E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

**中图法分类号**:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2023)06-1792-19

**论文引用格式**:Wu H Z, Zhang J, Li Y, Yin Z X, Zhang X P, Tian H, Li B, Zhang W M and Yu N H. 2023. Overview of artificial intelligence model watermarking. Journal of Image and Graphics, 28(06):1792-1810(吴汉舟, 张杰, 李越, 殷赵霞, 张新鹏, 田晖, 李斌, 张卫明, 俞能海. 2023. 人工智能模型水印研究进展. 中国图象图形学报, 28(06):1792-1810)[DOI:10.11834/jig.230010]

# 人工智能模型水印研究进展

吴汉舟[1],张杰[2],李越[3],殷赵霞[4],张新鹏[1,5*],田晖[3],李斌[6],张卫明[2],俞能海[2]

1. 上海大学通信与信息工程学院,上海　200444; 2. 中国科学技术大学网络空间安全学院,合肥　230027;
3. 华侨大学计算机科学与技术学院,厦门　361021; 4. 华东师范大学通信与电子工程学院,上海　200240;
5. 复旦大学计算机科学技术学院,上海　200438; 6. 深圳大学电子与信息工程学院,深圳　518060

**摘　要**:以神经网络为代表的人工智能技术在计算机视觉、模式识别和自然语言处理等诸多应用领域取得了巨大的成功,包括谷歌、微软在内的许多科技公司都将人工智能模型部署在商业产品中,以提升服务质量和经济效益。然而,构建性能优异的人工智能模型需要消耗大量的数据、计算资源和专家知识,并且人工智能模型易于被未经授权的用户窃取、篡改和贩卖。在人工智能技术迅速发展的同时,如何保护人工智能模型的知识产权具有显著学术意义和产业需求。在此背景下,本文主要介绍基于数字水印的人工智能模型产权保护技术。通过与传统多媒体水印技术进行对比,首先概述了人工模型水印的研究意义、基础概念和评价指标;然后,依据水印提取者是否需要掌握目标模型的内容细节以及是否需要和目标模型进行交互,从"白盒"模型水印、"黑盒"模型水印、"无盒"模型水印3个不同的角度分别梳理了国内外研究现状并总结了不同方法的差异,与此同时,对脆弱模型水印也进行了分析和讨论;最后,通过对比不同方法的特点、优势和不足,总结了不同场景下模型水印的共性技术问题,并对发展趋势进行了展望。

**关键词**:模型水印;数字水印;信息隐藏;人工智能安全;知识产权保护

# Overview of artificial intelligence model watermarking

Wu Hanzhou[1], Zhang Jie[2], Li Yue[3], Yin Zhaoxia[4], Zhang Xinpeng[1,5*], Tian Hui[3], Li Bin[6], Zhang Weiming[2], Yu Nenghai[2]

*1. School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China;*
*2. School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230027, China;*
*3. School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China;*
*4. School of Communication and Electronic Engineering, East China Normal University, Shanghai 200240, China;*
*5. School of Computer Science, Fudan University, Shanghai 200438, China;*
*6. School of Electronic and Information Engineering, Shenzhen University, Shenzhen 518060, China*

**Abstract**: The deep neural networks (DNNs)-relevant artificial intelligence (AI) technique has been developing intensively in the context of such domains like computer vision, pattern analysis, natural language processing, bioinformatics,

**收稿日期**:2023-01-06;**修回日期**:2023-02-28;**预印本日期**:2023-03-06

*通信作者:张新鹏　xzhang@shu.edu.cn

**基金项目**:国家自然科学基金项目(U22B2047);上海市科学技术委员会资助项目(21010500200);CCF-腾讯犀牛鸟基金项目(CCF-Tencent RAGR20220126)

**Supported by**: National Natural Science Foundation of China (U22B2047); Science and Technology Commission of Shanghai Municipality (21010500200); CCF-Tencent Rhino-Bird Young Faculty Open Research Fund (CCF-Tencent RAGR20220126)

and games. Especially, AI models have been widely deployed in the cloud by technology companies to provide smart and personalized services. However, creating state-of-the-art AI models requires a lot of high-quality data, powerful computing resources and expert knowledge of the architecture design. Furthermore, AI models are threatened to be copied, tampered and redistributed in an unauthorized manner. It indicates that it is necessary to protect the AI models against intellectual property infringement, which yields researchers to concern about the intellectual property protection. Current techniques are concerned of digital watermarking for intellectual property protection of AI models, which is referred to AI model watermarking. The core of AI model watermarking is to embed a secret watermark revealing the ownership of the AI model to be protected into the AI model through an imperceptible way. However, unlike many multimedia watermarking methods that treat media data as a static signal, AI model watermarking is required to embed information into an AI model with a specific task. We cannot directly apply conventional multimedia watermarking methods to AI models since simply modifying a given AI model may significantly impair the performance of the AI model on its original task. It motivates people to design watermarking methods specifically for AI models. For embedding a watermark, the performance of the watermarked AI model on its original task should not be degraded significantly and the embedded watermark concealed in the watermarked AI model should be able to be extracted to identify the ownership of the AI model when disputes arise. Considering whether the watermark extractor should know the internal details of the target AI model or not, we can divide existing methods into two categories, i. e. , white-box AI model watermarking and black-box AI model watermarking. For white-box AI model watermarking, the watermark extractor should know the internal details of the target watermarked AI model so that he can extract the embedded watermark from model parameters or model structures. For black-box AI model watermarking, the watermark extractor does not know the internal details of the target watermarked AI model, but he has the ability to query the prediction results of the target model in correspondence to a set of trigger samples. The trigger samples are carefully crafted. By checking whether the prediction results are consistent with the pre-specific labels of the trigger samples, the watermark extractor is capable of determining the ownership of the target watermarked AI model. A special case of black-box AI model watermarking is box-free AI model watermarking, in which the watermark extractor has no access to the target model. It means that the watermark extractor cannot know the internal details of the model and interact with the model. However, the watermark extractor can extract the watermark from any sample generated by the target model. The ownership can be verified via extracting the watermark from the output. In addition, fragile AI model watermarking has also been investigated recently. Unlike many methods that focus on robust verification, fragile model watermarking enables us to detect whether the target model was modified or not, thereby achieving integrity verification of the target model. To review the latest developments and trends, advanced methodologies in AI model watermarking are analyzed as mentioned below: 1) the aims and objectives, basic concepts, evaluation metrics and technical classification of AI model watermarking are introduced. 2) Current development status of AI model watermarking is summarized and analyzed. 3) Such pros and cons are compared and analyzed as well. 4)The prospects for the development trend of AI model watermarking and the potentials of AI model watermarking in relevance to AI security are provided.

**Key words**: model watermarking; digital watermarking; information hiding; artificial intelligence security; intellectual property protection

## 0 引 言

2021年3月，新华社公布了《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》，提出了以人工智能为代表的新型基础设施建设政策，标志着人工智能发展进入技术持续创新和应用广泛深化的新阶段。然而，人工智能在相关行业和人民社会生活中深度应用的同时，也带来国家安全、社会伦理、网络安全、人身安全和隐私保护等多个层面的风险和挑战。如何研判和防范人工智能发展所带来的潜在风险对维护人民利益和国家社会安全具有重要的战略意义。

得益于计算机硬件和大数据的发展，以神经网络为代表的人工智能技术在诸多应用领域取得了巨大成功，包括图像识别、自然语言处理、自动驾驶、生物医学和金融决策等。以谷歌、微软、腾讯和百度为代表的大型科技公司都将人工智能作为战略性技

术，通过将人工智能模型部署在商业产品中，显著提升了服务质量和经济效益。然而，构建这些性能优异的人工智能模型不仅需要设计者的智慧，而且需要消耗大量的训练数据和计算资源。因此，人工智能模型作为一种昂贵的数字资产，如何保护其知识产权不受侵害十分重要。尤其是在人工智能技术迅速发展的同时，研究人工智能模型的知识产权保护具有显著意义。主要体现在：1）该研究领域方兴未艾，其基础理论与关键方法中蕴含重要的科学问题，极具学术研究价值；2）对人工智能模型标识所有者、使用者、版本号、传播路径并进行篡改检测，建立人工智能模型全生命周期的秩序保障，可为人工智能的发展和应用提供必不可少的良好环境；3）该研究是利用人工智能技术保护人工智能模型，其成果将辐射和带动人工智能安全技术进步。

在此背景下，为适应人工智能的发展需求，近年来研究人员主要运用数字水印技术保护人工智能模型的知识产权。如图1所示，数字水印是一种将特定信息（称为“水印”）以不损害信号使用价值的方式隐藏在数字信号中的安全技术，该信号可以是图像、视频和音频等任意数字产品（孙杉 等，2022；王翌妃 等，2022；郑钢 等，2021）。倘若拷贝含有水印的信号，水印也会一并被拷贝。当水印包含了产权信息，一旦信号被泄露，通过在泄露信号中重构水印，可确定其产权。显然，作为一种数字产品，人工智能模型也可承载水印以保护其知识产权，称为人工智能模型水印（简称模型水印）（Uchida 等，2017）。

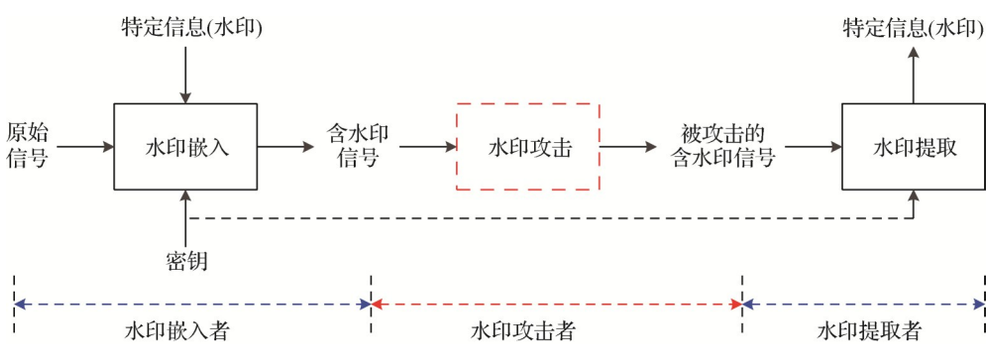


图1　数字水印的基本框架示意图

Fig. 1　General framework for digital watermarking

传统数字水印技术主要针对多媒体数据。多媒体数据的静态特性和冗余特性使多媒体的价值不因水印嵌入而遭受明显损害。然而，不同于多媒体数据，人工智能模型需要完成特定任务（如图像分类、人脸识别等），这种功能特性使得简单地将传统多媒体水印技术应用于人工智能模型会降低人工智能模型在特定任务上的计算性能，损害人工智能模型的使用价值，甚至使水印嵌入失去意义。因此，人工智能模型水印原则上应当确保水印嵌入不会严重损害人工智能模型在特定任务上的计算性能。在此基础上，人工智能模型水印还需要考虑水印的保真度、水印的嵌入量、水印的鲁棒性、水印的隐蔽性、水印的安全性、计算复杂度和普适性等评价指标，如表1所示。需要指出的是，在具体算法的设计过程中，还需要根据具体的应用场景在不同的评价指标间取得平衡，以满足实际的应用需求。

