基于差分隐私的深度伪造指纹检测模型版权保护算法

袁程胜[1,2]，郭强[1,2]，付章杰[1,2]

（1. 南京信息工程大学计算机学院、软件学院、网络空间安全学院，江苏 南京 210044；
2. 南京信息工程大学数字取证教育部工程研究中心，江苏 南京 210044）

摘 要：提出了一种基于差分隐私的深度伪造指纹检测模型版权保护算法，在不削弱原始任务性能的同时，实现了深度伪造指纹检测模型版权的主动保护和被动验证。在原始任务训练时，通过添加噪声以引入随机性，利用差分隐私算法的期望稳定性进行分类决策，以削弱对噪声的敏感。在被动验证中，利用 FGSM 生成对抗样本，通过微调决策边界以建立后门，将后门映射关系作为植入水印实现被动验证。为了解决多后门造成的版权混淆，设计了一种水印验证框架，对触发后门加盖时间戳，借助时间顺序来鉴别版权。在主动保护中，为了给用户提供分等级的服务，通过概率选择策略冻结任务中的关键性神经元，设计访问权限实现神经元的解冻，以获得原始任务的使用权。实验结果表明，不同模型性能下的后门验证依然有效，嵌入的后门对模型修改表现出稳健性。此外，所提算法不但能抵挡攻击者策反合法用户实施的合谋攻击，而且能抵挡模型修改发动的微调、压缩等攻击。

关键词：版权保护；对抗样本；差分隐私；模型水印；伪造指纹检测

中图分类号：TP391

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2022184

Copyright protection algorithm based on differential privacy deep fake fingerprint detection model

YUAN Chengsheng[1,2], GUO Qiang[1,2], FU Zhangjie[1,2]

1. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China
2. Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China

Abstract: A copyright protection algorithm based on differential privacy for deep fake fingerprint detection model (DFFDM) was proposed, realizing active copyright protection and passive copyright verification of DFFDM without weakening the performance of the original task. In the original task training, noise was added to introduce randomness, and the expected stability of the differential privacy algorithm was used to make classification decisions to reduce the sensitivity to noise. In passive verification, FGSM was used to generate adversarial samples, the decision boundary was fine-adjusted to establish a backdoor, and the mapping was used as an implanted watermark to realize passive verification. To solve the copyright confusion caused by multiple backdoors, a watermark verification framework was designed, which stamped the trigger backdoors and identified the copyright with the help of time order. In active protection, to provide users with hierarchical services, the key neurons in the task were frozen by probabilistic selection strategy, and the access rights were designed to realize the thawing of neurons, so as to obtain the right to use the original task. Experimental results show that the backdoor verification is still effective under different model performance, and the embedded backdoor shows a certain robustness to the model modification. Also, the proposed algorithm can resist not only the collusion attack by the attacker to recruit legitimate users, but also the fine-tuning and compression attacks caused by the model modification.

Keywords: copyright protection, adversarial samples, differential privacy, model watermark, fake fingerprint detection

收稿日期：2022−06−01；**修回日期：**2022−09−08

基金项目：国家自然科学基金资助项目（No.62102189）；江苏省自然科学基金资助项目（No.BK20200807, No.BK20200039）；国防科技大学科研计划基金资助项目（No.JS21-4）；浙江省科技厅公益性科技产业基金资助项目（No.LGF21F020006）

Foundation Items: The National Natural Science Foundation of China (No.62102189), The Natural Science Foundation of Jiangsu Province (No.BK20200807, No.BK20200039), NUDT Scientific Research Program (No.JS21-4), Public Welfare Technology and Industry Project of Zhejiang Provincial Science Technology Department (No.LGF21F020006)

0 引言

随着大数据技术和数字经济的蓬勃发展，互联网每天都会产生海量的数据，部分数据不仅关系到个人隐私权益，还会涉及国家安全和社会公共利益。为了避免敏感隐私数据外泄，对其进行安全访问至关重要。生物识别技术（借用人体生理或行为特性）作为一种新颖的身份识别模式，逐渐替代传统的密码验证。在现有的生物特征中，指纹因具有唯一性、稳定性和长久不变性的特性，应用更普及。截至 2021 年，指纹识别占据全球生物识别的大部分市场份额[1]。但是，该技术存在严重的安全隐患，借助硅胶、树脂、明胶等材料伪造的指纹能够成功欺骗指纹识别系统。因此，伪造指纹[2]检测技术被提出。

近些年，随着机器学习尤其是深度学习的迅速发展，人工智能技术被广泛应用在无人驾驶[3-4]、计算机视觉[5-6]、自然语言处理[7-8]等领域，基于深度学习的深度伪造指纹检测方法[9]也相继被提出。但是，训练一个鉴别指纹真假的模型除依赖超强的算力和专业的领域知识外，还需要海量优质指纹数据的加持，并且一旦模型滥用势必会导致用户隐私的泄露和知识产权侵犯风险，对训练好的深度伪造指纹检测模型进行版权保护迫在眉睫。

深度伪造指纹检测（后文简称为深伪检测）模型在确保指纹识别系统完整和隐私数据安全访问方面的作用是无法替代的[10-12]，尤其是对具有较高隐私的深伪检测模型的保护极为重要。现有的知识产权保护对象更多是新媒体内容，鲜有对深伪检测模型版权保护的研究，并且无法直接将现有方法用于深伪检测模型的版权保护任务中。文献[13]提出一种构造零比特水印的版权保护方法，当模型所有者对知识产权和经济利益产生纠纷时，可调用远程应用程序接口（API, application programming interface）来获得模型的访问权限，并通过远程操作从神经网络模型中提取嵌入的水印实现版权验证，通过对抗训练操作生成触发集以调整分类决策边界，并依据触发集输出的特定标签对模型版权归属进行验证。该方法在 MNIST 数据集上表现出较好的性能，而对指纹数据集保护的效果如何有待考究。文献[14]提出一种抗伪造攻击的神经网络水印协议，通过引入单向哈希函数，确保所有权的触发样本形成单向链，且触发样本的标签也被赋值，其认为攻击者无法拥有训练权限，因此该协议能够抵挡伪造攻击。但是现实中，攻击者通过非法手段能够获得模型的训练权限，该方法将不适用。文献[15]提出一种深度模型的分发机制，能够为用户提供分等级的服务，但是当用户与攻击者发动合谋攻击后，该模型将会被攻击者窃取和非法使用。

针对神经网络模型知识产权侵权问题，本文提出了一种基于差分隐私的深度伪造指纹检测模型版权保护算法。首先，通过构建的触发集微调模型的分类决策边界以建立后门，实现模型版权的被动验证。然后，为了最小化原始任务在非触发集中的误差，在深伪检测模型中设计一个噪声层模块，充分利用差分隐私算法的期望稳定性进行分类决策，让模型在训练时降低对噪声的敏感度。当模型授权给用户后，攻击者与用户发动合谋攻击，以非法获得使用权，此时仅需通过给模型嵌入的后门来验证模型版权。即使攻击者伪造一批与触发集样本同分布的数据来混淆模型版权，所有者依然可通过给触发集加盖时间戳的方式，抵抗混淆版权的恶意攻击。本文的主要贡献如下。

1) 提出一种主动保护和被动验证相结合的深度伪造指纹检测模型版权保护框架。主动保护通过设计一组访问权限，利用概率选择策略将冻结的关键性神经元进行不同程度的解冻，以实现对该模型的授权分发和用户的身份管理。即使攻击者与用户发动合谋攻击，所有者依然可以通过后门映射关系来进行版权的被动验证。

2) 改进了传统的决策边界微调算法。通过给深度伪造指纹检测模型引入随机性，借助差分隐私算法的期望稳定性进行分类决策，以降低模型对噪声的敏感度，从而让模型的分类决策边界更加稳定。确保在后门嵌入时，决策边界不会发生大幅变化而影响原始任务。

3) 在 3 个公开的指纹数据集上进行了性能测试，实验结果表明，主动保护并不会影响后门验证，对于不同模型任务后门依然有效，嵌入的后门对模型修改同样具有稳健性。此外，所提算法能够抵挡攻击者发起的合谋攻击，也能够抵挡模型修改带来的微调、压缩等常见攻击。

1 模型版权保护算法分类

通过对现有模型版权算法进行归纳发现[16]，主要分为三类，即白盒水印算法、黑盒水印算法和无

盒水印算法。白盒水印算法实现流程如图 1 所示，利用白盒水印进行版权验证时，所有者能够访问模型的结构和参数，通过修改神经网络模型的权值实现水印的嵌入和提取。黑盒水印算法实现流程如图 2 所示，在无法获悉神经网络模型的结构和参数时，通过生成的触发集让模型输出预期的分类结果，以实现水印的提取和对比。无盒水印算法实现流程如图 3 所示，利用生成式模型让输出的图像中含有水印。在版权验证时，利用提取网络完成水印的提取和版权归属验证。

