DOI:10.7544/issn1000-1239.20211115
59(5):953-977,2022

深度神经网络模型版权保护方案综述

樊雪峰[1] 周晓谊[1] 朱冰冰[1] 董津位[2] 牛 俊[3] 王 鹤[2]
[1](海南大学网络空间安全学院 海口 570228)
[2](西安电子科技大学网络与信息安全学院 西安 710126)
[3](西安电子科技大学计算机科学与技术学院 西安 710071)
(xffan98@163.com)

Survey of Copyright Protection Schemes Based on DNN Model

Fan Xuefeng[1], Zhou Xiaoyi[1], Zhu Bingbing[1], Dong Jinwei[2], Niu Jun[3], and Wang He[2]
[1](*School of Cyberspace Security*, *Hainan University*, *Haikou* 570228)
[2](*School of Cyber Engineering*, *Xidian University*, *Xi'an* 710126)
[3](*School of Computer Science and Technology*, *Xidian University*, *Xi'an* 710071)

Abstract Emerging technologies such as the deep neural network (DNN) have been rapidly developed and applied in industrial Internet security with unprecedented performance. However, training a DNN model needs to capture a large number of proprietary data in different scenarios in the target application, to require extensive computing resources, and to adjust the network topology with the assistance of experts to properly train the parameters. As valuable intellectual property, DNN model should be technically protected from illegal reproduction, redistribution or abuse. Inspired by the classical watermarking technologies which protect intellectual property rights related to multimedia content, neural network watermarking is currently the DNN model copyright protection method most concerned by researchers. So far, there is no complete description of the application of neural network watermarking in the protection of intellectual property of DNN models. We investigate the relevant work of CCF recommended journals and conferences in recent five years. From the perspective of watermark embedding and extraction, based on the original classification of white box and black box watermarking, the neural network watermarking is extended to gray box and null box. The white box and black box watermarkings are summarized in details according to their different ideas and various task models, and the performances of the four classifications are compared. Finally, we discuss the future challenges and research directions of neural network watermarking, aiming to provide guidance to further promote such technologies for DNN model copyright protection.

Key words copyright protection; deep neural network (DNN); white box watermarking; black box watermarking; gray box watermarking; null box watermarking

摘 要 深度神经网络(deep neural network, DNN)等新兴技术以前所未有的性能在工业互联网安全中得到广泛发展和应用.然而,训练DNN模型需要在目标应用程序中捕获大量不同场景的专有数据、广泛

收稿日期:2021-11-16;**修回日期**:2022-01-19
基金项目:海南省高水平人才计划项目(2019RC044);国家重点研发计划项目(2018YFB2100805,2018YFB0804701)
This work was supported by the High-level Talent Program of Hainan Province (2019RC044) and the National Key Research and Development Program of China (2018YFB2100805, 2018YFB0804701).
通信作者:周晓谊(xy.zhou.xy@gmail.com)

的计算资源,以及在专家的协助下调整网络拓扑结构并正确训练参数.因此,DNN 模型应当作为有价值的知识产权,从技术上保护其不被非法复制、重新分发或滥用.受经典水印技术被用于保护与多媒体内容相关的知识产权的启发,神经网络水印是目前最受研究者关注的 DNN 模型版权保护方法.迄今为止,学术界对神经网络水印在 DNN 模型知识产权保护中的应用尚缺乏完整描述.调研了近 5 年 CCF 推荐期刊和会议等关于该领域的相关工作,从水印的嵌入和提取的视角,将神经网络水印在原有的白盒水印和黑盒水印分类的基础上,扩充了灰盒水印和无盒水印 2 种分类,对白盒水印和黑盒水印方法根据其水印嵌入的不同思路和不同任务模型进行了更详细的分类总结,并对 4 类水印方法的性能进行了对比.最后,探讨了神经网络水印未来面临的挑战和可研究的方向,旨在为学者进一步推动基于神经网络水印的 DNN 模型版权保护的发展提供指导.