**表1　人工智能模型水印的常见评价指标**

**Table 1　Common evaluation indicators for artificial intelligence model watermarking**

| 评价指标 | 含义 |
|---|---|
| 保真度 | 含水印模型在原始任务上的计算性能及重构水印的质量 |
| 嵌入量 | 水印承载的信息量 |
| 鲁棒性 | 从被攻击的模型中重构出正确水印的可能性 |
| 隐蔽性 | 水印不被察觉的能力 |
| 安全性 | 水印不被恶意检测、伪造和篡改等的可能性 |
| 计算复杂度 | 水印嵌入和水印提取时的计算开销 |
| 普适性 | 算法的适用范围和可扩展性 |

2017年以来，人工智能模型水印得到了学术界和工业界越来越多的关注，各国研究机构和大型科技公司开始投入大量资源开展人工智能模型水印研究，相关成果数量成逐年增长趋势。因此，有必要梳理人工智能模型水印的发展现状、前沿动态、热

点问题和发展趋势。这对相关机构决策、相关领域从业者和研究人员了解进展有重要意义。为此，本文以神经网络为载体，对人工智能模型水印的国内外研究现状进行相对全面的介绍，按照模型水印的算法原理和特点对主流方法进行分类和比较，通过分析和对比各种模型水印算法的优缺点，结合实际应用场景需求，对模型水印的发展趋势进行展望。

## 1 国际研究现状

主流人工智能模型水印算法以神经网络为保护对象。神经网络是一种模仿生物中枢神经系统的结构和功能的数学计算模型，由有限多个神经元联结起来进行计算，每个神经元的功能是计算加权向量经非线性映射后的结果。依据水印提取者是否需要掌握神经网络的内部细节，可将当前国际主流的模型水印算法大致归类为白盒模型水印和黑盒模型水印。接下来，首先介绍两类主流模型水印算法，然后概述其他算法。

### 1.1 白盒模型水印

白盒模型水印算法强调水印在提取的过程中对目标模型内部细节的可获取性，通过获取目标模型的内部结构和参数等信息，实现对目标模型的版权鉴定。如图2所示，目前，国际上主要是通过调整模型的参数或者模型的网络结构来实现白盒模型水印技术。

2017年，Uchida等人(2017)首次提出了一种将数字水印技术用于神经网络模型版权保护的白盒方

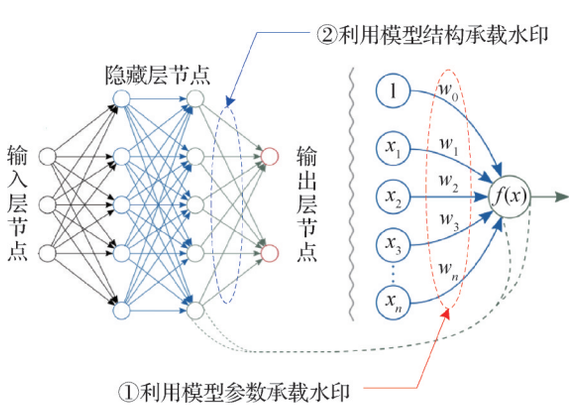


图2 白盒模型水印中两种常见的水印承载方式

Fig. 2 Two common watermark carrying methods in white-box model watermarking

案。该方案选择神经网络模型的权重承载水印信息，并利用激活函数将自定义的扩频矩阵与权重的乘积值约束在特定区间内，通过训练基于交叉熵函数的正则项，实现该乘积值序列与待嵌入水印序列的一致性，从而实现水印被嵌入到模型权重的目的。在水印提取的过程中，借助扩频矩阵计算得到水印信息实现模型版权的鉴定，由于扩频矩阵兼具了扩频和密钥的功能，该方案的鲁棒性和安全性得到了保障。实验结果表明，该方案不仅不会对模型分类的准确率产生影响，反而因为正则项的引入避免了过拟合问题，使得模型的性能得到提升，但该方案不能抵御水印复写攻击。

Uchida等人(2017)的方案是在已训练完备的权重上提取水印信息，这种静态水印无法与模型输入产生映射关系，缺乏灵活性。为此，Rouhani等人(2019)提出了Deepsigns方案，使与输入内容直接相关的层激活输出为水印的嵌入位置。为了避免直接修改激活输出所带来的模型分类准确率下降，Deepsigns借鉴了Uchida等人(2017)的水印嵌入方案，先将水印嵌入到满足高斯分布的自定义的激活输出上，再通过损失函数的约束，达到目标模型中的激活输出与含水印的自定义激活输出高度相似的效果。实验结果表明，该方案鲁棒性良好，但水印容量较小。为了增强安全性，Chen等人(2019a)提出了Deepmarks，在已有的水印嵌入方案基础上，利用抗合谋攻击码生成水印信息。由于不同的模型使用者被分发的水印不同，该方案在抵御共谋攻击的同时也实现了溯源。

然而，Wang和Kerschbaum(2019)通过对上述方案水印嵌入前后模型的权重分析发现，含水印权重和不含水印权重存在显著的统计差异，这不仅削弱了水印的隐蔽性，甚至可以使攻击者通过统计分析探知水印长度，从而实现精准的水印复写攻击。为了解决这一问题，Wang等人(2021)借助生成对抗网络，采用对抗训练机制，约束水印在嵌入过程中对权重分布的影响，使含水印的权重与原始策略训练后的权重的分布保持一致，增强了这一类方案中水印的不可见性。

Cortiñas-Lorenzo和Pérez-González(2020)发现在模型的训练过程中使用Adam(adaptive moment estimation)进行优化也会对训练后含水印的权重分布产生巨大影响，削弱水印的隐蔽性。为了解决这个

问题，他们设计了一种投影和反投影方案，通过在每次迭代中对投影后的权值和梯度运行Adam优化算法，消除权重分布中不希望出现的异常峰值。优化完成后，投影后的权值和梯度将进一步反投影到原始空间中，在此过程中，块正交投影的运用减少了该方案的计算开销和内存空间的消耗。实验结果证实，该水印方案具有较好的不可检测性，提高了对模型剪枝攻击的鲁棒性。

除了从模型的角度提升水印的隐蔽性，Kuribayashi等人（2020）借助传统数字水印中的量化索引调制技术提升了模型水印的性能。具体的操作方法是，采样得到部分全连层的权重，在使用量化索引调制技术对这部分权重进行水印嵌入后，用含水印的权重替换掉原始权重。然而，该方案仅适用于基于微调的水印嵌入方式，并且为了保证模型计算精度，一般只对模型的部分结构层进行微调。

为了进一步增强水印的鲁棒性，使鲁棒性不受嵌入容量的困扰，Tartaglione等人（2021）提出白盒模型零水印方案，在模型训练前随机选取模型中的部分权重并进行标记，利用约束条件确保该部分权重在训练过程中不伴随梯度下降而进行权值更新。水印检测时，通过对该部分权重进行提取并进行权值比较，即可判定模型的版权归属。实验结果证实了该方案在抵御模型微调攻击方面有着较强的鲁棒性，且对目标模型性能几乎不产生影响。然而，由于零水印方案无真实数据嵌入，该方案的应用场景受到了一定限制。

除了以参数为载体的白盒模型水印方案，Lou等人（2022）选择以模型结构作为水印载体，该方案对基于参数修改的攻击具有固有的鲁棒性。其利用神经网络结构搜索来获取可嵌入水印的模型结构，固定部分边（即计算节点之间的连接）和操作以缩小搜索空间，所选的边和操作被视为水印信息。在较小的搜索空间中，通过神经网络结构搜索方法进一步找到最优的网络结构，并在原始数据集上训练以获得含水印模型。在验证阶段，使用侧信道捕获泄露信息以便得到模型结构从而恢复水印。然而，Lou等人（2022）的方法仅适用于从起始阶段开始训练的模型，无法在已训练好的模型上实施水印嵌入。

上述白盒模型水印方案侧重于保护卷积神经网络模型，应用的范围多局限于图像分类任务。然而，网络结构的差异性使得适用于卷积神经网络的模型水印方案很难直接迁移至其他类型的网络模型。因此，Ong等人（2021）提出了针对生成对抗网络模型的版权保护方案。该方案综合考虑了白盒和黑盒两种应用场景。对于白盒场景，该方案选取生成对抗网络（generative adversarial network，GAN）模型中生成器的标准化层为水印的嵌入位置，通过定义正则化项实现对标准化层中缩放参数的约束，迫使缩放参数依照水印序列的正负性进行二极化。通过仿真实验，该方案被证实在DCGAN（deep convolutional generative adversarial network）、SRGAN（super resolution generative adversarial network）和Cycle-GAN等多种GAN模型上有效，同时该方案可以抵御权重微调和水印复写在内的多种攻击。受Ong等人（2021）的启发，Lim等人（2022）提出了循环神经网络模型水印方案，采用与Ong等人（2021）相同的约束函数，趋使模型中的超参数依据水印序列的正负性实现二值化。不同点在于，该超参数不再来自标准化层，而是取自循环神经模型单元输出的隐层状态中。Chen等人（2020）也将提出的Deepmarks算法（Chen等，2019a）拓展应用到语音识别系统。

随着受保护模型类型的扩展，白盒模型水印的应用范围也在拓宽。Chen等人（2021b）依据“彩票假设”提出一种针对“中奖彩票”类稀疏网络的版权保护方案。彩票假设来源于模型剪枝问题，虽然模型剪枝技术在保证模型性能基础上极大压缩了模型的规模，但是被剪枝后得到的稀疏网络存在无法从起始阶段开始训练的问题。因此，Frankle和Carbin（2019）提出一种彩票假设理论，即在密集的、随机初始化的神经网络里面存在着一些子网络，从起始阶段开始独立训练时可以在相似的迭代次数内达到与原始网络相当的计算精度。假设大的网络里面包含着千千万万的子网络，这些子网络可以看做是一张张彩票，满足上述特殊性质的稀疏网络就称为“中奖彩票”。然而找到这些“中奖彩票”在时间和算力上的消耗并不亚于从起始阶段开始训练一个密集型的神经网络模型。因此，高品质“中奖彩票”也是一种需要被保护的知识产权。中奖彩票的搜寻过程需要将所定义的不同稀疏掩码叠加至原始模型权重进行处理，Chen等人（2021b）提出以找到中奖彩票的稀疏掩码作为密钥，其中包含了以QR（quick response）编码方式嵌入的版权所有者的水印信息。合法的使用者利用完整的稀疏掩码可获取到“中奖彩票”的全

部权重，通过“中奖彩票”模型的性能良好与否即可判定模型的版权归属。

### 1.2 黑盒模型水印

白盒水印存在着一些局限性，例如验证阶段必须获取可疑模型的内部信息，这在实际验证过程中是难以保证的，原因是模型窃取者可能并不会配合取证。深度模型的训练依托数据集，一般通过有监督的方式学习一个从输入到输出的映射。以分类任务 $D=\{X, Y\}$ 为例，模型训练者拥有带标签的原始训练数据集，其中 $X$ 表示所有的训练数据，$Y$ 表示相应的标签，其目的是在一定的损失函数下学习到由 $X$ 指向 $Y$ 的映射。然而，相关研究表明，即使训练数据被赋予了错误的标签，深度模型仍能很好地拟合这些异常关系（Zhang 等，2021a）。黑盒水印利用了模型的这种冗余特性，记住一些额外添加的特殊输入输出关系，进而使用这些独有的关系作为模型所有者的版权标识。图 3 给出了黑盒模型水印的一般框架，可以简单描述如下：水印嵌入者利用正常图像和触发图像训练待保护的模型，训练好的模型即为含有水印的模型；在验证阶段，水印提取者通过获取目标模型在触发图像的预测结果，并与预先指定的标签进行一致性分析，可以验证其产权。