1.1　白盒水印算法

Uchida 等[11]在 2017 年首次提出模型水印的概

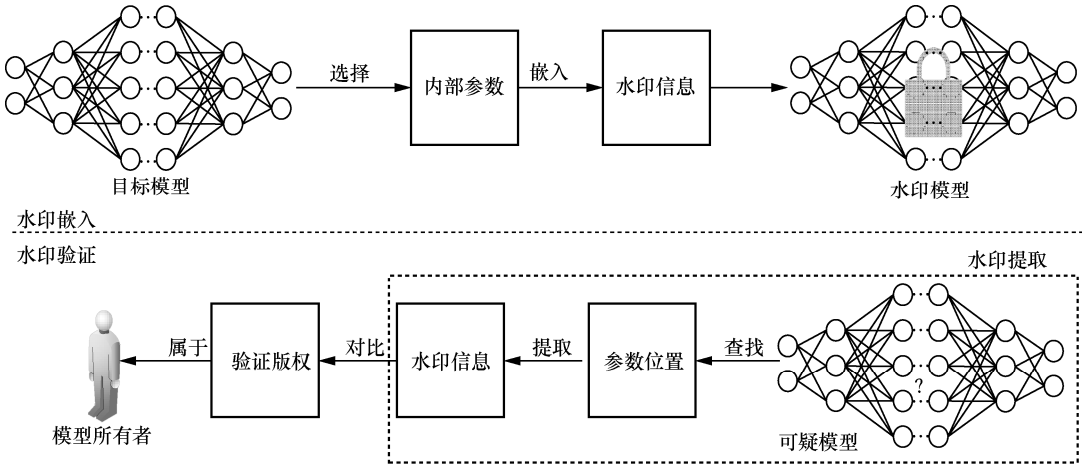


图 1　白盒水印算法实现流程

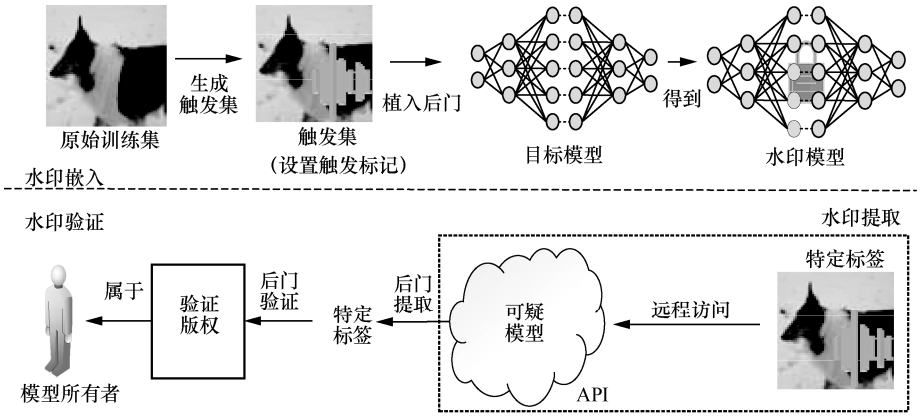


图 2　黑盒水印算法实现流程

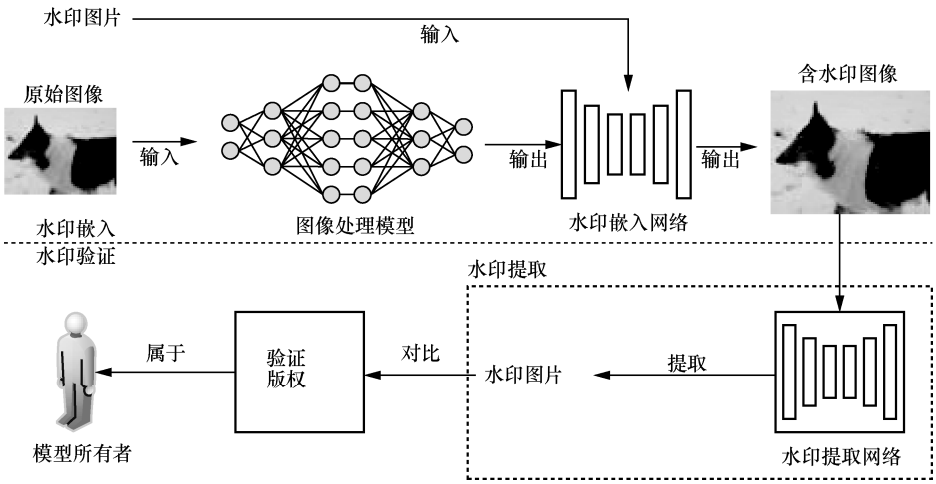


图 3　无盒水印算法实现流程

念。在训练过程中，利用构造的投影矩阵将水印植入模型权重中，通过提取网络实现水印信息的提取和版权验证。具体实现过程如下，首先，随机选择某一卷积层进行平均操作，并将其转化成一维向量；然后，将投影矩阵与一维向量的乘积输入激活函数中，构造一个二值比特流；最后，将水印转化成二进制向量，利用交叉熵损失函数最小化水印信息和激活后的二值比特流，以实现模型水印信息的植入。水印验证是将投影矩阵与权重执行乘法操作，利用阶跃函数进行水印提取，通过与嵌入的二值水印信息进行对比实现模型版权的归属确权。若权重变化幅度较大，便能检测到水印的存在。因此，Kuribayashi 等[17]提出了一种基于全连接层权重的量化水印嵌入方法，该方法通过在训练过程改变模型参数，量化水印对模型的影响，从而确保植入的水印引起的参数变化很小。

Rouhani 等[18]提出一种通用水印版权保护方案，将其命名为 DeepSigns。DeepSigns 能够生成一个受保护的神经网络模型，模型所有者设计一个水印签名，将其植入不同激活图的概率密度函数中，并使用密钥记录嵌入位置。当版权验证时，首先通过密钥获取水印的位置信息；然后提取水印签名；最后比较所提取水印与真实签名之间的误差，若小于阈值，则提取成功。Feng 等[19]提出一种带有补偿机制的水印嵌入方案，为了让嵌入的水印位置更隐蔽，抵抗覆写攻击，选取随机权重；然后将权重进行正交变换，并通过二值化操作向系数中嵌入水印，再通过逆正交变换得到新的含水印的权重；最后使用其他权重作为补偿来微调模型，以消除二值化对性能的影响。

Fan 等[20]指出，现有的水印算法易受到伪造攻击。为此，他们提出一种基于护照的水印策略，即让预先训练好的模型在正确的护照下保持任务性能。当面对伪造或修改的护照，原始任务性能会大幅下降。该方法在不同卷积层后添加了一个护照层，类似于归一化层，区别在于护照层的权重和偏置由特殊的护照决定，而归一化层的权重和偏置是为了保证中间层的变化幅度不能过大而抵消部分归一化操作对模型的影响。因为模型训练过程与护照紧密耦合在一起，所以模型的性能受到护照的控制。若攻击者通过逆向工程伪造一个新的护照来盗取模型，则必须从头训练模型。因此，该方法能够有效抵挡混淆攻击。文献[20]仅适用于一些特殊的归一化层，为了让归一化层都植入水印，Zhang 等[21]在原始任务中引入了一个护照感知分支，通过设计一个秘密护照让护照感知分支与原始模型联合训练。仅当验证模型版权的归属时才提供护照和护照感知分支，而其他时候只将原始模型提供给用户使用。在验证过程中，正确的护照能使模型正常工作，伪造护照则不能。

针对模型版权归属确权问题，目前的白盒水印算法虽然能够很好地解决，但是模型内部结构信息需要公开，攻击者能够轻易地训练一个模型。因此，模型所有者更期望将持有的模型封装成黑盒，通过提供 API 完成指定任务。