关键词 版权保护;深度神经网络;白盒水印;黑盒水印;灰盒水印;无盒水印

中图法分类号 TP391

工业互联网(industrial Internet)是新一代信息通信技术与工业经济深度融合的新型基础设施、应用模式和工业生态,通过“人、机、物”的全面互联,全要素、全产业链、全价值链的全面连接,推动制造业生产方式和企业形态变革.安全是工业互联网领域关注的核心问题之一,是实现工业互联网高质量发展的前提条件,具有重要的价值.虽然 DNN 技术在提升工业互联网安全防护水平中发挥了重要作用,但其固有的高维线性、大量依赖数据、可解释性不佳等特性,为工业互联网埋下了安全隐患.目前针对 DNN 模型的攻击层出不穷,如利用构建检测器等方法对所有权验证进行的攻击等[1],对 DNN 模型及其所有者的权益造成了极大的侵害.同时,由于训练 DNN 模型所花费的人力和物力资源巨大,因此,有必要设计 DNN 模型的知识产权保护技术,从而使侵权盗版者无计可施,并保持 DNN 模型所有者在商业竞争中的优势.

1 从多媒体版权保护到 DNN 模型版权保护

数字资产的知识产权保护[2-7]研究工作始于 20 世纪 50 年代,目前已经有 70 年左右的历史.其中,数字水印[8-12]作为最受研究者关注的多媒体版权保护方法,被应用于数字图像、音频、视频等多媒体产品上进行版权保护以及验证多媒体数据的完整性[13].水印是一种基于内容的信息隐藏技术[14-16],最早可追溯到唐朝我国造纸工匠用于防止假冒和美化纸张所发明的水印.大约 7 个世纪以前,意大利的 Fabriano 镇出现了纸水印[17].1954 年,Muzak 公司的工程师将数字水印应用于保护音乐作品的方案中.从那时起,人们开始进行了大量的水印技术研究与应用.多媒体水印是将所有者信息嵌入载体数据中,且不影响载体数据的正常使用.当发生版权纠纷时,版权所有者从多媒体数据中提取出事先嵌入的水印,用以证明所有者版权,因此,这就要求嵌入的水印不易被检测和去除,具有一定的鲁棒性.数字水印的分类如表 1 所示:

Table 1 Digital Watermarking Classification

表 1 数字水印分类

依据	分类
载体形式	图像水印[18-20]、视频水印[21-23]、音频水印[24-26]、文本水印[27-29]、用于三维网格模型的网格水印[30-31]
是否需要原始载体	非盲水印[32-33]、盲水印[34-37]
鲁棒性	鲁棒水印[38-39]、脆弱水印[40-41]、半脆弱水印[42-43]
嵌入位置	空间域[44-46](时域)水印、变换域[47-49](频域)水印

近年来,DNN 在图像和语音处理等领域被广泛应用,如自然语言处理[50-52]、计算机视觉[53-55]、图像识别[56-57]、图像处理[58]、对象检测[59-61]、语音识别[62-63]等.一些先进的神经网络模型如 LeNet[64],AlexNet[56],VGGNet[65],GoogLeNet[57] 和 ResNet[53] 表现出了优异的性能.微软、百度和谷歌等互联网公司已经在其产品服务中布局 DNN 模型,用以提供智能化和高质量的服务,同时 DNN 也逐渐成为目前提升工业互联网安全防护水平的主要技术之一.然而,在享受着智能化服务的同时,如何保护 DNN 模型的版权不被非法盗取和剽窃逐渐成为一个具有挑战性的问题.与传统的多媒体数据不同,训练一个好的 DNN 模型所花费的代价是巨大的,这需要用到大规模的数据集、庞大的计算资源和人力成本[66-67].因此,保护 DNN 模型的版权不受侵犯变得尤为重要.