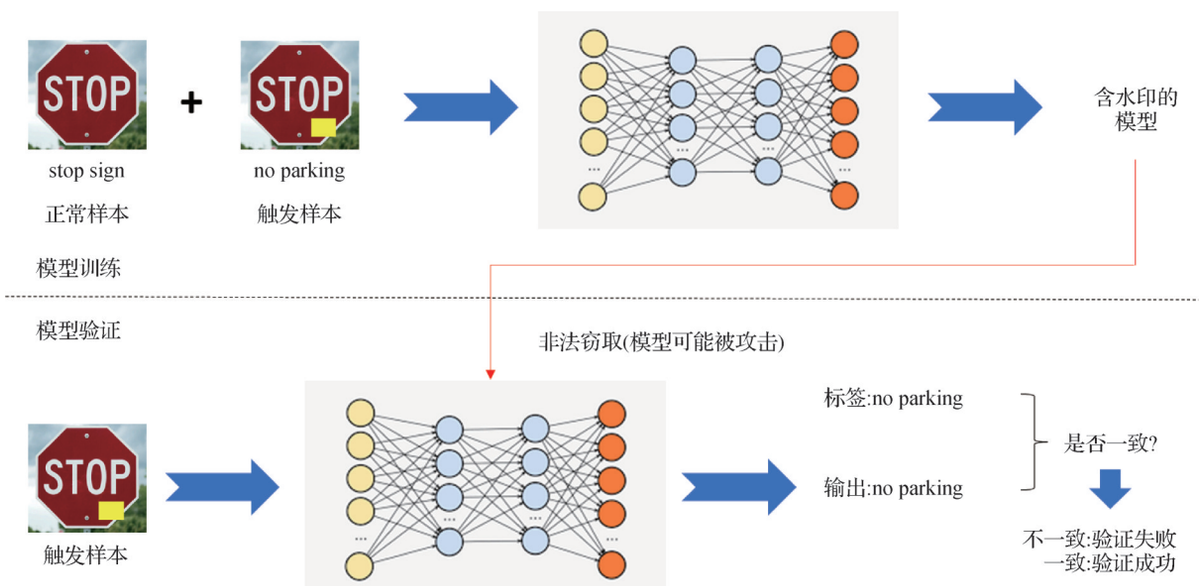


图 3 黑盒模型水印的一般框架示意图（以分类模型为例）

Fig. 3 General framework for black-box model watermarking (classification as an example)

与白盒模型水印相比，黑盒模型水印仅需对模型进行访问，从而获得相应的输出来进行所有权的验证，也就是说黑盒模型在验证阶段不需要模型的内部信息，甚至可对可疑模型进行远程验证。Facebook 团队的 Adi 等人（2018）提出了第一个基于后门攻击的黑盒模型水印方法。深度模型中后门攻击的想法来源于传统系统安全中的后门概念，即可以将模型看做是一个具备输入和输出的系统，对其进行后门植入，而此处的后门可以看做是上述提到的特殊的输入输出关系。Adi 等人（2018）直接巧妙地将后门攻击的思想用于模型的版权认证，具体而言，随机挑选一些抽象的图像作为触发后门，并使用事先定义的目标标签对其进行标注，随后将其混入原始训练集中训练目标模型。训练好的模型在正常输入上表现正常，在遇到选定的触发图像时，模型会输出指定的目标标签。在验证阶段，只需对模型远程输入这些抽象的触发图像，如果得到事先指定的标签输出，则可证明模型的版权所属。为了保证产权验证的安全性，Adi 等人（2018）进一步提出了一种基于零知识证明的公开验证方案。在此工作基础上，IBM 团队的 Zhang 等人（2018）继续探究了使用不同类型的触发后门的可能性，包括选择与训练数据分布无关的图像、使用一些随机噪声或将特定内容字符串添加在干净图像上。然而，上述触发模式也有一些局限性。如无关图像易于被检测发现；随机噪声虽不易察觉但不能代表模型所有者的身份；内容字符串不能绕过人工检查或机器检测。Guo 和 Potkonjak（2018）沿用了在训练数据上添加噪声的方

式,但该噪声不是随机生成,而是由用户信息指导生成,从而建立了不可察觉的噪声与模型拥有者之间的联系,实现了明确的所有权归属划分。Guo和Potkonjak(2021)还采用遗传算法,搜寻到了触发图像的最优模式以及该模式应该添加的最优位置。

上述的触发图像都是事先人工设计的,没有依赖目标模型的知识,这样的模型水印策略对于目标模型本身的容量和复杂度具有很高的要求。最近的一些工作开始利用模型自身的一些特性,例如,法国国家信息与自动化研究所Le Merrer等人(2020)提出了微调模型决策边界的方式来表示水印信息。具体来说,受对抗样本的启发,首先找到一些真实的对抗样本(被原始模型错误分类)和错误的对抗样本(仍然由原始模型正确预测),而这两种样本都刚好处在原始模型的决策边界附近。之后,通过对抗训练将真实的对抗样本进一步分类为真实标签,从而稍微调整决策边界,使得添加水印后的模型具有特有的决策边界。版权验证过程与基于后门的模型水印方法类似。基于Le Merrer等人(2020)的方法,Chen等人(2019b)提出了BlackMarks,将模型预测类别平均分为两组,分别代表"0"和"1"。也就是说,被错误分类到属于"0"组标签的对抗样本用于表示比特信息"0",比特信息"1"的表示方式类似,之后同样通过对抗训练来微调原始模型。通过这种方式提高了水印的容量,更多的水印信息被嵌入到原始模型中,可以用于身份表示或安全约束。但该方法对噪声扰动的松弛处理降低了其触发图像的隐蔽性。对于使用触发后门和对抗样本进行黑盒验证的情况,日本筑波大学的Namba和Sakuma(2019)指出通过对访问样本进行修改或检测,可使验证过程失效。为此,Namba和Sakuma(2019)使用干净图像作为访问图像,但是为其标注错误的标签,可以看做是异常数据。为了记住这些异常数据,模型往往会产生过拟合,即只有少量的模型权值在进行预测时被激活。现有的一些模型修改方式如剪枝或微调可以减弱过拟合,从而使模型版权验证失效。为了解决这个问题,模型训练时要求赋予这些异常数据更多的激活权值。

在黑盒验证框架的基础上,近期工作从不同方面对黑盒模型水印进一步改进。为了提高保真度,即减少对原始任务性能的影响,澳大利亚迪肯大学的Zhong等人(2020)在上述方法的基础上,使用一个额外的类别作为触发图像的类别,但该策略并不适用于严格的黑盒验证场景。同样是对模型输出进行设计,Sakazawa等人(2019)训练了一个维度拓展模型,将返回的1维预测值转化为2维图像,通过一个触发图像集的访问,得到一组输出图像,这些图像通过累加的方式得到最终的版权图像,使得验证方式可以直接可视化。Jebreel等人(2021)在原始模型的输出添加了一个额外网络,对于触发图像,原始模型仍然输出其正确标签,随后额外网络输出指定标签。

为了提高鲁棒性,美国杜克大学的Jia等人(2021b)通过使用纠错码使水印提取过程更加鲁棒;新加坡南洋理工大学的Chattopadhyay和Chattopadhyay(2021)通过固定网络层微调的方法使网络的每一层都学习对应的触发样本,从而使简单的模型修改无法擦除嵌入的后门水印。攻击者除了通过直接修改目标模型进行水印擦除外,还可以通过模型窃取攻击得到一个与目标模型功能近似但结构完全不同的替代模型。为了抵抗窃取攻击,芬兰阿尔托大学的Szyller等人(2021)提出了DAWN(dynamic adversarial watermarking of network)的防御方法,在攻击者进行数据访问时,对返回结果进行相应修改,其本质是在攻击者窃取的模型中注入后门从而进行后续的版权认证。类似地,华为加拿大实验室的Charette等人(2022)在输出中注入了特定的信号分布(如余弦分布),使窃取模型也能学习到这种分布。加拿大多伦多大学的Jia等人(2021a)采用了不同策略,同样在原始模型中嵌入后门,不同的是添加了额外的损失函数来约束网络使用同样的特征表达原始数据和后门数据。换句话说,后门特征被紧紧地耦合在正常特征表达中,当窃取模型想要学习到原始特征必然同时学习到后门特征。基于Jia等人(2021a)耦合特征的想法,美国莱斯大学的Tang等人(2021)采用了不同的训练策略,只在原始数据上对训练特征提取器,而使用混合数据训练最后的分类层,换言之,干净数据和后门数据使用了同一个特征提取器,相比Jia等人(2021a)的方法,实验证明该方法性能更优异。新加坡国立大学的Yang等人(2019)采用蒸馏的方法将一个在混合数据集上训练好的含水印模型蒸馏到目标模型上,以此来将水印知识更好地嵌入目标模型。

在安全性方面,黑盒模型水印算法需要考虑伪

造攻击问题。美国芝加哥大学的Li等人(2020a)在模型初始训练阶段使模型的性能和水印模式建立了很强的联系,使模型的局部最优解和水印模式紧密相关。如果攻击者想伪造自己的水印模式,需要花费巨大的计算代价跳出局部最优解,该代价与攻击者自己重新训练一个模型相当,但是只能适用于从起始阶段开始训练的策略,不能应用于已经训练好的模型。Kapusta等人(2021)关注了实际交易平台版权认证环节的伪造问题,并提出了相应的水印协议来解决该问题。与许多攻击防御问题一样,模型水印的嵌入和攻击者的擦除也存在博弈关系,受对抗攻击中可证防御思想的启发,美国马里兰大学的Bansal等人(2022)利用随机平滑策略提出了一种可证明的模型水印方法,证明了当模型权值的改变小于一定范围时,所嵌入的水印不会被擦除。

除了分类任务之外,黑盒模型水印也得以应用于不同的任务中。例如图像处理任务(Quan等,2021)、图像生成任务(Yu等,2021)、语音识别任务(Rathi等,2022)、自然语言处理任务(Yadollahi等,2021; Venugopal 等 , 2011; He 等 , 2022; Tan 等 , 2022)、图片标注(Lim等,2022)、图神经网络(Xu和Picek,2021)、联邦学习(Atli等,2021)、强化学习(Behzadan和Hsu,2019;Chen等,2021a)以及对大型预训练模型的保护(Liu等,2022;Sha等,2022;Cong等,2022)等。可以说,与白盒水印相比,黑盒水印得到了学术界和工业界更多的关注。

### 1.3 脆弱模型水印

上述白盒算法和黑盒算法均侧重水印的鲁棒性,即水印在经受恶意攻击之后仍能可靠地被重构,以鉴定目标模型的版权。在实际应用场景中,除了鉴定模型的版权,还需要对模型的完整性进行验证。为此,研究人员提出了适用的“脆弱”模型水印技术。这里的“脆弱”是指对目标模型的轻微修改必将导致水印难以完美重构,这种“不完美”可以用于鉴别目标模型的完整性。基于此,美国普林斯顿大学的研究人员提出了一种基于敏感样本的黑盒脆弱水印算法(He等,2019),在不需要了解目标模型结构和参数的情况下能够检测木马攻击、数据投毒等引起的模型变化。其思路与对抗样本类似,通过不断对原始图像增加细小的扰动以生成敏感样本(脆弱样本),然后使用敏感样本和初始预测标签组成的数据对来标记目标分类模型;验证时,如果目标分类模型被篡改,其对敏感样本的预测标签将与初始标签不一致。借助模型的改变会影响其对于敏感样本的分类结果来实现篡改检测的黑盒验证,因此,此类方法的本质是寻找或生成尽量靠近分类模型决策边界的敏感样本,一旦模型受到篡改或攻击导致分类边界发生变化,对于这些敏感样本的预测结果就会发生变化。