1.2 黑盒水印算法

Zhang 等[22]提出一种新颖的模型版权认证方法，即黑盒水印算法，并分别给出 3 种黑盒水印算法：第一种算法将特定的文本信息嵌入图像中作为水印；第二种算法将无意义的噪声嵌入图像作为水印；第三种算法将不相关图像分配错误标签后作为水印。上述方法均能通过植入后门映射关系实现模型的版权归属认证。Adi 等[23]通过后门植入法研究版权的归属问题。首先，从原始数据中选定部分样本作为触发集，并进行标记；然后，通过训练让模型拟合触发样本的特性。在版权验证时，所有者将触发样本输入 API 中，通过观察预测结果是否为预设的标签。为了降低误报率，Guo 等[24-25]提出一种基于进化算法的水印生成和黑盒水印优化算法，将版权所有者的签名植入数据集中，使用含签名信息的数据集来训练一个神经网络模型。当嵌入签名的数据被输入时，预设的临时模式将会运行，以此验证模型的版权。Jia 等[26]发现现有的水印嵌入方法大多与主任务无关，可通过模型微调和压缩来盗取版权，为此提出纠缠水印的概念，即将水印嵌入和原始任务紧密耦合。此外，在后门植入时，触发集输出错误的标签会导致决策边界发生变化，影响原始任务的性能。Zhong 等[27]设计一种全新的黑盒水印算法，在后门嵌入过程中，决策边界并不会发生变化。Quan 等[28]设计一种用于保护图像处理模型的黑盒触发式水印，通过微调操作使模型改变特定域内的预测结果，为了让微调后的模型输出图像和事先预定义的图像接近，将触发图像和初始验证图像一并输入模型中训练，用触发图像的预测结果来更新验证图像。在验证水印时，当所有者把触发图像输入模型后，如果输出结果与验证图像相同则验证成功。Ong 等[29]提出一种用于保护生成对抗网络的水印方法，核心是当输入一个触发图像时，模型

会生成一个包含水印的图像来验证版权。

由于深度模型易遭受数据中毒和后门攻击的威胁，确保神经网络模型在部署后的完整性对黑盒模型极其重要。Zhu 等[30]提出一种基于黑盒的脆弱水印方法来检测恶意微调，水印处理分为以下 3 个步骤：用户首先用一个特定的密钥来构造一组触发集；然后，用交替训练的方式对训练集和触发集进行分类；最后，对训练好的 DNN 模型进行微调。

黑盒水印算法是目前主流的模型版权保护方法。所有者仅需通过 API 访问远程模型便能完成版权验证，不需要像白盒水印算法那样将内部结构公开给第三方。虽然黑盒水印算法提升了模型的安全，但是攻击者依然可通过伪造触发集的方式混淆版权。

1.3　无盒水印算法

无盒水印算法是一种新颖的、不需要人模交互、不需要获悉模型细节和不需要构建特定触发集的生成式版权保护方法，核心操作是在模型训练损失函数中引入一个水印损失项，使输出样本中包含水印信息，最终通过水印提取和比对实现模型版权的归属确权。Zhang 等[31]通过在模型后引入一个与原始任务无关的水印模块，提出一种端到端的水印信息嵌入算法。具体地，在原始任务后设计一个水印嵌入模块，通过迭代优化将水印信息嵌入图像中。为了从水印图像中提取水印，便于后续水印比对和版权归属认证，同时训练一个由密钥控制的水印提取子网。当攻击者利用该框架训练的模型进行代理模型攻击时，表征归属的水印信息将被嵌入该代理模型中。此外，还通过对抗训练提升模型的稳健性，以提高水印防御代理模型攻击的性能。Wu 等[32]提出一种全新数字水印框架，设计一个水印损失组合损失函数来训练模型，使输出的图像中包含一个定制化的水印信息，后续神经网络模型版权需要归属确权时，只要通过检测输出图像中的水印，便能够判断图像是否来自该神经网络模型。实验结果显示，该方法在面对各种图像处理操作时，如图像着色、超分、编辑及语义分割等，均表现出良好的稳健性。

无盒水印的版权验证算法是通过在输出图像中植入定制化数字水印信息来保护模型版权的，为深伪检测模型的版权保护研究提供了新思路。

2　本文算法

2.1　融合主动保护和被动验证的版权保护框架

本文提出的融合主动保护和被动验证的深伪检测模型版权保护框架如图 4 所示。其包含以下 3 个功能：第一，可以防止未授权用户使用深伪检测模型，仅授权用户能使用该模型，而未授权者将会得到一个与任务无关的输出；第二，能够对深伪检测模型进行分级保护，越忠诚的用户获得的访问等级越高；第三，当攻击者策反授权用户发起合谋攻击时，模型所有者可通过后门映射关系对该模型进行版权归属确权。

首先，模型所有者将预训练好的深伪检测模型使用后门嵌入模块来微调决策边界，让模型获得后门映射关系。然后，通过概率选择模块筛选模型中的关键性神经元，以确保选定的神经元不影响后门触发。接着，冻结模块将筛选出的关键性神经元进行冻结，以降低模型的可用性，让未授权用户无法使用该任务模型。最后，分发模块对冻结模块中的神经元授予不同等级，依据授权等级从冻结模块中解冻不同数量的神经元，以执行相应的功能，该操

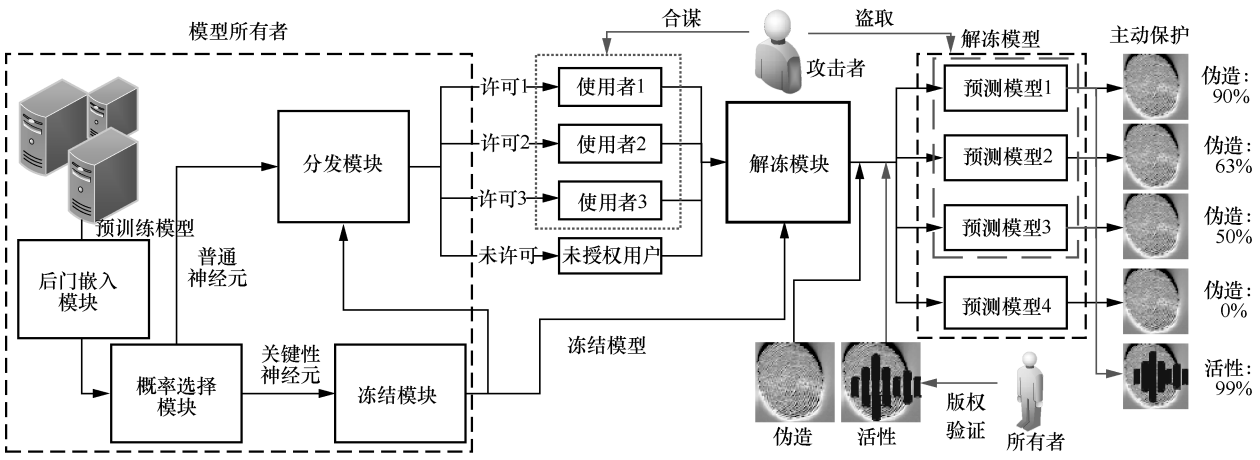
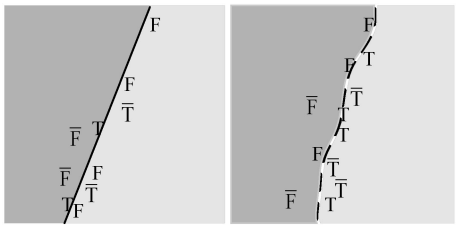


图 4　融合主动保护和被动验证的深伪检测模型版权保护框架

作能够确保最忠诚的用户解冻全部的神经元。在上述操作过程中，模型所有者仅需执行一次后门嵌入模块、概率选择模块和冻结模块的调用操作，当用户需要使用模型时可以多次调用分发模块。被动验证是当用户被攻击者策反后，导致模型被非法使用时所采取的事后验证操作。综上，本文所提的版权保护框架能够为深度伪造指纹检测模型提供一个系统完整和版权归属确权的解决方案。

2.2 决策边界的构建

本文使用 FGSM（fast gradient sign method）[33]生成对抗性指纹集，并在原始任务模型上进行白盒测试，当模型输出标签发生错误时，则定义为成功对手。由于 FGSM 通过逐批添加扰动的方式来生成对抗性指纹，因此那些测试错误的对抗指纹被视为失败对手。选择失败对手作为触发集的目的是限制决策边界[13]，让成功对手返回原先正确分类的类别时变化更少。此外，失败对手还具有表征模型边界形状的作用，从而提升模型的稳健性。决策边界微调的前后对比如图 5 所示，标签 T 和标签 F 分别表示类别为 True 和 Fake 的成功对手，而 $\overline{\mathrm{T}}$ 和 $\overline{\mathrm{F}}$ 分别表示类别为 True 和 Fake 的失败对手。



(a) 原始任务的决策边界 (b) 微调后的决策边界

图 5 决策边界微调的前后对比

为了实现深伪检测模型的版权保护，使用对抗性指纹的原始标签作为后门映射关系进行被动验证。具体地，利用决策边界微调算法让深度伪造指纹检测模型的决策边界发生轻微改变，以实现模型的水印植入。由于触发集是由对抗性指纹构成的集合，当进行决策边界微调时，深伪检测模型能够学习到触发集图像中对抗指纹的特殊特征，使原始任务输出错误结果。为了解决上述问题，受差分隐私和对抗样本防御具有一定联系[34]的启发，让深伪检测模型的决策边界更加稳定，不易受触发集的对抗性扰动影响，本文通过在原始任务中添加噪声层，从而在前向传播过程中引入随机性，使模型能够利用差分隐私算法的期望稳定性进行最终的决策，并能够有效降低深伪检测模型对噪声的敏感性。此外，还能够确保在用触发集进行后门嵌入的过程中，决策边界不会因为扰动对模型的影响而大幅变化。