一种保护 DNN 模型版权比较有效的方法是将数字水印引入其中.但 DNN 模型强大的学习能力导致其与传统多媒体相比,版权保护工作更具有挑战性.不仅如此,DNN 模型不具备多媒体直观的特性,导致传统的数字水印不能直接应用于 DNN 模型的保护.近年来,研究者针对 DNN 模型的版权保护提出了很多优秀的方案.文献[68]从神经网络水印技术的相关基础、白盒和黑盒水印方法的梳理对比、针对水印的攻击方法等方面总结了神经网络水印技术的研究进展,并对未来的工作进行了展望,为该领域的发展提供了极大的参考价值,但是该文献缺乏对图像处理任务 DNN 模型保护工作的梳理,而且对神经网络水印的分类存在不足.文献[69]总结了现阶段神经网络水印的研究成果,论述了当前主流的神经网络水印算法,对基于内部权重水印算法(正则项水印)等 4 类典型算法进行了复现比较,但该文献对当前针对神经网络水印的攻击方法等描述不够完善.文献[70]从场景、机制、容量、类型、功能和目标模型 6 个属性提出了 DNN 模型版权保护方法的分类,并将针对 DNN 模型保护方法的攻击分为 3 个级别,从弱到强分别是模型修改、被动攻击、主动攻击,但该文献对已有工作的梳理不够详细,对现有工作的描述较为简单.文献[71]引入了一种新的神经网络水印分类法,将神经网络水印分类为静态水印算法和动态水印算法,为我们梳理相关工作提供了一种新的思路,同时文章给出了属于每一类的几个示例性方法,但没有梳理迄今为止提出的所有方法.因此,我们在过去的基础上,不仅对图像处理任务 DNN 的保护工作进行了梳理,而且对白盒和黑盒方法进行更详细的分类总结,之后将神经网络水印的分类进行了扩充,对当前神经网络水印的性能指标和攻击方法进行了整理汇总.另外,我们针对未来 DNN 模型的版权保护工作也提出了自己的一些观点.

我们调研了网络安全领域的四大顶级会议 USENIX Security Symposium, Network and Distributed System Security Symposium (NDSS), Conference on Computer and Communications Security (ACM CCS), IEEE Symposium on Security and Privacy (S&P)以及其他期刊和会议近 3~5 年的相关文章,并在 5 个领域的期刊或会议检索到相关工作的文献(如人工智能领域 NeurIPS/KSEM/TPAMI/TNNLS/NCA、网络与信息安全领域 USENIX Security/ACSAC/AsiaCCS/TrustCom/IH&MMSec、计算机图形学与多媒体领域 ICMR/ICASSP/TCSVT、并行与分布计算领域 ASPLOS/ICCAD、数据挖掘领域 PAKDD/arXiv 等).我们依据这些文章,对近年来研究者针对保护 DNN 模型版权所提出的神经网络水印方法进行了梳理,调研的神经网络水印相关文献来源分析如图 1 所示.可以看到,虽然总体的研究文献还不是很多,但是近几年发表的相关文献数量一直在增加,这说明 DNN 模型的版权保护工作越来越受到广大学者和科研人员的重视.

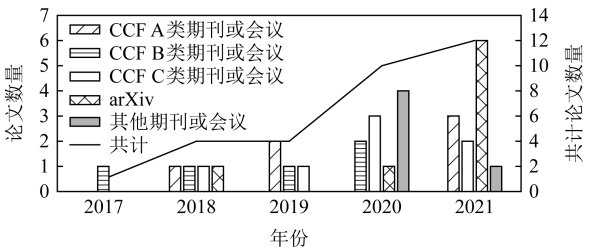


Fig. 1 Source analysis of research literature

图 1 调研文献来源分析

我们的主要贡献有 3 个方面:

1) 针对截至 2021 年 11 月之前研究者所提出的用来保护 DNN 模型版权的水印方法,从水印嵌入和提取的不同方式出发,将神经网络水印在原有的白盒水印和黑盒水印分类基础上,扩充了灰盒水印和无盒水印 2 种分类,并对白盒水印和黑盒水印方法根据其水印嵌入的不同思路和不同任务模型进行了更详细的分类总结,对 4 类水印方法的性能进行了对比,对当前神经网络水印的性能指标和攻击方法进行了整理汇总.