受上述算法使用敏感样本思想的启发,澳大利亚迪肯大学的Kuttichira等人(2022)提出一种利用贝叶斯优化和变分自编码器找出全局最优的敏感样本的算法,使用变分自编码器将高维数据映射到一个非线性的低维空间,从而生成敏感样本,然后选择贝叶斯优化进行训练找全局最优的敏感样本,对几乎任何种类的攻击都能敏感检测,不足之处是检测效率不够高,一般需要至少10个样本完成检测。

美国马里兰大学帕克分校的Aramoon等人(2021)深入分析了样本与神经元的关联,设计了一个简单有效的损失函数,使敏感样本与模型的大部分参数相关联,提高了可验证模型参数的比例,并增强了输出标签对权重修改的敏感性,使得仅通过访问TOP-1标签就能获得很好的验证效果,不足之处是样本生成需要的迭代数较高。与该方法不同,意大利都灵大学的Botta等人(2021)提出一种具有篡改定位功能的模型参数篡改检测脆弱水印算法。将模型中的参数抽取并按块分割,利用哈希工具、线性变换以及遗传算法将水印嵌入参数块,保证了水印的脆弱性、隐蔽性和保真性。此外,也有研究人员将模型隐私保护的方法引入完整性验证,设计了外包模型完整性验证机制(Dong等,2021)。外包服务器使用双线性配对的方法构建一个模型参数的密文证明,借鉴模型隐私保护的方法对整个模型中的参数进行加密并发给验证者。由于该模型可验证加密算法无法对非线性激活函数及稍复杂的网络结构做出合适的处理,该工作仅对简单的全连接网络有效。

Lao等人(2022)提出了一种能够从原始的预训练模型生成不同版本的受保护模型的工作,可以在验证模型完整性的同时完成用户身份认证。核心思想是利用输入空间、隐藏空间以及逻辑空间制作敏感样本(脆弱样本),从而产生签名,可以从一个原始的预训练模型中生成不同版本的受保护模型,每个模型将对这些敏感样本做出独特的反应,从而在每个神经网络模型中嵌入唯一但脆弱的签名,不足之处在于生成的独特样本有被误判的可能。Li等人

(2021a)提出了一种针对分类神经网络权值的渐进式比特翻转攻击的检测和恢复算法。比特翻转攻击可以通过翻转非常少量的比特导致分类精度的灾难性下降。针对该攻击,算法首先将模型权重参数分组,并采用加法校验为每个组派生2比特签名。随后使用交错权重和掩膜验证,实现了对比特反转攻击的高精度检测。若是被攻击,则将该组中的所有权重归零,以减轻恶意比特翻转造成的准确性下降。算法针对此种特定的比特翻转攻击具有100%的检测率,而且运行速度快,另外还具有一定程度的篡改恢复功能,但通用性方面有待进一步改进。

研究人员还研究了结合频域信息隐藏的模型脆弱水印算法(Abuadbba等,2021),以检测数据投毒、微调以及后门等攻击。该算法首先将神经网络中每一层的参数抽取转化为2维数据并做离散小波变换,然后利用信息隐藏技术将哈希码嵌入到不重要的系数中,最后逆变换转换为空域参数并放回原模型中。在验证端,用户通过提取和对比哈希码来判断模型是否被篡改。图4给出了该方法对应的一般性框架,该脆弱水印算法可归类为白盒水印,不需要对模型进行重新训练,可以检测出比较细微的攻击,水印容量大且对模型的原始任务的影响较小。然而,在完成验证时,用户需要验证每一个层是否被篡改,影响了验证效率;如果攻击方知道水印的嵌入机制,可以将层间的顺序进行替换造成模型功能下降,从而能有效逃避检测。

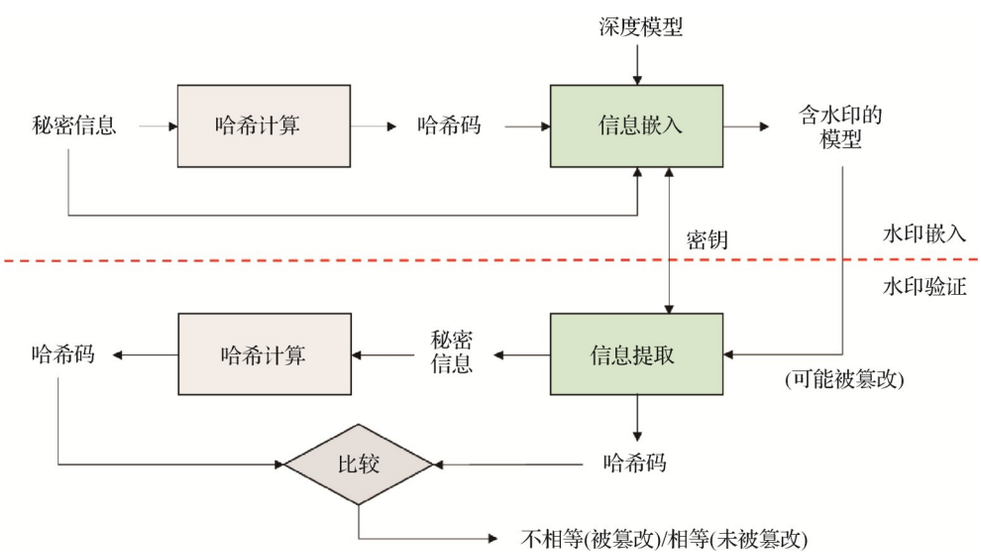


图4 基于哈希校验的脆弱模型水印基本框架示意图

Fig. 4 General framework for fragile model watermarking based on hash verification

## 2 国内研究进展

国内的高校和研究机构在模型水印方面也做出了重要贡献,推动了该领域的发展。接下来,梳理国内相关研究工作。

### 2.1 白盒模型水印

Uchida等人(2017)的方案虽然涉及了扩频水印技术,却未充分利用数字水印理论对白盒模型水印方案性能进行提升。Kuribayashi等人(2020)虽然尝试结合QIM(quantization index modulation)以提升性能,但因其仅选择全连接层作为水印嵌入位置,性能提升不明显且限制了模型水印的嵌入方式。为此,Li等人(2021c)结合扩展变换抖动调制(spread-transform dither-modu-lation, ST-DM),提出了一种可作用在卷积层且不限制嵌入方式的白盒模型水印方案。Uchida等人(2017)提出的水印嵌入方案中,激活函数的选取会对模型水印的性能产生直接影响。受Uchida等人(2017)的启发,Li等人(2021c)构造了基于ST-DM理论的激活函数。ST-DM技术因其优异的隐蔽性和良好的鲁棒性在传统数字水印领域得到广泛应用,然而其中的抖动调制函数是一个高度非线性的函数,直接定义为激活函数无法满足神经网络在训练过程中的反向传播要求。Li等人(2021c)通过大量实验的参数调制,使自定义的激活函数实现对ST-DM的线性拟合。实验结果证明,拟合后的

激活函数同样具有ST-DM的优良性质，在保证鲁棒性的前提下，提升了水印的隐蔽性，扩充了水印的嵌入容量。

为了提升水印的隐蔽性和鲁棒性，Wang等人(2020)提出基于反向传播实现的模型水印方案。通过研究Uchida等人(2017)的方案，Wang等人(2020)发现若是将水印嵌入在已收敛的权重，水印方案具有更好的隐蔽性和鲁棒性。因此，构造了单独的水印嵌入提取模型，以手动或自动模式选取目标模型中的已收敛权重，再通过反向传播完成权重更新，从而实现水印嵌入。同样是选择模型内部参数作为嵌入位置，相较于Uchida等人(2017)的随机权重选取，Liu等人(2021a)提出一种权重选择方案，并将水印嵌入到由筛选出的权重构造的残差值中。由于被挑选的权重对数值修改等多种攻击并不敏感，嵌入其中的水印也具有较强的鲁棒性。

Rouhani等人(2019)提出的Deepsigns方案将水印嵌入在层激活输出部分，但是没有充分利用该部分丰富的数据冗余特性。为了探究该部分的最大嵌入容量，Li等人(2021b)选择卷积层的激活输出，即与输入内容直接相关的特征图为水印的嵌入位置。为保证水印算法的鲁棒性，通过自定义扩频矩阵，水印在嵌入时被分散在整幅特征图上，而不是只影响其中的部分数值。由于特征图随着输入内容的类别变化而改变，该方案还分析了不同类型的输入对水印算法性能的影响。

上述白盒模型水印方案可以较好地抵御权重微调和模型剪枝攻击，却忽略了生成伪水印数据以谎报版权的混淆攻击。为了抵御混淆攻击，Fan等人(2019，2022)在模型卷积层后添加“护照层”。护照层的构造以模型的常规标准化层为基础，其中的缩放参数和平移参数由该层权重和“护照”信息(即水印信息)共同决定。由于模型分类的准确性依赖于水印信息的正确性，在不能提供正确水印信息的情况下，模型分类的准确性会遭到严重的削弱。一个模型匹配一个“护照”，因此该方案可以抵御混淆攻击。同时，实验结果证实在抵御微调和剪枝攻击方面该方案均有良好表现。

然而，Fan等人(2022)强化了护照学习任务和原始任务之间的耦合关系，但造成使用方面的局限性。并且，方案涉及了对目标模型结构的修改，对目标模型的性能造成一定影响。因此，Zhang等人(2020b)提出了不改变模型结构的改进方案。通过在护照层中添加支持护照识别的额外分支，标准化层中仿射变换被分隔为两支，分别处理与护照层相关输入和与护照层无关输入。与护照相关分支在进行训练时，在独立计算自身的标准化统计信息后，再分别学习基于护照信息的仿射变换参数。模型所有者只将不带护照层的目标模型交付给终端用户进行使用，并保存这些秘密护照和额外添加的护照层。所有权验证阶段，将护照层插入可疑模型的相应位置，如果提供护照为真，模型性能将保持不变。

### 2.2 黑盒模型水印

国内学者在黑盒模型水印的研究方面做出了相应的贡献。山东大学的Li等人(2019)使用深度隐藏的方式，将版权信息嵌入在原始输出中以生成对应的触发图像，并约束触发图像和原始干净图像分布相似，进一步提升了触发图片的隐蔽性。同样为了提高触发图像的隐蔽性，武汉大学的Li等人(2020b)在图像频域中嵌入了触发模式。对于触发图像的标签标注问题，厦门大学的Zhang等人(2020c)提出了使用混沌算法进行触发图像的标签标注，来应对逆向分析，从而防御可能存在的伪造攻击。为了更好地满足保真度的要求，南京航空航天大学的Sun等人(2021)使用了额外的类别作为触发类别，但其选取与原始数据分布无关的数据作为触发数据以及基于最低有效比特位的身份信息嵌入策略，无法很好地满足隐蔽性和鲁棒性的要求。百度团队的Yang等人(2021)将黑盒模型水印的训练过程转化为一个双层优化问题，目的是找到和模型紧密绑定的验证样本以及使用尽可能少的权值改动来记住这些验证样本，相比之前的方法，在保真度和鲁棒性上都有了相应的提升。为了抵抗基于触发图像修改(如使用自编码器重构)的逃逸攻击，上海交通大学的Li和Wang(2021)将自编码器的重构过程引入了模型水印嵌入过程，从而提高了对此类逃逸攻击的鲁棒性，但自编码器的引入在一定程度上降低了模型原始任务的性能。为了同时满足触发图像隐蔽性和鲁棒性方面的需求，中国科学技术大学的Zhang等人(2022a)使用图像的边缘结构作为触发模式，并提出了名为“毒墨水”的触发图像生成策略。受传统数字水印中安全问题的启发，复旦大学的Zhu等人(2020)将混淆攻击问题引入到黑盒模型水印中来。为了使版权证明过程中出现纷争，提出了