2.3 噪声层的构建

差分隐私[35]是避免个人数据隐私泄露的一种防御方法，通过给数据添加噪声以引入随机性，使在数据集上的任何增加或删除操作记录都能被隐藏，用户无法通过查询的结果反推出隐私信息。设 D 为随机算法，输入域为 X，输出域为 R，任意 2 个相邻数据集 x，$x'\in X$，输出集合满足 $S\subseteq R$。若 x 和 x' 在算法 D 下满足式(1)，则称算法 D 满足 (ε,δ)-差分隐私。

$$P\left(D(x)\in S\right)\leqslant \mathrm{e}^{\varepsilon}P\left(D(x')\in S\right)+\delta \tag{1}$$

其中，ε 和 δ 均为控制差分隐私保护强度的超参数，ε 为隐私预算，δ 为隐私被泄露的概率，$P(\cdot)$ 为算法 D 的输出概率。

度量标准 ρ 用来表示敏感性，以记录 2 个查询之间的不同数目。在标准的差分隐私中，通常用汉明距离作为度量标准，以使数据库中单一数据的改变不会大幅修改输出的分布。而差分隐私也适用于对抗样本的范数度量。本文将深伪检测模型的输入图像构成的样本视为数据库，将图像中的像素视为记录，以建立起差分隐私与深度伪造指纹检测模型之间的联系。具体地，本文触发集的构建是使用 FGSM 来生成对抗性指纹的，通过微调建立后门映射，利用映射关系来验证版权归属。

本文利用了差分隐私的 2 个属性：1) 后处理性，即差分隐私算法之后模型的输出结果仍具有差分隐私的特性；2) 期望稳定性，即差分隐私算法之后模型的输出期望对输入的扰动变化不敏感。上述属性能够使差分隐私与模型的稳健性建立明确的联系。通过差分隐私的期望稳定性来进行决策，以降低分类决策边界的敏感性。在执行后门嵌入时，微小扰动对原始任务性能并不会产生太大影响。差分隐私的期望稳定性为

$$\mathrm{E}(D(x))\leqslant \mathrm{e}^{\varepsilon}\mathrm{E}(D(x+\alpha))+\delta \tag{2}$$

其中，α 表示添加的扰动，$D(x)$ 表示随机算法的输出，且 $D(x)\in[0,1]$。为了验证差分隐私的属性 2)，对其进行如下推理。连续性随机变量的输出期望为

$$\mathrm{E}(D(x))=\int_0^1 P(D(x)>t)\mathrm{d}t$$

将式(1)两边同时积分得

$$\mathrm{E}(D(x))\leqslant \mathrm{e}^{\varepsilon}\int_0^1 P(D(x+\alpha)>t)\mathrm{d}t+\int_0^1\delta\mathrm{d}t=$$

$$\mathrm{e}^{\varepsilon}\mathrm{E}(D(x+\alpha))+\int_0^1\delta\mathrm{d}t$$

其中，δ 是常数，可得

$$\int_0^1 \delta \mathrm{d}t = \delta$$

因此，$\mathrm{E}(D(x)) \leqslant \mathrm{e}^{\varepsilon}\mathrm{E}(D(x+\alpha)) + \delta$ 得证。

深度伪造指纹检测模型的稳健性是指模型输入的轻微改变并不会影响原始任务的性能，在基于标签输出概率的深度伪造指纹检测模型中，稳健性应该满足

$$y_f(x+\alpha) > \max_{n:n\neq f} y_n(x+\alpha) \tag{3}$$

其中，y 为模型分类结果，f 和 n 为标签类别。

本文将检测模型 SoftMax 层的决策转化为随机的 $D(x)$，利用噪声的输出期望 $\mathrm{E}(D(x))$ 作为决策概率，以挑选最大的概率标签，如式(4)所示。

$$\mathrm{E}(D_f(x)) > \mathrm{e}^{2\varepsilon} \max_{n:n\neq f} \mathrm{E}(D_n(x)) + (1+\mathrm{e}^{\varepsilon})\delta \tag{4}$$

式(4)为稳健性条件，若满足条件，则输出期望 $\mathrm{E}(D(x))$ 对微小扰动是稳健的。证明如下。

证明　根据式(2)可得

$$\mathrm{E}\left(D_f(x)\right) \leqslant \mathrm{e}^{\varepsilon}\mathrm{E}\left(D_f(x+\alpha)\right) + \delta \tag{5}$$

$$\mathrm{E}\left(D_n(x+\alpha)\right) \leqslant \mathrm{e}^{\varepsilon}\mathrm{E}\left(D_n(x)\right) + \delta, \quad n \neq f \tag{6}$$

式(5)给出了 $\mathrm{E}\left(D_n(x+\alpha)\right)$ 的下界，式(6)给出了 $\max_{n\neq f}\mathrm{E}\left(D_f(x+\alpha)\right)$ 的上界。式(4)中标签 n 的期望值下限严格高于其他标签的期望值上限。满足式(3)的稳健性条件，从而建立差分隐私与模型稳健性联系，实现模型输出的稳健性。当满足式(4)时，可得稳健性为

$$\mathrm{E}\left(D_f(x+\alpha)\right) \overset{式(5)}{\geqslant} \frac{\mathrm{E}\left(D_f(x)\right) - \delta}{\mathrm{e}^{\varepsilon}} \overset{式(4)}{\geqslant} \frac{\mathrm{e}^{2\varepsilon}\max\limits_{n:n\neq f}\mathrm{E}\left(D_n(x)\right) + (1+\mathrm{e}^{\varepsilon})\delta - \delta}{\mathrm{e}^{\varepsilon}} = \mathrm{e}^{\varepsilon}\max_{n:n\neq f}\mathrm{E}\left(D_n(x)\right) + \delta \overset{式(6)}{\geqslant} \max_{n:n\neq f}\mathrm{E}\left(D_n(x+\alpha)\right)$$

则深度伪造指纹检测模型稳健性结论为

$$\mathrm{E}\left(D_f(x+\alpha)\right) > \max_{n:f\neq n}\mathrm{E}\left(D_n(x+\alpha)\right) \tag{7}$$

深度伪造指纹检测模型的实现流程如图 6 所示。在进行模型训练时，通过添加噪声层以引入高斯噪声，使深度伪造指纹检测模型的分类决策获得随机性。添加噪声后的训练相对复杂，无法直接计算该输出的期望。因此，本文采用蒙特卡罗估计来近似原有的期望值。具体地，在原始任务中添加决策层，反复调用预测来计算噪声对 SoftMax 层输出结果并取平均操作得到期望的估计值。利用差分隐私算法的期望稳定性进行最终决策以降低该模型对噪声的敏感度。如式(7)所示，对于添加噪声后的指纹图像，该模型的输出期望依然比较稳定。

2.4　主动保护框架

对于训练好的深度伪造指纹检测模型，训练的参数中一部分神经元对任务的决策至关重要，若剔除则会影响任务的输出，被视为关键性神经元；而有些神经元有无与否对任务并无影响，被视为普通神经元。本文提出的主动保护框架通过冻结模型中关键性神经元，以禁止未授权用户的使用。由于大部分神经元位于卷积层，仅冻结卷积层中的关键性