2) 现有的大多数黑盒水印方法针对图像分类模型进行保护,我们扩展梳理了通过构造触发图像和通过在输出图像中嵌入水印的方法实现对图像处理模型的版权保护工作,以及通过构造触发文本等方法实现对文本处理任务模型的版权保护.

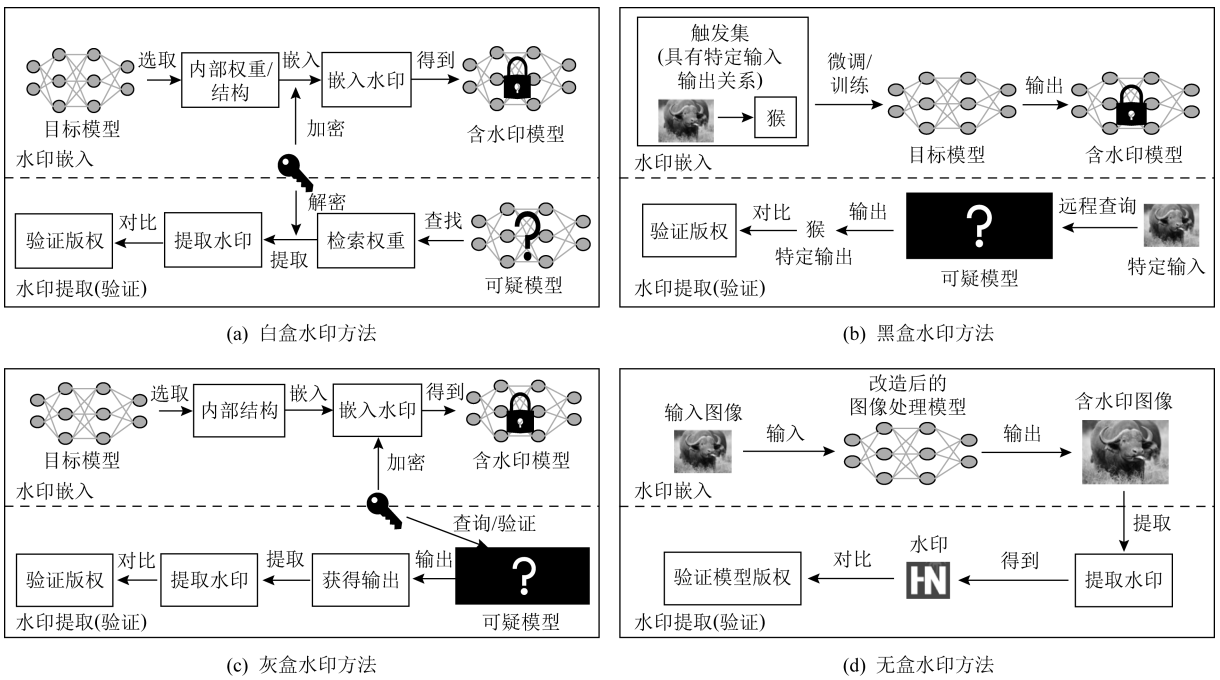
3) 结合现有神经网络水印的研究工作,从嵌入速度的提升、在音频处理等其他 DNN 模型中的应用、DNN 模型冗余和水印嵌入的不同位置与水印容量之间数学关系的理论证明、访问控制和主动防护的实现、DNN 模型的完整性保护以及形成统一的神经网络水印评估标准 6 个角度探讨了下一步神经网络水印的研究方向.

2 深度神经网络水印分类

我们从水印的嵌入和提取的不同方式出发,在原有的白盒水印和黑盒水印分类基础上,扩充了灰盒水印和无盒水印2种分类,如图2所示.

1) 白盒水印方法,如图2(a)所示.在白盒的场景下,模型所有者在目标模型的内部嵌入水印,提取水印时,目标模型的网络结构和内部权重等信息是已知的.因此,白盒水印假设DNN模型的所有者可以访问可疑目标模型的内部结构和权重,通过提取嵌入在模型内部的水印验证模型的所有权.