一系列连续的单向哈希函数对触发图像进行置乱，并对所有置乱的图像赋予相同的预定义标签，水印嵌入过程遵循传统的后门水印方式。由于单向哈希函数的不可逆性，攻击者很难学习到触发模式，进而使验证过程中的触发模式变得独特且安全。但是经过哈希函数置乱后的触发模式接近噪声模式，在验证阶段很难抵抗攻击者基于异常检测的逃逸攻击。

传统的黑盒模型水印基于异常标签的后门攻击方法，但是在为模型嵌入水印的过程中，也会引入相应的安全性问题。例如，当攻击者知道触发模式以及对应目标标签时，攻击者可以触发后门攻击。为了解决这个问题，武汉理工大学的Xu等人(2020b)仍使用触发图像本身的真实标签作为目标标签，但利用触发图像输出中每一类别的置信度分数作为水印信息，不但解决了模型本身的安全隐患，同时增大了水印的容量。相比于传统的黑盒模型水印，特殊的映射关系由分类任务变为了回归任务，为了能够更好地学习到该任务，进一步引入了无关任务情况下的多任务学习框架。同样为了解决上述安全性问题，清华大学的Li等人(2022c)使用无目标的后门攻击方法进行水印嵌入，削弱了特定触发模式和目标标签之间的联系。

同样，国内学者将黑盒模型水印应用到了各种各样的任务上。例如图像处理任务(Zhang等，2020a，2021b，2022b；Zhang，2022c；Wu等，2021)、图像生成任务(Fei等，2022)、语音识别任务(Wang等，2021；Wang和Wu，2022；Chen等，2022)、自然语言处理任务(Dai等，2022；Xiang等，2021；Yang等，2022)、图神经网络(Zhao等，2021)、联邦学习(Li等，2022a；Li，2022b；Liu等，2021b)以及对大型预训练模型的保护(Zhang等，2022d；Wu等，2022)等。

### 2.3 无盒模型水印

除了白盒水印和黑盒水印，国内研究人员提出了无盒模型水印框架(Wu等，2021；Zhang等，2020a)。与白盒水印相比，无盒水印不要求提取者掌握目标模型的内部细节。与黑盒水印相比，无盒水印不要求提取者与目标模型进行直接交互。因此，相对于白盒/黑盒水印，无盒水印框架中提取者掌握的边信息更少。无盒水印主要面向具有生成任务的模型，如图像分割、图像染色和超分辨等。例如，Wu等人(2021)提出了一种适用于云端服务场景、面向图像生成模型的无盒水印算法。如图5所示，该算法联合了两个神经网络(即载体网络和水印提取网络)，通过在模型训练的过程中优化两个网络的参数，使得受保护的网络在完成训练后不仅可以完成原始任务，而且允许验证者利用密钥从输出图像中检测出水印，实现图像和模型的双重产权保护。

中国科学技术大学的Zhang等人(2020a，2022b)也关注了图像处理模型的版权保护问题，通过将水印信息嵌入到图像处理模型的输出分布中，并使用相应的提取网络对水印进行提取。除了考虑针对传统数字图片处理的鲁棒性外，该方案还考虑了对模型窃取攻击的鲁棒性。为了应对攻击者在模型窃取攻击时对获取数据进行预处理的情况，Zhang等人(2021b)还提出了基于结构一致性的模型水印框架。就本质而言，无盒水印在黑盒水印框架上进行了更强的条件约束，可视做黑盒水印的特例。

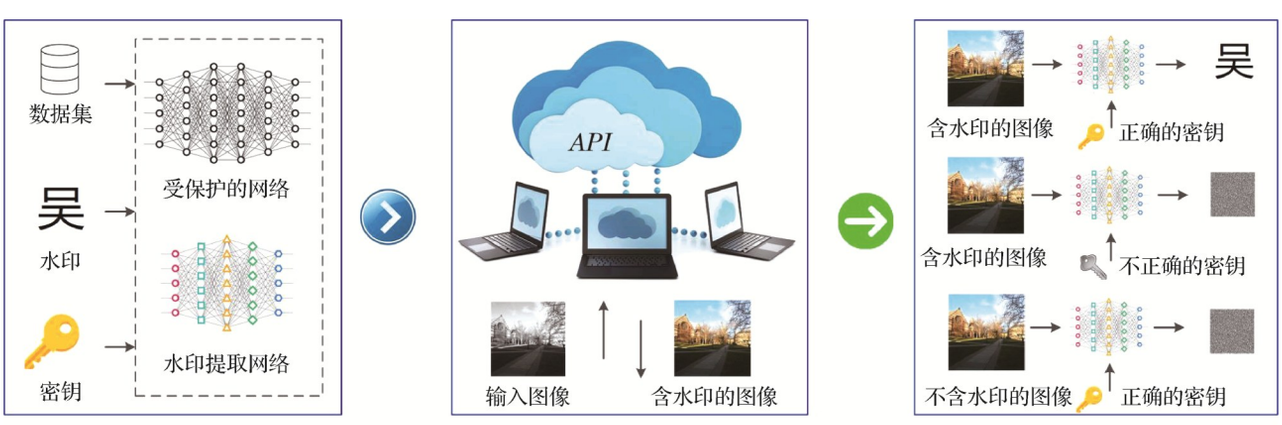


图5 无盒模型水印应用场景示例

Fig. 5 Application scenario for box-free model watermarking

### 2.4 脆弱模型水印

国内学者在脆弱模型水印方面也取得了一些进展。例如，在基于敏感样本的脆弱模型水印方面，电子科技大学的Xu等人(2020a)提出了一种新的敏感样本生成方法，通过结合同态加密实现了隐私保护验证，可以高精度地验证外包给服务器的模型参数的完整性。复旦大学的Zhu等人(2021)提出了一种基于触发集的脆弱模型水印算法。该算法随机产生一些触发图像，并根据密钥标注对应的标签，从而生成触发集。通过两步骤交替训练，即在原始训练数据中加入触发集的数据集和仅包含触发集的数据集上训练，达到预设条件后即完成水印嵌入。两步骤交替训练可以保证含水印模型的性能以及触发图像的有效触发；此外加入了对触发图像预测值求方差的正则化项使触发图像尽可能地向决策边界靠拢，这样模型一旦经过恶意微调，边界就发生变化，对应触发图像标签即发生变化，由此判定模型被投毒攻击或恶意微调等篡改。该工作提出了新颖的两步骤交替训练模式，但该过程也会对目标模型的性能产生轻度影响。

上述方法提供了较好的完整性认证性能，但对受保护的原始模型造成了不可抹去的扰动。为了避免对原始模型造成这种永久性扰动，国内学者探索了可逆模型水印(或称无损模型水印)技术。例如，中国科学技术大学的Guan等人(2020)提出了一种基于可逆水印的模型完整性认证算法，首先利用哈希算法得到模型参数的哈希值，然后结合熵剪枝技术构建卷积层的宿主序列，利用直方图平移技术将哈希值嵌入到宿主序列中。该算法通过利用哈希函数的单向性和低碰撞性实现了模型的完整性认证，并且利用熵剪枝衡量模型中参数的重要性程度将水印嵌入到不重要的参数，从而保证了水印嵌入的保真性。此外，在水印嵌入时采用了可逆信息隐藏中经典的直方图平移算法确保了可逆性。

与上述白盒算法不同，Yin等人(2022)针对黑盒场景提出了一种基于触发集的无损式脆弱水印算法，其验证方式同Zhu等人(2021)的算法。不同的是，该算法不直接使用固定的触发集，而是先训练触发图像生成器，利用PGAN(progressive growing generative adversarial network)以及分类结果的方差损失函数训练触发集生成网络。从决策边界的角度分析，该算法的原理是通过训练触发图像生成器使触发图像不断向目标模型的决策边界边缘靠拢，而不需要对目标模型进行重训练或任何调整，从而能够实现无损认证。

此外，Xiong等人(2022)将感知哈希引入到模型保护中，提出一种具有盗版模型识别和篡改定位能力的模型保护方法。首先分别定义原始模型、微调模型以及篡改模型，设计一种双分支网络结构提取不同模型的特征，用三元组的思想来训练并得到盗版识别哈希序列。同时，通过将模型参数分块，同样用双分支网络提取特征，得到篡改定位哈希序列，最后将两种序列进行拼接。三元组损失函数使得非篡改模型的哈希码与原始模型距离更近，篡改模型的哈希码距离更远。在训练阶段需要给不同类别的模型贴标签，且训练成本较高。该方法不局限于单一形式的模型，对不同大小、结构的卷积神经网络模型结构均适用，且不需要对模型进行人工预处理，直接输入特征提取网络即可提取模型的特征，具有一定的实用性。

Zhao等人(2022)提出一种自嵌入脆弱模型水印方案，不仅能识别和定位模型中被篡改的参数块，还能准确地恢复被破坏的参数。该方案首先根据目标模型特征生成认证数据和恢复数据，然后采用引用共享机制将生成的数据嵌入到模型中，在不影响模型的任务功能的条件下，实现了篡改定位检测和参数恢复功能。

## 3 国内外研究进展比较

### 3.1 白盒模型水印

自2017年日本学者提出白盒模型水印以来，国内外学者围绕白盒模型水印展开了较多的研究。其原因在于，一方面，白盒水印允许水印提取者完全知悉目标模型的内部细节，这为水印的验证提供了巨大的便利条件；另一方面，在白盒场景下，研究人员可以借鉴传统多媒体水印的嵌入思想，为设计白盒模型水印算法提供了技术渠道。从目前已报导的研究成果数量上看，国内外学者在白盒模型水印方面呈现了并驾齐驱的竞争态势。但从抵御特定攻击的角度分析，国内学者的研究较国外学者更为深入。例如，国内学者提出了利用模型结构中的标准化层承载水印，构造了可以抵御混淆攻击的水印嵌入方

案，这种嵌入思想得到了国外学者的借鉴。从国内研究团队来看，上海大学、复旦大学和中国科技大学等高校深耕于白盒模型水印的研究，有一定的竞争优势。但是，相较于国外的研究人员，国内学者在原创性工作方面稍显不足，这为未来的研究目标指明了方向。

### 3.2 黑盒模型水印

黑盒模型水印的基本框架也是由国外学者率先提出，包括Facebook、IBM等知名科技公司的研究团队，研究成果引起了学术界和工业界的广泛关注。表2展示了国内外黑盒模型水印的定性比较。在黑盒模型水印发展的过程中，国内的研究团队做出了