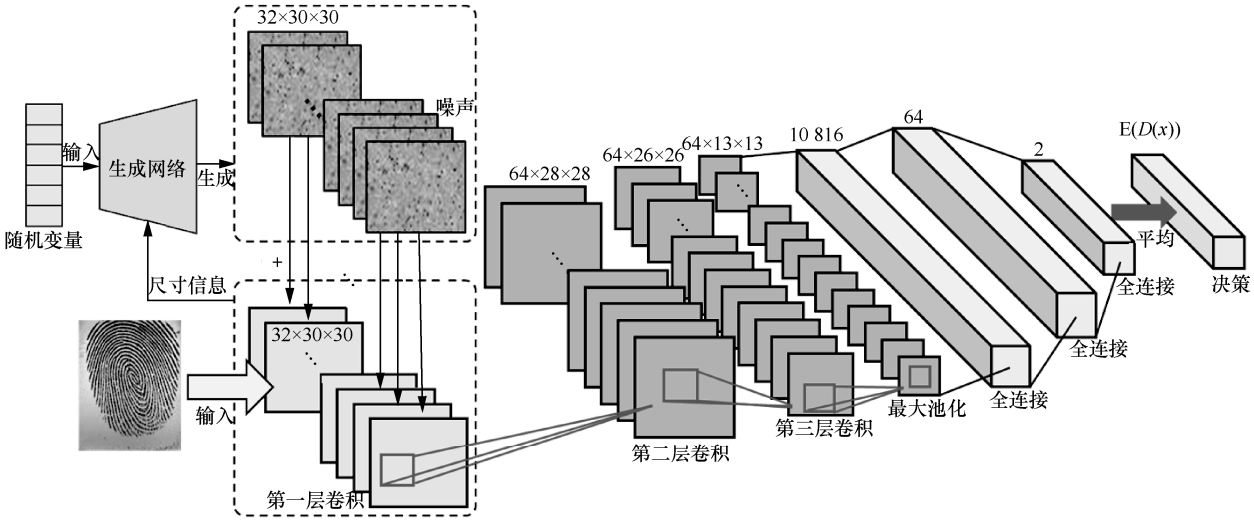


图 6　深度伪造指纹检测模型的实现流程

神经元。首先，通过概率选择策略筛选出关键性神经元以便于后续的冻结操作。概率选择是通过观察丢弃某个神经元后，对模型性能的变化程度，若明显，则相应的神经元极为重要。另外，输入样本不同，对神经元产生的刺激也不同。假设输入样本为 $x_n(n=1,\cdots,N)$，训练的参数中必然存在一些关键性神经元构成集合 $\hat{\theta}$，剔除 $\hat{\theta}$ 则会使性能陡然下降。由于 $\hat{\theta}$ 会随着 x_n 的变化而变化。因此，每输入一批样本，模型就会产生一批不同的集合 $\hat{\theta}$。为了消除不同输入产生的随机性，当所有样本输入后，统计每个神经元入选集合 $\hat{\theta}$ 的次数，并用 p_θ 表示被选定的概率。若神经元总在 $\hat{\theta}$ 中，则为关键性神经元，并赋值 1。本文将概率选择操作转化为式(8)所示的优化问题，通过优化操作筛选出关键性神经元，在不影响后门验证的同时，还能确保选取的关键性神经元尽可能少。

$$\min_{p_\theta\in[0,1]}\frac{1}{N}\sum_{n=1}^{N}l_1\left(f_\theta\left(x_n,\left(I-B^{(n)}\right)\theta\right),y_n\right)+\lambda_1\left\|B^{(n)}\right\|_0+\lambda_2 l_2\left(f_\theta\left(\left(x_i',\left(I-B^{(n)}\right)\theta\right),y_i'\right)\right) \tag{8}$$

其中，$B^{(n)}=\left\{b_\theta^{(n)}\right\}$，$b_\theta^{(n)}$ 满足伯努利分布，记作 $\mathrm{Bern}\left(b_\theta,p_\theta\right)$，$b_\theta^{(n)}$ 为二元随机变量 b_θ 的一个样本；θ 为模型全部神经元构成的集合；I 为所有元素都为 1 的集合，且元素个数与 θ 相同；λ_1 和 λ_2 为比例因子，用于控制正则化项权重。为了利用概率选择剔除掉关键性神经元，先将 I 与 $B^{(n)}$ 对应位置的元素进行相减，再利用逐元素乘法去除被筛选出的元素。当 $b_\theta=1$ 时，则将神经元从 θ 中剔除。为了对剔除参数后的模型性能进行评估，设计 $l_1(\cdot)$ 函数，本文使用负向的交叉熵损失函数来约束模型性能的下降幅度。$\left\|B^{(n)}\right\|_0$ 用来统计非零元素的个数，以控制神经元的数量。$l_2(\cdot)$ 为嵌入触发集后的性能评价指标，使用正向的交叉熵损失函数来控制嵌入后门的触发精度，确保选取的关键性神经元不会干扰后门验证。x_i' 为后门图像，y_i' 为后门标签。由于离散型的二元随机变量不便优化，根据文献[36]中的方法，将离散 B_θ 转化成连续型随机变量 $\tilde{B}_\theta$，即

$$\begin{aligned}s_\theta(u)&=\mathrm{Sigmoid}\left(\frac{\left(\log\left(\frac{u}{(1-u)}\right)+\log\left(\frac{p_\theta}{(1-p_\theta)}\right)\right)}{\beta}\right)\\ \tilde{s}_\theta(u)&=s_\theta(u)(\alpha-\gamma)+\gamma\\ \tilde{B}_\theta&=\min\left(1,\max\left(0,\tilde{s}_\theta(u)\right)\right)\end{aligned} \tag{9}$$

其中，β 用来衡量 $\tilde{B}_\theta$ 与 B_θ 的逼近程度，u 符合均匀分布 $U(0,1)$，$\alpha>0$ 和 $\gamma>0$ 为 $\tilde{B}_\theta\in[0,1]$ 的可调整参数。由于 B_θ 在式(7)中已被 $\tilde{B}_\theta$ 放宽，且 $\left\|B^{(n)}\right\|_0$ 中零元素相对较少。采用文献[37]中的累加分布函数，即

$$\mathcal{Q}\left(\tilde{s}_\theta(u)\right)=\mathrm{Sigmoid}\left(\left(\log\left(\frac{\tilde{s}_\theta(u)-\gamma}{\alpha-\tilde{s}_\theta(u)}\right)\right)\beta-\log\left(\frac{p_\theta}{1-p_\theta}\right)\right) \tag{10}$$

$\left\|B^{(n)}\right\|_0$ 可以被放宽到可微分的程度，如式(11)所示。

$$P\left\{\tilde{B}_\theta\neq 0\right\}=1-P\left\{\left(\tilde{s}_\theta(u)\leqslant 0\right)\right\}=1-\mathcal{Q}(0)=\mathrm{Sigmoid}\left(\log\left(\frac{p_\theta}{1-p_\theta}\right)-\beta\log\frac{-\gamma}{\alpha}\right) \tag{11}$$

通过式(9)和式(11)将式(8)放宽，并将式(8)的优化问题转化为

$$\min_{p_\theta\in[0,1]}\frac{1}{N}\sum_{n=1}^{N}\left(l_1\left(f_\theta\left(x_n,\left(\left(I-\tilde{B}^{(n)}\right)\theta\right)\right),y_n\right)\right)+\lambda_1\sum_\theta\mathrm{Sigmoid}\left(\log\left(\frac{p_\theta}{1-p_\theta}\right)-\beta\log\frac{-\gamma}{\alpha}\right)+\lambda_2 l_2\left(f_\theta\left(\left(x_i',\left(I-\tilde{B}^{(n)}\right)\theta\right),y_i'\right)\right) \tag{12}$$

利用式(12)对模型进行优化，以完成关键性神经元的筛选和保障后门的验证；利用概率选择操作将关键性神经元进行冻结，并限制未授权用户的使用；将冻结的神经元划分不同子集，让每个子集均包含不同数量的神经元，图 4 中的分发模块通过选择不同的子集来控制深度伪造指纹检测模型的性能，而冻结模块和解冻模块分别用来进行关键性神经元的冻结和解冻操作。

2.5 时间戳

任意数据经过哈希运算[38]后均生成一个定长的输出。该运算是单向的，即无法依据哈希值反推出原始数据。因此，使用时间戳对电子数据产生的时间进行签名认证，以证明其在某个伪造签名之前就已存在。时间戳的生成操作如下。

1) 使用时间戳服务中心（TSSC, time stamp service center）提供的时间戳软件，将电子数据加盖时间戳并输出哈希值 A。

2) 将生成的哈希值 A 发送给 TSSC，由 TSSC 记录下生成的时间点，并将哈希值 A 与时间点拼接

组成的新数据输入哈希函数，构建新的哈希值 B。

3) TSSC 利用私钥将哈希值 B 进行加密操作以防止 B 的泄露，将加密后的哈希值与时间点绑定封装来生成时间戳，并返还给申请者保管。

时间戳的验证步骤如下。

1) 将原有电子数据作为输入，使用时间戳软件求得哈希值 A。

2) 把哈希值 A 与时间点作为输入，得到哈希值 B。

3) 利用 TSSC 提供的公钥将使用者保管的加密内容进行解密，得到哈希值 B'。

4) 通过对比哈希值 B 和哈希值 B'，来判断原有数据的时间点是否一致。

3　具体实施

3.1　数据集介绍

为了验证本文所提算法的性能，使用的数据集分别来自 2015 年、2017 年和 2019 年的指纹活性检测竞赛，公开发布 3 个指纹集 LivDet2015、LivDet2017 和 LivDet2019。其中，LivDet2015 指纹数据集中的图像使用 4 种不同的光学传感器 GreenBit、Biometrika、DigitalPersona 和 Crossmatch 采集构建而成，每类传感器采集的指纹数量约为 4 000 张。LivDet2017、LivDet2019 数据集中的图像则是由 GreenBit、DigitalPersona、Orcanthus 这 3 种不同光学设备所采集，LivDet2017 和 LivDet2019 中每类光学传感器采集的指纹约为 6 000 张和 4 000 张。

