.

[68] 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.

[69] ZHANG J, CHEN D D, LIAO J, et al. Model watermarking for image processing networks[C]//Proceedings of the 34th AAAI conference on artificial intelligence. New York:

AAAI, 2020: 12805–12812.

[70] 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, 2022, 44(8): 4005−4020.

[71] YU N, SKRIPNIUK V, ABDELNABI S, et al. Artificial fingerprinting for generative models: rooting deepfake attribution in training data[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 14428–14437.

[72] FEI J W, XIA Z H, TONDI B, et al. Supervised GAN watermarking for intellectual property protection[C]//2022 IEEE International Workshop on Information Forensics and Security (WIFS). Shanghai: IEEE, 2022: 1–6.

[73] ZHAO Y Q, PANG T Y, DU C, et al. A recipe for watermarking diffusion models[EB/OL]. (2023-03-17). https://arxiv.longhoe.net/abs/2303.10137

[74] YU N, SKRIPNIUK V, CHEN D F, et al. Responsible disclosure of generative models using scalable fingerprinting[C]//The Tenth International Conference on Learning Representations(Virtual). ICLR, 2022.

[75] ABDELNABI S, FRITZ M. Adversarial watermarking transformer: Towards tracing text provenance with data hiding[C]//2021 IEEE Symposium on Security and Privacy (SP). San Francisco: IEEE, 2021: 121–140.

[76] HE X L, XU Q K, LYU L J, et al. Protecting intellectual property of language generation APIs with lexical watermark[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2022: 10758–10766.

[77] HE X L, XU Q K, ZENGT Y, et al. CATER: Intellectual property protection on text generation APIs via conditional watermarks[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc. , 2022: 392.

[78] YOO K Y, AHN W, JANG J, et al. Robust multi-bit natural language watermarking through invariant features[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto: ACL, 2023: 2092–2115.

[79] YANG X, CHEN K J, ZHANG W M, et al. Watermarking text generated by black-box language models[EB/OL]. (2023-05-14). https://arxiv.longhoe.net/abs/2305.08883

[80] KIRCHENBAUER J, GEIPING J, WEN Y X, et al. A watermark for large language models[C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu: ICML, 2023: 17061–17084.

[81] KIRCHENBAUER J, GEIPING J, WEN Y X, et al. On the reliability of watermarks for large language models[EB/OL]. (2023-06-07). https://arxiv.longhoe.net/abs/2306.04634

[82] ZHAO X D, ANANTH P, LI L, et al. Provable robust watermarking for AI-generated text[EB/OL]. (2023-06-30). https://arxiv.longhoe.net/abs/2306.17439

[83] CHRIST M, GUNN S, ZAMIR O. Undetectable watermarks for language models[EB/OL]. (2023-05-25). https://arxiv.longhoe.net/abs/2306.09194

[84] WANG L A, YANG W K, CHEN D L, et al. Towards codable watermarking for injecting multi-bits information to LLMs[EB/OL]. (2023-07-29). https://arxiv.longhoe.net/abs/2307.15992

[85] YAN Y F, PAN X D, ZHANG M, et al. Rethinking white-box watermarks on deep learning models under neural structural obfuscation[C]//32nd USENIX Security Symposium. Anaheim: USENIX Association, 2023: 2347–2364.

[86] LUKAS N, JIANG E, LI X D, et al. SoK: How robust is image classification deep neural network watermarking?[C]//2022 IEEE Symposium on Security and Privacy (SP). San Francisco: IEEE, 2022: 787–804.

[87] WANG R, LI H X, MU L Z, et al. Rethinking the vulnerability of DNN watermarking: are watermarks robust against naturalness-aware perturbations?[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 1808–1818.

[88] CHEN Y M, TIAN J Y, CHEN X Y, et al. Effective ambiguity attack against passport-based DNN intellectual property protection schemes through fully connected layer substitution[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 8123–8132.

（编辑：丁红艺）