**表2 国内外黑盒模型水印的定性比较**

**Table 2 Qualitative comparison of domestic and foreign black-box model watermarking**

| 方法 | 改进点 | 核心思想 | 侧重点 |
|---|---|---|---|
| Adi等人(2018) | 嵌入模式 | 使用抽象图像 | 黑盒验证 |
| Zhang等人(2018) | 嵌入模式 | 使用随机噪声、无关图像和可见标识 | 通用性 |
| Guo和Potkonjak(2018) | 嵌入模式 | 添加与用户信息有关的噪声 | 可解释 |
| Guo和Potkonjak(2021) | 嵌入模式 | 采用遗传算法搜寻最优模式 | 隐蔽性 |
| Le Merrer等人(2020) | 嵌入模式 | 选择与模型相关的对抗样本 | 完整性 |
| Chen等人(2019b) | 嵌入模式 | 对类别进行分组，表示0,1比特 | 嵌入量 |
| Namba和Sakuma(2019) | 训练策略 | 添加损失约束减弱过拟合 | 鲁棒性 |
| Zhong等人(2020) | 输出模式 | 使用额外类别作为目标类别 | 保真度 |
| Sakazawa等人(2019) | 输出模式 | 使用2维图像进行可视化验证 | 可解释 |
| Jebreel等人(2021) | 输出模式 | 利用额外网络作为验证 | 保真度 |
| Jia等人(2021b) | 训练策略 | 使用水印纠错码 | 鲁棒性 |
| Chattopadhyay和Chattopadhyay(2021) | 训练策略 | 训练网络每一层都嵌入水印 | 鲁棒性 |
| Szyller等人(2021) | 输出模式 | 指定模型输出规则，防止窃取攻击 | 鲁棒性 |
| Charette等人(2022) | 输出模式 | 向模型输出中注入指定信号 | 鲁棒性 |
| Jia等人(2021a) | 训练策略 | 将水印特征和原始任务特征进行耦合 | 鲁棒性 |
| Tang等人(2021) | 训练策略 | 将水印任务和原始任务在分类层进行耦合 | 鲁棒性 |
| Yang等人(2019) | 训练策略 | 使用蒸馏的方式嵌入水印 | 鲁棒性 |
| Li等人(2020a) | 嵌入模式 | 建立水印模式和模型性能的强相关 | 安全性 |
| Kapusta等人(2021) | 嵌入模式 | 采用水印协议解决验证伪造 | 安全性 |
| Bansal等人(2022) | 训练策略 | 可证明的黑盒模型水印 | 安全性 |
| Li等人(2019) | 嵌入模式 | 使用深度网络隐蔽嵌入用户信息 | 隐蔽性 |
| Li等人(2020b) | 嵌入模式 | 将用户信息在图像频域中 | 隐蔽性 |
| Zhang等人(2020c) | 输出模式 | 使用混沌算法选择目标类别 | 安全性 |
| Sun等人(2021) | 输出模式 | 使用额外类别作为目标类别 | 保真度 |
| Yang等人(2021) | 嵌入模式 | 优化寻找与模型紧密绑定的验证样本 | 鲁棒性 |
| Li等人(2021) | 训练策略 | 训练过程纳入自编码器 | 鲁棒性 |
| Zhang等人(2022a) | 嵌入模式 | 使用彩色图像边缘 | 鲁棒性 |
| Zhu等人(2020) | 嵌入模式 | 使用单向哈希函数处理验证模式 | 安全性 |
| Xu等人(2020) | 输出模式 | 使用类别置信度分数 | 安全性 |
| Li等人(2022c) | 输出模式 | 采用无目标攻击 | 安全性 |

重要的贡献。例如，中国科学技术大学的张卫明团队和复旦大学的张新鹏团队，提出了黑盒模型水印在更加真实场景下需要解决的鲁棒性和安全性方面的特殊问题，包括如何抵抗模型窃取攻击和混淆攻击等。张卫明团队提出了一种抵抗窃取攻击的图像处理模型的黑盒模型水印算法，张新鹏团队提出了利用连续单向哈希进行触发图像置乱的验证协议，进而解决混淆攻击的问题。

在黑盒模型的应用方面，国内外的研究进展相差不大，已经探究了在大量任务上的可能性，包括图像处理任务、图像生成任务、语音识别任务、自然语言处理任务、图神经网络、联邦学习和大型预训练模型等。国内工业界如微众银行也展开了对黑盒模型水印的研究，基于其在联邦学习领域的研究经验，主要探讨了联邦学习中如何使用黑盒模型水印进行版权保护。目前，黑盒模型水印的理论研究还非常少，值得一提的工作是来自美国马里兰大学的 Tom Goldstein 团队，利用随机平滑的思想提出了可证明的黑盒模型水印方法，即当模型参数在一定范围内改动时，相应的水印信息仍然能够保持。因此，黑盒模型水印的理论研究值得在未来重点探索。

### 3.3 脆弱模型水印

目前，脆弱模型水印方面的研究成果还相对较少。国外代表性的研究成果包括美国普林斯顿大学提出的基于敏感样本的脆弱水印、意大利都灵大学提出的具有篡改定位功能的脆弱水印、美国蒙特克莱尔州立大学提出的外包模型完整性验证和隐私保护以及美国克莱门森大学提出的能够从原始的预训练模型生成带有不同水印的受保护模型等。国内的研究团队也做出了重要的贡献，例如，复旦大学提出了基于触发集的脆弱水印；华东师范大学实现了无损脆弱水印；中国科学技术大学提出了可逆水印；上海理工大学实现了篡改定位和恢复功能等。总体而言，国内外在脆弱模型水印方面处于并跑的地位。尽管国内外学者在脆弱模型水印方面已经取得了积极进展，但这一研究方向目前还处在发展初期，值得进一步研究。

## 4 发展趋势与展望

### 4.1 白盒模型水印

从隐蔽性、鲁棒性和嵌入量这 3 个主要指标来看，白盒模型水印方案设计时隐蔽性和鲁棒性的优先级要高于嵌入量。考虑到版权保护的应用场景，水印的鲁棒性是需要重点考量的指标。目前，白盒水印方案的鲁棒性验证多侧重于权重微调、模型剪枝等基础攻击，对抵御如混淆攻击、共谋攻击和窃取攻击等进阶攻击方式的研究相对较少。未来研究中，借鉴传统数字水印理论和深度神经网络可解释性理论或将有助于模型水印鲁棒性的提升。其次，考虑到白盒模型水印方案的特性，当前主流的白盒模型水印方案存在验证即暴露的问题。验证过程中，所有者对模型内部信息使用权的让渡，会直接导致不可信的第三方或攻击者对模型的滥用，无法达到模型版权保护的初衷。借鉴零知识证明来设计强隐蔽性的白盒模型水印方案或许是解决该问题的潜在技术之一。此外，主流白盒水印方案多侧重于基于卷积神经网络的图像分类任务。后续方案可拓宽应用范围，针对不同模型结构的特点进行调整，提出相适应的白盒模型水印方案。

### 4.2 黑盒模型水印

就黑盒模型水印而言，在迁移学习中攻击者窃取的模型会用于新的任务，现有方法还不能有效解决该场景下的所有权认证；此外，黑盒模型水印的水印容量相对白盒模型水印而言是有限的。因此，如何增强黑盒模型水印在更多真实场景下的鲁棒性以及如何满足大容量需求是值得研究的方向。此外，深度模型版权保护研究可以借鉴相关交叉领域的研究，诸如模型的可解释性、模型或数据的隐私保护、模型或数据的公平性。这些方向与模型版权保护或多或少地存在着一定的联系。例如，对模型训练的数据添加水印是否会带来公平性的问题，模型水印的相关技术是否增加了模型隐私被窃取的风险等。因此，梳理各方向之间的关系，设计可控制可调节的模型版权保护技术也是一个值得研究的方向。对于可证明的黑盒模型水印而言，对于网络参数修改的限定并不一定符合模型窃取过程中的参数的真实改变情况。窃取模型相对原始模型在什么层面的改变是更容易约束的，利用这些特性变化实现可证明黑盒模型水印也是未来值得研究的方向。

### 4.3 脆弱模型水印

脆弱水印是篡改检测和完整性认证领域重要的主动防御手段。回顾多媒体脆弱水印技术的演进可以看到，该技术从最初的检测是否被篡改发展到能

够定位具体篡改位置且定位精度不断提高,然后结合可逆信息隐藏技术发展出能够恢复篡改内容的脆弱水印。就多媒体脆弱水印而言,首先,具有篡改定位并恢复功能的脆弱水印具有其特有的应用场景和价值,是白盒条件下模型脆弱水印值得研究的课题;其次,设计和验证能够对模型篡改分级、检测模型不同篡改程度的半脆弱水印是值得研究的另一个课题;最后,现有的脆弱水印算法大都是面向分类网络,面向生成式网络等其他任务的人工智能模型脆弱水印也是值得期待的研究。

**致　谢**　本文由中国图象图形学学会数字媒体取证与安全专委会组织撰写,该专委会更多详情请见链接:http://www.csig.org.cn/detail/2450。

## 参考文献(References)

Abuadbba A, Kim H and Nepal S. 2021. DeepiSign: invisible fragile watermark to protect the integrity and authenticity of CNN//The 36th Annual ACM Symposium on Applied Computing. [s.l.]: ACM: 952-959 [DOI: 10.1145/3412841.3441970]

Adi Y, Baum C, Cissé M, Pinkas B and Keshet J. 2018. Turning your weakness into a strength: watermarking deep neural networks by backdooring//The 27th USENIX Security Symposium (USENIX Security 18). Baltimore, USA: USENIX Association: 1615-1631

Aramoon O, Chen P Y and Qu G. 2021. AID: attesting the integrity of deep neural networks//Proceedings of the 58th ACM/IEEE Design Automation Conference (DAC). San Francisco, USA: IEEE: 19-24 [DOI: 10.1109/DAC18074.2021.9586290]

Atli B G, Xia Y X, Marchal S and Asokan N. 2021. WAFFLE: watermarking in federated learning [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2008.07298.pdf

Bansal A, Chiang P Y, Curry M J, Jain R, Wigington C, Manjunatha V, Dickerson J P and Goldstein T. 2022. Certified neural network watermarks with randomized smoothing//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: ICML: 1450-1465

Behzadan V and Hsu W. 2019. Sequential triggers for watermarking of deep reinforcement learning policies [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/1906.01126.pdf

Botta M, Cavagnino D and Esposito R. 2021. NeuNAC: a novel fragile watermarking algorithm for integrity protection of neural networks. Information Sciences, 576: 228-241 [DOI: 10.1016/j.ins.2021.06.073]

Charette L, Chu L Y, Chen Y Z, Pei J, Wang L J and Zhang Y. 2022. Cosine model watermarking against ensemble distillation//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI Press: 9512-9520 [DOI: 10.1609/aaai.v36i9.21184]

Chattopadhyay N and Chattopadhyay A. 2021. ROWBACK: robust watermarking for neural networks using backdoors//Proceedings of the 20th IEEE International Conference on Machine Learning and Applications (ICMLA). Pasadena, USA: IEEE: 1728-1735 [DOI: 10.1109/ICMLA52953.2021.00274]

Chen H L, Rouhani B D, Fu C, Zhao J S and Koushanfar F. 2019a. DeepMarks: a secure fingerprinting framework for digital rights management of deep learning models//Proceedings of 2019 on International Conference on Multimedia Retrieval. Ottawa, Canada: ACM: 105-113 [DOI: 10.1145/3323873.3325042]

Chen H L, Rouhani B D and Koushanfar F. 2019b. Blackmarks: Black-box multibit watermarking for deep neural networks [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/1904.00344.pdf

Chen H L, Rouhani B D and Koushanfar F. 2020. SpecMark: a spectral watermarking framework for IP protection of speech recognition systems//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 2312-2316 [DOI: 10.21437/Interspeech.2020-2787]