3.2　性能评价指标

本文采用的深度伪造指纹检测模型版权保护算法的性能指标如下[13]。

1) 保真度。在神经网络模型中植入水印后，不能影响原始检测模型的性能。

2) 高效性。植入的神经网络水印应避免模型版权验证时响应时间过长。

3) 有效性。神经网络水印必须长期有效，对每个用户保持独一无二性。

4) 稳健性。神经网络水印遭受常见的恶意攻击后，依然存在且能用于后续的模型版权归属确权。

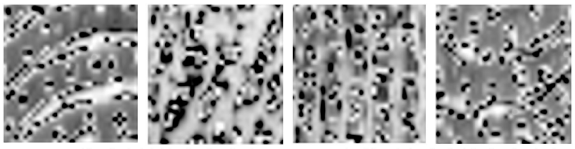
5) 安全性。用于模型版权验证的水印不易被伪造、访问和读取。

3.3　后门的嵌入和提取

黑盒水印主要是通过构造触发集来为模型嵌入后门的。当模型版权归属发生纠纷时，拥有者通过触发集的特定输出实现模型版权认定，而伪造者无法提供该证明。触发集的生成如式(13)所示。

$$x' = x + \varepsilon \operatorname{sign}\left(\nabla_x J(\theta, x, y)\right) \tag{13}$$

其中，θ 为检测模型的相关参数，$J(\theta, x, y)$ 为训练该模型的损失函数，∇ 为梯度，sign(·) 为符号函数，ε 为添加的扰动大小。通过逐渐添加扰动的方式使构造的触发集位于决策边界附近。成功对手和失败对手的对抗性指纹示例如图 7 所示，当触发集中样本数量为 4 时，图 7(a)为越过边界的成功对手，图 7(b)为保持原有分类的失败对手。



(a) 越过边界的成功对手　　(b) 保持原有分类的失败对手

图 7　成功对手和失败对手的对抗性指纹示例

后门的提取通过输入触发集样本使成功对手和失败对手都能被深度伪造指纹检测模型正确分类，最理想的状态是所有触发样本的标签与模型预测结果之间的距离为零。但是，由于深度伪造指纹检测模型存在被攻击和嵌入时触发精度对原始任务的影响，导致误报，因此，通过设计阈值 θ 对水印提取的性能进行评估。为了将错误率控制在 0.05 以内，嵌入成功和失败的概率均为 0.5，触发样本服从二项分布 $B\left(|T|, \frac{1}{2}\right)$，即

$$2^{-|T|} \sum_{z=0}^{\theta} \binom{|T|}{z} < 0.05 \tag{14}$$

其中，T 为触发集样本的个数，为了使水印验证有效，只需误报数小于阈值，便认为成功提取水印。如当 T=50 时，阈值为 19，只要触发集的标签与预测标签最大误差小于 19，则表明水印提取成功。

3.4　实验结果

本文在不同数据集下进行了算法性能测试，在 LivDet2017 中的 Orcanthus 对不同卷积层模型分类精度进行了分析，不同卷积层数下的分类精度如图 8 所示。卷积层数为 3 时，深伪检测任务的性能最佳，因此在后续的实验中，均采用 4 个卷积层和 3 个全连接层结构，利用期望值进行最终决策，且差分隐私算法的强度分别设为 ε=1.0，δ=0.05。目前，基于深层的伪造指纹检测虽然已取得极高的性能，但是本文还是使用一个轻量级的模型进行版

权保护：一方面，本文主要研究的是深伪检测模型的版权保护任务，有别于传统的深伪检测任务，因而选用的是参数较少、层数较浅的模型；另一方面，鉴于移动终端的有限算力，本文期望设计的轻量化深伪指纹检测模型能够应用在小型化的设备提供服务，如手机、平板等移动终端。

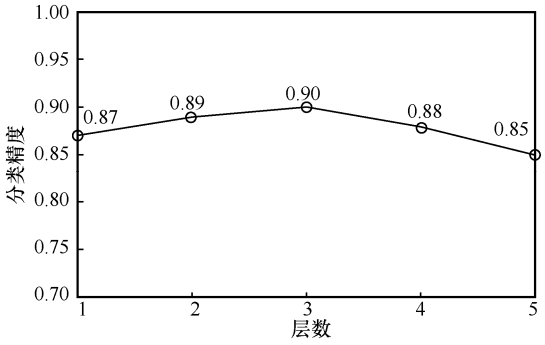


图 8 不同卷积层数下的分类精度

本文将 LivDet2015 中的真假指纹图像进行再整合，用于真伪检测模型的构建，差分隐私对深度伪造指纹检测模型分类精度的保护效果如图 9(a)所示，原始任务检测模型分类精度为 92%。当使用触发集微调决策边界时，原始任务的分类精度下降了 22%。虽然能够鉴别模型版权归属，但此时的性能下降较明显，不满足版权保护算法中的保真度。通过引入噪声层的方式重新构造决策边界，在完成版权归属确权同时，能够减弱对任务性能的影响，仅下降 3%。此外，本文还测试了 LivDet2017 和 LivDet2019 中在不同传感器下的保护效果，如图 9(b)所示。结果表明嵌入的后门不仅能够被成功触发，并且优化后的决策边界微调算法还能对原始任务起到较好的保护作用。为了进一步验证所提框架的有效性，本文测试了不同用户等级下的模型分类精度，如图 10 所示，结果表明所提算法依旧有效。

若未授权用户尝试访问深度伪造指纹模型，将拒绝提供服务，即使提供输入数据，输出的结果也将无任何参考价值。真伪指纹鉴别是一个二分类问题，性能最低为 50%，相当于随机采样的概率。为了验证本文采用的概率选择策略能够降低模型分类精度，采用了以下 4 种不同的策略对神经元进行冻结。1) 随机，随机性地冻结神经元。2) 均值，围绕总体神经元的均值冻结。3) 升序，按照神经元的值从小到大冻结。4) 降序，按照神经元的值从大到小冻结。将第 2 个卷积层中的神经元进行不同程度的冻结，模型的性能如图 11 所示，可观察到本文采用的概率选择策略仅需冻结 4%左右的神经元就能快速降低模型的分类精度，而其他 4 种策略则需要冻结 20%左右的神经元。

3.5 模型稳健性验证

为了验证深度伪造指纹检测模型修改后是否具有稳健性，本文还在 LivDet2015 数据集下进行了稳健性实验。通过对参数进行修剪，将绝对值最小的权重剔除来模拟压缩攻击，结果表明构造的触发后门能够抵挡模型压缩攻击。对模型压缩的稳健性如表 1 所示，深度伪造指纹检测模型能抵挡 50%左右的压缩攻击。