Chen H Z, Zhang W M, Liu K L, Chen K J, Fang H and Yu N H. 2022. Speech pattern based black-box model watermarking for automatic speech recognition//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 3059-3063 [DOI: 10.1109/ICASSP43922.2022.9747044]

Chen K J, Guo S W, Zhang T W, Li S X and Liu Y. 2021a. Temporal watermarks for deep reinforcement learning models//Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. [s.l.]: International Foundation for Autonomous Agents and Multiagent Systems: 314-322

Chen X X, Chen T L, Zhang Z Y and Wang Z Y. 2021b. You are caught stealing my winning lottery ticket! Making a lottery ticket claim its ownership//Proceedings of the 35th Advances in Neural Information Processing Systems. [s.l.]: [s.n.]: 1780-1791

Cong T S, He X L and Zhang Y. 2022. SSLGuard: a watermarking scheme for self-supervised learning pre-trained encoders//Proceedings of 2022 ACM SIGSAC Conference on Computer and Communications Security. Los Angeles, USA: ACM: 579-593

Cortiñas-Lorenzo B and Pérez-González F. 2020. Adam and the ants: on the influence of the optimization algorithm on the detectability of DNN watermarks. Entropy, 22(12): #1379 [DOI: 10.3390/e22121379]

Dai L, Mao J R, Fan X F and Zhou X Y. 2022. DeepHider: a multi-module and invisibility watermarking scheme for language model [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2208.04676v1.pdf

Dong B X, Zhang B and Wang H. 2021. VeriDL: integrity verification of outsourced deep learning services//Proceedings of 2021 Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Bilbao, Spain: Springer: 583-598 [DOI: 10.1007/978-

3-030-86520-7_36]

Fan L X, Ng K W and Chan C S. 2019. Rethinking deep neural network ownership verification: embedding passports to defeat ambiguity attacks//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 4714-4723

Fan L X, Ng K W, Chan C S and Yang Q. 2022. DeepIPR: deep neural network ownership verification with passports. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10): 6122-6139 [DOI: 10.1109/TPAMI.2021.3088846]

Fei J W, Xia Z H, Tondi B and Barni M. 2022. Supervised GAN watermarking for intellectual property protection//2022 IEEE International Workshop on Information Forensics and Security (WIFS). Shanghai, China: IEEE: #9975409 [DOI: 10.1109/WIFS55849.2022.9975409]

Frankle J and Carbin M. 2019. The lottery ticket hypothesis: finding sparse, trainable neural networks//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ICLR

Guan X Q, Feng H M, Zhang W M, Zhou H, Zhang J and Yu N H. 2020. Reversible watermarking in deep convolutional neural networks for integrity authentication//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2273-2280 [DOI: 10.1145/3394171.3413729]

Guo J and Potkonjak M. 2018. Watermarking deep neural networks for embedded systems//Proceedings of 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). San Diego, USA: IEEE: #3740862 [DOI: 10.1145/3240765.3240862]

Guo J and Potkonjak M. 2021. Evolutionary trigger set generation for DNN black-box watermarking [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/1906.04411.pdf

He X L, Xu Q K, Lyu L J, Wu F Z and Wang C G. 2022. Protecting intellectual property of language generation APIs with lexical watermark//Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI: 10758-10766 [DOI: 10.1609/aaai.v36i10.21321]

He Z C, Zhang T W and Lee R. 2019. Sensitive-sample fingerprinting of deep neural networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4724-4732 [DOI: 10.1109/CVPR.2019.00486]

Jebreel N M, Domingo-Ferrer J, Sánchez D and Blanco-Justicia A. 2021. KeyNet: an asymmetric key-style framework for watermarking deep learning models. Applied Sciences, 11(3): #999 [DOI: 10.3390/app11030999]

Jia H R, Choquette-Choo C A, Chandrasekaran V and Papernot N. 2021a. Entangled watermarks as a defense against model extraction//The 30th USENIX Security Symposium. [s.l.]: USENIX Association: 1937-1954

Jia J Y, Wang B H and Gong N Z. 2021b. Robust and verifiable information embedding attacks to deep neural networks via error-correcting codes//Proceedings of 2021 ACM Asia Conference on Computer and Communications Security. Hong Kong, China: ACM: 2-13 [DOI: 10.1145/3433210.3437519]

Kapusta K, Thouvenot V, Bettan O, Beguinet H and Senet H. 2021. A protocol for secure verification of watermarks embedded into machine learning models//Proceedings of 2021 ACM Workshop on Information Hiding and Multimedia Security. Brussels, Belgium: ACM: 171-176 [DOI: 10.1145/3437880.3460409]

Kuribayashi M, Tanaka T and Funabiki N. 2020. DeepWatermark: embedding watermark into DNN model//Proceedings of 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Auckland, New Zealand: IEEE: 1340-1346

Kuttichira D P, Gupta S, Nguyen D, Rana S and Venkatesh S. 2022. Verification of integrity of deployed deep learning models using Bayesian optimization. Knowledge-Based Systems, 241: #108238 [DOI: 10.1016/j.knosys.2022.108238]

Lao Y J, Zhao W J, Yang P and Li P. 2022. DeepAuth: a DNN authentication framework by model-unique and fragile signature embedding//Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI: 9595-9603 [DOI: 10.1609/aaai.v36i9.21193]

Le Merrer E, Perez P and Trédan G. 2020. Adversarial frontier stitching for remote neural network watermarking. Neural Computing and Applications, 32(13): 9233-9244 [DOI: 10.1007/s00521-019-04434-z]

Li B W, Fan L X, Gu H L, Li J and Yang Q. 2022a. FedIPR: ownership verification for federated deep neural network models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4521-4536 [DOI: 10.1109/TPAMI.2022.3195956]

Li F Q and Wang S L. 2021. Persistent watermark for image classification neural networks by penetrating the autoencoder//Proceedings of 2021 IEEE International Conference on Image Processing (ICIP). Anchorage, USA: IEEE: 3063-3067 [DOI: 10.1109/ICIP42928.2021.9506368]

Li F Q, Wang S L and Liew A W C. 2022b. Watermarking protocol for deep neural network ownership regulation in federated learning//Proceedings of 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Taipei, China: IEEE: #9859395 [DOI: 10.1109/ICMEW56448.2022.9859395]

Li H Y, Wenger E, Shan S, Zhao B Y and Zheng H T. 2020a. Piracy resistant watermarks for deep neural networks [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/1910.01226.pdf

Li J T, Rakin A S, He Z Z, Fan D L and Chakrabarti C. 2021a. RADAR: run-time adversarial weight attack detection and accuracy recovery//Proceedings of 2021 Design, Automation and Test in Europe Conference and Exhibition (DATE). Grenoble, France: IEEE: 790-795 [DOI: 10.23919/DATE51398.2021.9474113]

Li M, Zhong Q, Zhang L Y, Du Y J, Zhang J and Xiang Y. 2020b. Protecting the intellectual property of deep neural networks with watermarking: the frequency domain approach//Proceedings of the 19th

IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Guangzhou, China: IEEE: 402-409 [DOI: 10.1109/TrustCom50675.2020.00062]

Li Y, Abady L, Wang H X and Barni M. 2021b. A feature-map-based large-payload DNN watermarking algorithm//Proceedings of the 20th International Workshop on Digital Watermarking. Beijing, China: Springer: 135-148 [DOI: 10.1007/978-3-030-95398-0_10]

Li Y, Tondi B and Barni M. 2021c. Spread-transform dither modulation watermarking of deep neural network. Journal of Information Security and Applications, 63: #103004 [DOI: 10.1016/j.jisa.2021.103004]

Li Y M, Bai Y, Jiang Y, Yang Y, Xia S T and Li B. 2022c. Untargeted backdoor watermark: towards harmless and stealthy dataset copyright protection [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2210.00875.pdf

Li Z, Hu C Y, Zhang Y and Guo S Q. 2019. How to prove your model belongs to you: a blind-watermark based framework to protect intellectual property of DNN//Proceedings of the 35th Annual Computer Security Applications Conference. San Juan, USA: ACM: 126-137 [DOI: 10.1145/3359789.3359801]

Lim J H, Chan C S, Ng K W, Fan L X and Yang Q. 2022. Protect, show, attend and tell: empowering image captioning models with ownership protection. Pattern Recognition, 122: #108285 [DOI: 10.1016/j.patcog.2021.108285]

Liu H W, Weng Z Y and Zhu Y S. 2021a. Watermarking deep neural networks with greedy residuals//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: ICML: 6978-6988

Liu X Y, Shao S, Yang Y, Wu K M, Yang W Y and Fang H. 2021b. Secure federated learning model verification: a client-side backdoor triggered watermarking scheme//Proceedings of 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Melbourne, Australia: IEEE: 2414-2419 [DOI: 10.1109/SMC52423.2021.9658998]

Liu Y P, Jia J Y, Liu H B and Gong N Z. 2022. StolenEncoder: stealing pre-trained encoders in self-supervised learning//Proceedings of 2022 ACM SIGSAC Conference on Computer and Communications Security. Los Angeles, USA: ACM: 2115-2128 [DOI: 10.1145/3548606.3560586]

Lou X X, Guo S W, Li J W and Zhang T W. 2022. Ownership verification of DNN architectures via hardware cache side channels. IEEE Transactions on Circuits and Systems for Video Technology, 32(11): 8078-8093 [DOI: 10.1109/TCSVT.2022.3184644]

Namba R and Sakuma J. 2019. Robust watermarking of neural network with exponential weighting//Proceedings of 2019 ACM Asia Conference on Computer and Communications Security. Auckland, New Zealand: ACM: 228-240 [DOI: 10.1145/3321705.3329808]

Ong D S, Chan C S, Ng K W, Fan L X and Yang Q. 2021. Protecting intellectual property of generative adversarial networks from ambiguity attacks//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3629-3638 [DOI: 10.1109/CVPR46437.2021.00363]

Quan Y H, Teng H, Chen Y X and Ji H. 2021. Watermarking deep neural networks in image processing. IEEE Transactions on Neural Networks and Learning Systems, 32(5): 1852-1865 [DOI: 10.1109/TNNLS.2020.2991378]

Rathi P, Bhadauria S and Rathi S. 2022. Watermarking of deep recurrent neural network using adversarial examples to protect intellectual property. Applied Artificial Intelligence, 36(1): #2008613 [DOI: 10.1080/08839514.2021.2008613]

Rouhani B D, Chen H L and Koushanfar F. 2019. DeepSigns: an end-to-end watermarking framework for ownership protection of deep neural networks//Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems. Providence, USA: ACM: 485-497 [DOI: 10.1145/3297858.3304051]

Sakazawa S, Myodo E, Tasaka K and Yanagihara H. 2019. Visual decoding of hidden watermark in trained deep neural network//Proceedings of 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). San Jose, USA: IEEE: 371-374 [DOI: 10.1109/MIPR.2019.00073]

Sha Z Y, He X L, Yu N, Backes M and Zhang Y. 2022. Can't steal? Cont-steal! Contrastive stealing attacks against image encoders [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2201.07513.pdf

Sun S, Zhang W M, Fang H and Yu N H. 2022. Automatic generation of Chinese document watermarking fonts. Journal of Image and Graphics, 27(1): 262-276 (孙杉, 张卫明, 方涵, 俞能海. 2022. 中文水印字库的自动生成方法. 中国图象图形学报, 27(1): 262-276) [DOI: 10.11834/jig.200695]

Sun S C, Xue M F, Wang J and Liu W Q. 2021. Protecting the intellectual properties of deep neural networks with an additional class and steganographic images [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2104.09203.pdf

Szyller S, Atli B G, Marchal S and Asokan N. 2021. DAWN: dynamic adversarial watermarking of neural networks//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 4417-4425 [DOI: 10.1145/3474085.3475591]

Tan Z Q, Wong H S and Chan C S. 2022. An embarrassingly simple approach for intellectual property rights protection on recurrent neural networks//Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. [s.l.]: Association for Computational Linguistics: 93-105

Tang R X, Jin H Y, Wigington C, Du M N, Jain R and Hu X. 2021. Was my model stolen? Feature sharing for robust and transferable watermarks[EB/OL]. [2023-01-06]. https://openreview.net/forum?id=XHxRBwjpEQ

Tartaglione E, Grangetto M, Cavagnino D and Botta M. 2021. Delving

in the loss landscape to embed robust watermarks into neural networks//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 1243-1250 [DOI: 10.1109/ICPR48806.2021.9413062]

Uchida Y, Nagai Y, Sakazawa S and Satoh S. 2017. Embedding watermarks into deep neural networks//Proceedings of 2017 ACM on International Conference on Multimedia Retrieval. Bucharest, Romania: ACM: 269-277 [DOI: 10.1145/3078971.3078974]

Venugopal A, Uszkoreit J, Talbot D, Och F J and Ganitkevitch J. 2011. Watermarking the outputs of structured prediction with an application in statistical machine translation//Proceedings of 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK: ACL: 1363-1372

Wang J F, Wu H Z, Zhang X P and Yao Y W. 2020. Watermarking in deep neural networks via error back-propagation. Electronic Imaging, 32 (4): #00003 [DOI: 10.2352/ISSN. 2470-1173.2020.4.MWSF-022]

Wang T H and Kerschbaum F. 2019. Attacks on digital watermarks for deep neural networks//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE: 2622-2626 [DOI: 10.1109/ICASSP.2019.8682202]

Wang T H and Kerschbaum F. 2021. RIGA: covert and robust white-box watermarking of deep neural networks//Proceedings of 2021 Web Conference. Ljubljana, Slovenia: ACM: 993-1004 [DOI: 10.1145/3442381.3450000]

Wang Y F, Zhou Y M, Qian Z X, Li S and Zhang X P. 2022. Review of robust video watermarking. Journal of Image and Graphics, 27(1): 27-42 (王翌妃, 周杨铭, 钱振兴, 李晟, 张新鹏. 2022. 鲁棒视频水印研究进展. 中国图象图形学报, 27(1): 27-42) [DOI: 10.11834/jig.210437]

Wang Y M and Wu H Z. 2022. Protecting the intellectual property of speaker recognition model by black-box watermarking in the frequency domain. Symmetry, 14 (3): #619 [DOI: 10.3390/sym14030619]

Wang Y M, Ye J Y and Wu H Z. 2021. Generating watermarked speech adversarial examples//Proceedings of 2021 ACM Turing Award Celebration Conference-China (ACM TURC 2021). Hefei, China: ACM: 254-260 [DOI: 10.1145/3472634.3474080]

Wu H Z, Liu G, Yao Y W and Zhang X P. 2021. Watermarking neural networks with watermarked images. IEEE Transactions on Circuits and Systems for Video Technology, 31(7): 2591-2601 [DOI: 10.1109/TCSVT.2020.3030671]

Wu Y T, Qiu H, Zhang T W and Qiu M K. 2022. Watermarking pre-trained encoders in contrastive learning//Proceedings of the 4th International Conference on Data Intelligence and Security. Shenzhen, China: IEEE: 228-233 [DOI: 10.1109/ICDIS55630.2022.00042]

Xiang T, Xie C L, Guo S W, Li J W and Zhang T W. 2021. Protecting your NLG models with semantic and robust watermarks [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2112.05428.pdf

Xiong C, Feng G R, Li X R, Zhang X P and Qin C. 2022. Neural network model protection with piracy identification and tampering localization capability//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 2881-2889 [DOI: 10.1145/3503161.3548247]

Xu G W, Li H W, Ren H, Sun J F, Xu S M, Ning J T, Yang H M, Yang K and Deng R H. 2020a. Secure and verifiable inference in deep neural networks//Proceedings of 2020 Annual Computer Security Applications Conference. Austin, USA: ACM: 784-797 [DOI: 10.1145/3427228.3427232]

Xu J and Picek S. 2021. Watermarking graph neural networks based on backdoor attacks [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2110.11024v1.pdf

Xu X R, Li Y Q and Yuan C. 2020b. "Identity bracelets" for deep neural networks. IEEE Access, 8: 102065-102074 [DOI: 10.1109/ACCESS.2020.2998784]

Yadollahi M M, Shoeleh F, Dadkhah S and Ghorbani A A. 2021. Robust black-box watermarking for deep neural network using inverse document frequency//Proceedings of 2021 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). AB, Canada: IEEE: 574-581 [DOI: 10.1109/DASC-PICom-CBDCom-CyberSciTech52372.2021.00100]

Yang P, Lao Y J and Li P. 2021. Robust watermarking for deep neural networks via bi-level optimization//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14821-14830 [DOI: 10.1109/ICCV48922.2021.01457]

Yang X, Zhang J, Chen K J, Zhang W M, Ma Z H, Wang F and Yu N H. 2022. Tracing text provenance via context-aware lexical substitution. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10): 11613-11621 [DOI: 10.1609/aaai.v36i10.21415]

Yang Z Q, Dang H and Chang E C. 2019. Effectiveness of distillation attack and countermeasure on neural network watermarking [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/1906.06046.pdf

Yin Z X, Yin H and Zhang X P. 2022. Neural network fragile watermarking with no model performance degradation//Proceedings of 2022 IEEE International Conference on Image Processing (ICIP). Bordeaux, France: IEEE: 3958-3962 [DOI: 10.1109/ICIP46576.2022.9897413]

Yu N, Skripniuk V, Abdelnabi S and Fritz M. 2021. Artificial fingerprinting for generative models: rooting deepfake attribution in training data//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14428-14437 [DOI: 10.1109/ICCV48922.2021.01418]

Zhang C Y, Bengio S, Hardt M, Recht B and Vinyals O. 2021a. Under-

standing deep learning (still) requires rethinking generalization. Communications of the ACM, 64(3): 107-115 [DOI: 10.1145/3446776]

Zhang J, Chen D D, Huang Q D, Liao J, Zhang W M, Feng H M, Hua G and Yu N H. 2022a. Poison ink: robust and invisible backdoor attack. IEEE Transactions on Image Processing, 31: 5691-5705 [DOI: 10.1109/TIP.2022.3201472]

Zhang J, Chen D D, Liao J, Fang H, Ma Z H, Zhang W M, Hua G and Yu N H. 2021b. Exploring structure consistency for deep model watermarking [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2108.02360.pdf

Zhang J, Chen D D, Liao J, Fang H, Zhang W M, Zhou W B, Cui H and Yu N H. 2020a. Model watermarking for image processing networks. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 12805-12812 [DOI: 10.1609/aaai.v34i07.6976]

Zhang J, Chen D D, Liao J, Zhang W M, Feng H M, Hua G and Yu N H. 2022b. Deep model intellectual property protection via deep watermarking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(8): 4005-4020 [DOI: 10.1109/TPAMI.2021.3064850]

Zhang J, Chen D D, Liao J, Zhang W M, Hua G and Yu N H. 2020b. Passport-aware normalization for deep model protection//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #1896

Zhang J L, Gu Z S, Jang J, Wu H, Stoecklin M P, Huang H Q and Molloy I. 2018. Protecting intellectual property of deep neural networks with watermarking//Proceedings of 2018 on Asia Conference on Computer and Communications Security. Incheon, Korea (South): ACM: 159-172 [DOI: 10.1145/3196494.3196550]

Zhang L, Liu Y, Liu S T, Yang T S, Wang Y X, Zhang X P and Wu H Z. 2022c. Generative model watermarking based on human visual system [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2209.15268.pdf

Zhang T X, Wu H Z, Lu X F and Sun G L. 2022d. AWEncoder: adversarial watermarking pre-trained encoders in contrastive learning [EB/OL]. [2023-01-06]. https://arxiv.org/pdf/2208.03948.pdf

Zhang Y Q, Jia Y R, Wang X Y, Niu Q and Chen N D. 2020c. DeepTrigger: a watermarking scheme of deep learning models based on chaotic automatic data annotation. IEEE Access, 8: 213296-213305 [DOI: 10.1109/ACCESS.2020.3039323]

Zhao G J, Qin C, Yao H and Han Y F. 2022. DNN self-embedding watermarking: towards tampering detection and parameter recovery for deep neural network. Pattern Recognition Letters, 164: 16-22 [DOI: 10.1016/j.patrec.2022.10.013]

Zhao X Y, Wu H Z and Zhang X P. 2021. Watermarking graph neural networks by random graphs//The 9th International Symposium on Digital Forensics and Security (ISDFS). Elazig, Turkey: IEEE: 1-6 [DOI: 10.1109/ISDFS52919.2021.9486352]

Zheng G, Hu D H, Ge H and Zheng S L. 2021. End-to-end image steganography and watermarking driven by generative adversarial networks. Journal of Image and Graphics, 26(10): 2485-2502 (郑钢, 胡东辉, 戈辉, 郑淑丽. 2021. 生成对抗网络驱动的图像隐写与水印模型. 中国图象图形学报, 26(10): 2485-2502) [DOI: 10.11834/jig.200404]

Zhong Q, Zhang L Y, Zhang J, Gao L X and Xiang Y. 2020. Protecting IP of deep neural networks with watermarking: a new label helps//Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, Singapore: Springer: 462-474 [DOI: 10.1007/978-3-030-47436-2_35]

Zhu R J, Wei P, Li S, Yin Z X, Zhang X P and Qian Z X. 2021. Fragile Neural Network Watermarking with Trigger Image Set//Proceedings of the 14th International Conference on Knowledge Science, Engineering and Management. Tokyo, Japan: Springer: 280-293 [DOI: 10.1007/978-3-030-82136-4_23]

Zhu R J, Zhang X P, Shi M T and Tang Z J. 2020. Secure neural network watermarking protocol against forging attack. EURASIP Journal on Image and Video Processing, 2020(1): #37 [DOI: 10.1186/s13640-020-00527-1]

## 作者简介

吴汉舟,男,副教授,主要研究方向为信息隐藏。
E-mail: hanzhou@shu.edu.cn
张新鹏,通信作者,男,教授,主要研究方向为多媒体信息安全、人工智能安全和图像处理。E-mail: xzhang@shu.edu.cn
张杰,男,博士,主要研究方向为信息隐藏与人工智能安全。
E-mail: zjzac@ustc.edu.cn
李越,女,博士,主要研究方向为信息隐藏与人工智能安全。
E-mail: liyue_0119@hqu.edu.cn
殷赵霞,女,教授,主要研究方向为信息隐藏与人工智能安全。E-mail: zxyin@cee.ecnu.edu.cn
田晖,男,教授,主要研究方向为信息隐藏、人工智能安全和云计算安全。E-mail: htian@hqu.edu.cn
李斌,男,教授,主要研究方向为多媒体取证、多媒体安全和信息隐藏。E-mail: libin@szu.edu.cn
张卫明,男,教授,主要研究方向为信息隐藏与人工智能安全。E-mail: zhangwm@ustc.edu.cn
俞能海,男,教授,主要研究方向为多媒体内容安全。
E-mail: ynh@ustc.edu.cn