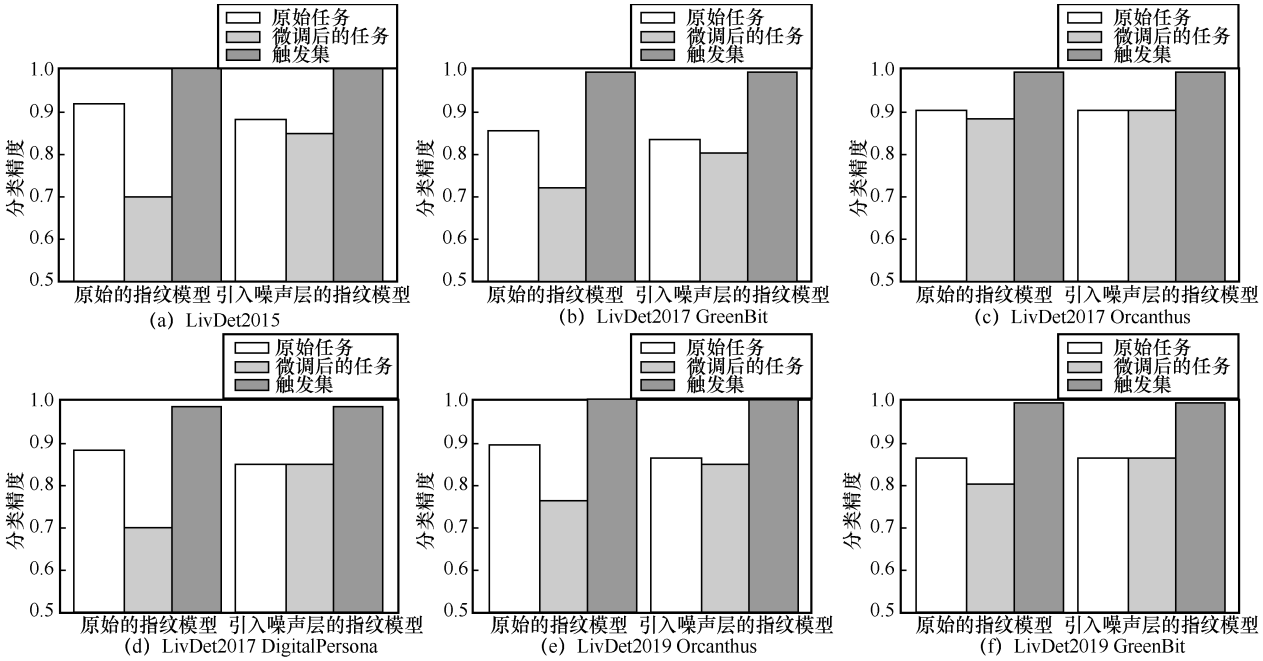


图 9 差分隐私对深度伪造指纹检测模型分类精度的保护效果

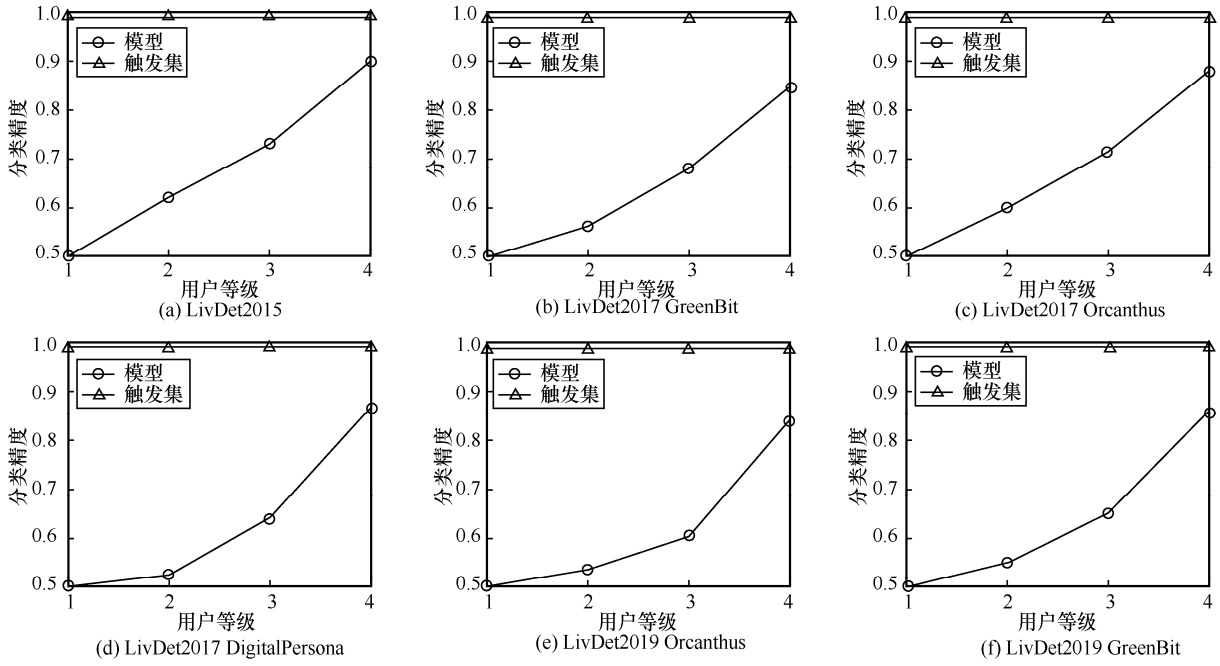


图 10　不同用户等级下的模型分类精度

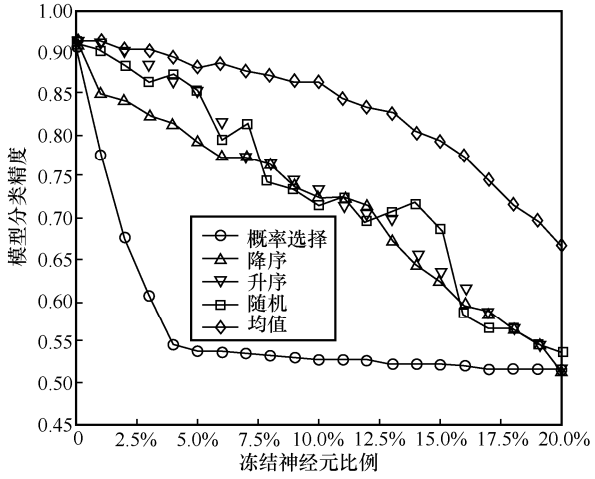


图 11　不同冻结神经元比例下的模型分类精度

表 1　　**对模型压缩的稳健性**

序号	模型压缩率	水印提取	模型分类精度
1	0.00	成功	90%
2	0.25	成功	84%
3	0.50	成功	70%
4	0.75	失败	50%
5	0.85	失败	50%
6	1.00	失败	0

通过构造大小不同的伪造触发集来微调训练好的深度伪造检测模型，对模型微调的稳健性如表 2 所示。原有的触发集能够对检测模型进行版权归属认证，表明本文提出的触发后门具有稳健性，与对抗样本难以防御的特性[39]相一致。

表 2　　**对模型微调的稳健性**

序号	触发集大小/个	水印提取	模型分类精度
1	20	成功	72%
2	50	成功	68%
3	100	成功	57%
4	120	成功	50%

除了上述 2 种攻击，攻击者还可能对冻结的关键性神经元进行再训练，尝试重现原始任务。由于攻击者事先无法获悉检测模型是在哪层执行关键性神经元冻结，只能通过固定其他层的神经元对该层进行重训练。对于未授权用户来讲，该尝试等同于重新训练一个原始任务模型。而攻击者是因为计算资源有限，为了降低模型的训练成本，才会盗取合法用户的知识产权。因此，攻击者不会花费更多训练代价来获取模型的使用权，使主动保护方案具有稳健性。即使使用者被策反，模型拥有者依然可借助时间戳进行模型版权归属确权，本文所提的框架依然具有稳健性。

3.6　水印验证框架

深度伪造指纹检测模型在抵抗模型微调攻击的同时，也会存在多个水印共存问题，间接发生水印的混淆。攻击者对模型进行微调攻击后，尽管模型性能会下降，但是攻击者宁愿牺牲检测模型的准确率。为了保障安全，本文设计了一种新的水印验证框架，当发生水印混淆时，由权威第三方提供时间戳的生成和认证服务，模型所有者仅需向权威第三方提供触发

集，使用 SHA-256 哈希运算来为触发集生成时间戳。当需要版权确权时，再次向权威第三方提供触发集，借助时间戳认证服务，通过时间先后来对混淆后的版权进行认证。模型所有者把争议模型以及触发集提供给权威第三方，权威第三方通过 SHA-256 哈希函数给触发集加盖时间戳，把生成的哈希值交还给模型所有者，形成证据。当发生水印混淆的时候，模型所有者和攻击者都需要把哈希值和触发集提交给权威第三方进行认证。最后权威第三方通过时间戳的哈希值，来判断时间节点的先后顺序，生成时间戳靠前的版权验证者为模型的真正所有者。

3.7 方法的通用性

除了在 3 个公开的指纹数据集上进行了版权归属验证，本文还在 Cifar10 数据集上进行了通用性测试，本文算法同样表现出较好性能，如表 3 所示。通过与现有的 5 种算法[22-23, 25]对比可知，当模型嵌入黑盒水印后，原始任务分类精度都会退化，相比较而言，本文算法分类精度降幅更小，基本上可忽略，对原始任务影响较小。

表 3 算法的通用性性能

算法	触发成功率	分类精度降幅
WMcontent[22]	0.99	0.001 9
WMunrelated[22]	1	0.004 8
WMnoise[22]	0.99	0.001 1
Fromscratch[23]	1	0.003 4
文献[25]算法	0.99	0.002 8
本文算法	0.99	0.000 7

4 结束语

在保密通信过程中，指纹识别是应用最广的身份识别技术，对保障隐私安全和查验用户身份的合法与否至关重要。近年来，研究者发现其易遭受伪造指纹的欺骗攻击，伪造指纹检测技术应运而生。但是训练一个鉴别真假指纹的深层模型需要海量的数据和超强的算力，高敏感型的指纹被收集后存在泄露风险，而深度伪造指纹检测模型的滥用势必会导致个人隐私的泄露和知识产权侵权风险，对深伪检测模型进行版权保护迫在眉睫。针对传统黑盒版权保护算法存在削弱原始任务性能且适用于模型事后确权的问题。本文提出一种基于差分隐私的深度伪造指纹检测模型版权保护算法，在实现版权的主动保护和被动验证的同时，能够兼顾原始任务分类精度。为解决传统的决策边界微调算法造成的原始任务分类精度下降问题，本文在检测模型中引入了噪声层模块，旨在特征提取过程中引入随机性，并利用差分隐私算法的期望稳健性进行最终的决策，以训练一个对噪声不敏感的深度伪造指纹检测模型。通过对抗训练来微调该模型的决策边界为其嵌入后门，使嵌入后门后的决策边界只发生轻微变化。采用概率选择策略对深度伪造指纹检测模型的神经元进行选择性冻结，让忠诚的用户可解冻更多数量的神经元，以实现对该模型的主动保护。此外，还设计了一种水印验证框架，攻击者通过伪造触发集来为模型植入后门水印，致使模型版权发生了混淆。当模型面对混淆攻击时，所有者可通过时间戳的顺序，对该模型版权进行验证。实验结果表明，本文设计的版权保护算法对多种不同攻击具有一定的稳健性。

由于模型版权保护研究还处于起步阶段，尤其是基于生物特征的神经网络模型，目前还没有统一的性能评价指标，如何为不同模型和不同的版权保护方法设计统一的指标是接下来需要研究的内容。

参考文献：

[1] YADAV J, JAFFERY Z A, SINGH L. A short review on machine learning techniques used for fingerprint recognition[J]. Journal of Critical Reviews, 2020, 7(13): 2768-2773

[2] YUAN C S, YU P P, XIA Z H, et al. FLD-SRC: fingerprint liveness detection for AFIS based on spatial ridges continuity[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(4): 817-827.

[3] HE Y, ZHAO N, YIN H X. Integrated networking, caching, and computing for connected vehicles: a deep reinforcement learning approach[J]. IEEE Transactions on Vehicular Technology, 2018, 67(1): 44-55.

[4] ZHAO D B, CHEN Y R, LV L. Deep reinforcement learning with visual attention for vehicle classification[J]. IEEE Transactions on Cognitive and Developmental Systems, 2017, 9(4): 356-367.

[5] LI X L, DING L K, WANG L, et al. FPGA accelerates deep residual learning for image recognition[C]//Proceedings of IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference. Piscataway: IEEE Press, 2017: 837-840.

[6] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv Preprint, arXiv: 1409.1556, 2014.

[7] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.

[8] BHUYAN M P, SARMA S K, RAHMAN M. Natural language processing based stochastic model for the correctness of Assamese sentences[C]//Proceedings of the 5th International Conference on Communication and Electronics Systems (ICCES). Piscataway: IEEE Press, 2020: 1179-1182.

[9] YUAN C S, JIAO S M, SUN X M, et al. MFFFLD: a multimodal-feature-fusion-based fingerprint liveness detection[J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(2): 648-661.

[10] CETINIC E, LIPIC T, GRGIC S. Fine-tuning convolutional neural networks for fine art classification[J]. Expert Systems with Applica-

tions, 2018, 114: 107-118.
[11] 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. New York: ACM Press, 2017: 269-277.
[12] LIU Z, SUN M, ZHOU T, et al. Rethinking the value of network pruning[J]. arXiv Preprint, arXiv: 1810.05270, 2018.
[13] LE M E, PÉREZ P, TRÉDAN G. Adversarial frontier stitching for remote neural network watermarking[J]. Neural Computing and Applications, 2020, 32(13): 9233-9244.
[14] ZHU R, ZHANG X, SHI M, et al. Secure neural network watermarking protocol against forging attack[J]. EURASIP Journal on Image and Video Processing, 2020, 2020(1): 1-12.
[15] TIAN J Y, ZHOU J T, DUAN J. Probabilistic selective encryption of convolutional neural networks for hierarchical services[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 2205-2214.
[16] 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.
FAN X F, ZHOU X Y, ZHU B B, et al. Survey of copyright protection schemes based on DNN model[J]. Journal of Computer Research and Development, 2022, 59(5): 953-977.
[17] KURIBAYASHI M, TANAKA T, FUNABIKI N. Deepwatermark: embedding watermark into DNN model[C]//Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Piscataway: IEEE Press, 2020: 1340-1346.
[18] 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. New York: ACM Press, 2019: 485-497.
[19] FENG L, ZHANG X. Watermarking neural network with compensation mechanism[C]//Proceedings of International Conference on Knowledge Science, Engineering and Management. Berlin: Springer, 2020: 363-375.
[20] FAN L, NG K W, CHAN C S. Rethinking deep neural network ownership verification: embedding passports to defeat ambiguity attacks[C]//Proceedings of Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2019: 4716-4725.
[21] ZHANG J, CHEN D, LIAO J, et al. Passport-aware normalization for deep model protection[J]. Advances in Neural Information Processing Systems, 2020, 33: 22619-22628.
[22] ZHANG J, GU Z, 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. New York: ACM Press, 2018: 159-172.
[23] ADI Y, BAUM C, CISSE M, et al. Turning your weakness into a strength: watermarking deep neural networks by backdooring[J]. arXiv Preprint, arXiv: 1802.04633, 2018.
[24] GUO J, POTKONJAK M. Evolutionary trigger set generation for DNN black-box watermarking[J]. arXiv Preprint, arXiv: 1906.04411, 2019.
[25] GUO J, POTKONJAK M. Watermarking deep neural networks for embedded systems[C]//Proceedings of IEEE/ACM International Conference on Computer-Aided Design. Piscataway: IEEE Press, 2018: 1-8.
[26] JIA H, CHOQUETTE-CHOO C A, CHANDRASEKARAN V, et al. Entangled watermarks as a defense against model extraction[C]//Proceedings of the 30th USENIX Security Symposium. Berkeley: USENIX Association, 2021: 1937-1954.
[27] ZHONG Q, ZHANG L Y, ZHANG J, et al. Protecting IP of deep neural networks with watermarking: a new label helps[C]//Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2020: 462-474.
[28] 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.
[29] ONG D S, SENG C C E, NG K W, et al. Protecting intellectual property of generative adversarial networks from ambiguity attacks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 3629-3638.
[30] ZHU R, WEI P, LI S, et al. Fragile neural network watermarking with trigger image set[C]//Proceedings of International Conference on Knowledge Science, Engineering and Management. Berlin: Springer, 2021: 280-293.
[31] 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.
[32] WU H, LIU G, YAO Y, et al. Watermarking neural networks with watermarked images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(7): 2591-2601.
[33] HUANG S, PAPERNOT N, GOODFELLOW I, et al. Adversarial attacks on neural network policies[J]. arXiv Preprint, arXiv: 1702.02284, 2017.
[34] LECUYER M, ATLIDAKIS V, GEAMBASU R, et al. Certified robustness to adversarial examples with differential privacy[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2019: 656-672.
[35] 刘艺菲，王宁，王志刚，等. 混洗差分隐私下的多维类别数据的收集与分析[J]. 软件学报, 2022, 33(3): 1093-1110.
LIU Y F, WANG N, WANG Z G, et al. Collecting and analyzing multidimensional categorical data under shuffled differential privacy[J]. Journal of Software, 2022, 33(3): 1093-1110.
[36] SHAYER O, LEVI D, FETAYA E. Learning discrete weights using the local reparameterization trick[J]. arXiv Preprint, arXiv: 1710.07739, 2017.
[37] LOUIZOS C, WELLING M, KINGMA D P. Learning sparse neural networks through L0 regularization[J]. arXiv Preprint, arXiv: 1712.01312, 2017.
[38] BOGDANOV A, KNEŽEVIĆ M, LEANDER G, et al. SPONGENT: a lightweight hash function[C]//Proceedings of International Workshop on Cryptographic Hardware and Embedded Systems. Berlin: Springer, 2011: 312-325.
[39] SHAFAHI A, HUANG W R, STUDER C, et al. Are adversarial examples inevitable?[J]. arXiv Preprint, arXiv: 1809.02104, 2018.

[作者简介]

袁程胜（1989– ），男，山东济宁人，南京信息工程大学副教授、硕士生导师，主要研究方向为信息隐藏、多媒体取证与 AI 安全。

郭强（1997– ），男，江苏南京人，南京信息工程大学硕士生，主要研究方向为信息安全和深度学习。

付章杰（1983– ），男，河南南阳人，南京信息工程大学教授、博士生导师，主要研究方向为区块链安全、数字取证、人工智能安全。