2) 黑盒水印方法,如图2(b)所示.与白盒情况不同,在黑盒情况下,模型所有者通过特定的输入输出构造触发集,用以改造模型.验证版权时,模型所有者不知道可疑目标模型的内部结构和权重,只能通过API来访问目标模型,从而获得特定的输出验证版权.事实上,模型被盗的大多数情况下,版权所有者只能通过API查询得到可疑模型的输出来验证模型的版权归属.



(a) 白盒水印方法 (b) 黑盒水印方法

(c) 灰盒水印方法 (d) 无盒水印方法

Fig. 2 Four types of watermarking methods

图2 4种类别水印方法

黑盒水印方法通常基于神经网络中的后门来构建,神经网络后门的一个简单示例如图3所示,原始样本输入分类网络后被正常分类为"牛",处理过的样本(如在图像中加入小方块)输入经过训练的分类网络后被错误分类为"猴".由此可见,通过构建具有特定输入输出关系的触发集(包含样本及指定的标签),并将该触发集与正常样本集一同训练,训练后特定的触发集由版权所有者保存,当出现可疑目标模型时,版权所有者将触发集中的样本输入可疑模型中,通过输出的指定标签来验证模型的版权.因此,黑盒水印方法多数情况下应用于分类任务的DNN模型中.

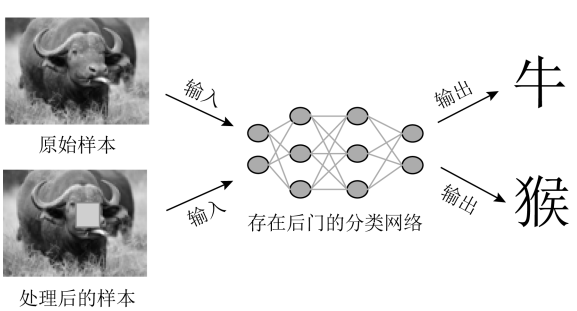


Fig. 3 Neural network backdoor example

图3 神经网络后门示例

3) 灰盒水印方法,如图2(c)所示.灰盒水印方法结合了白盒水印和黑盒水印方法的特点,既向模型的内部嵌入信息,又以黑盒的方式获得输出以验证模型版权.但与黑盒水印方法不同的是,黑盒水印方法的水印嵌入通过修改数据集继而调整模型实现在模型中嵌入水印,而灰盒水印方法则通过白盒思路直接在模型内部嵌入信息实现在模型中嵌入水印.

4) 无盒水印方法,如图 2(d)所示.无盒水印方法这个概念是我们在第一届中国媒体取证与安全大会(ChinaMFS 2020)[72]张新鹏教授的报告中首次获悉.无盒水印方法区别于白盒、黑盒、灰盒这 3 种水印方法,模型版权的验证既不需要在模型内部嵌入水印也不需要构建特定的输入输出对,即不再需要模型本身的参与.输入的图像经过模型输出后会携带水印信息,通过提取输出图像中的水印信息即可验证模型版权.截至目前,无盒水印方法主要关注图像处理任务的 DNN 模型.

本节从概念上对白盒、黑盒、灰盒和无盒 4 类水印方法进行了解释,并结合图示进行了说明.我们还统计了神经网络水印相关文献的发表数量,如图 4 所示;并以树状图的形式总结了神经网络水印的发展情况,如图 5 所示.可以发现,神经网络水印研究方向的大树在近 2 年逐渐枝繁叶茂.其中,黑盒水印方法因其验证的便捷性、无盒水印方法因其在图像处理领域 DNN 模型的应用,因此,在未来具有更好的发展优势.

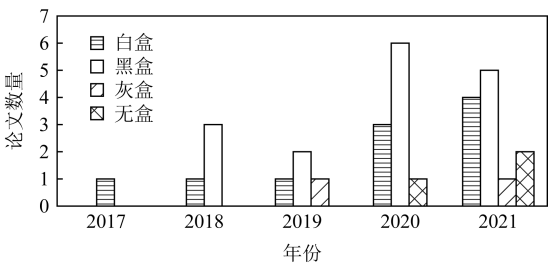


Fig. 4 Four categories of methods published in the past five years

图 4 4 种类别方法近 5 年文献发表量

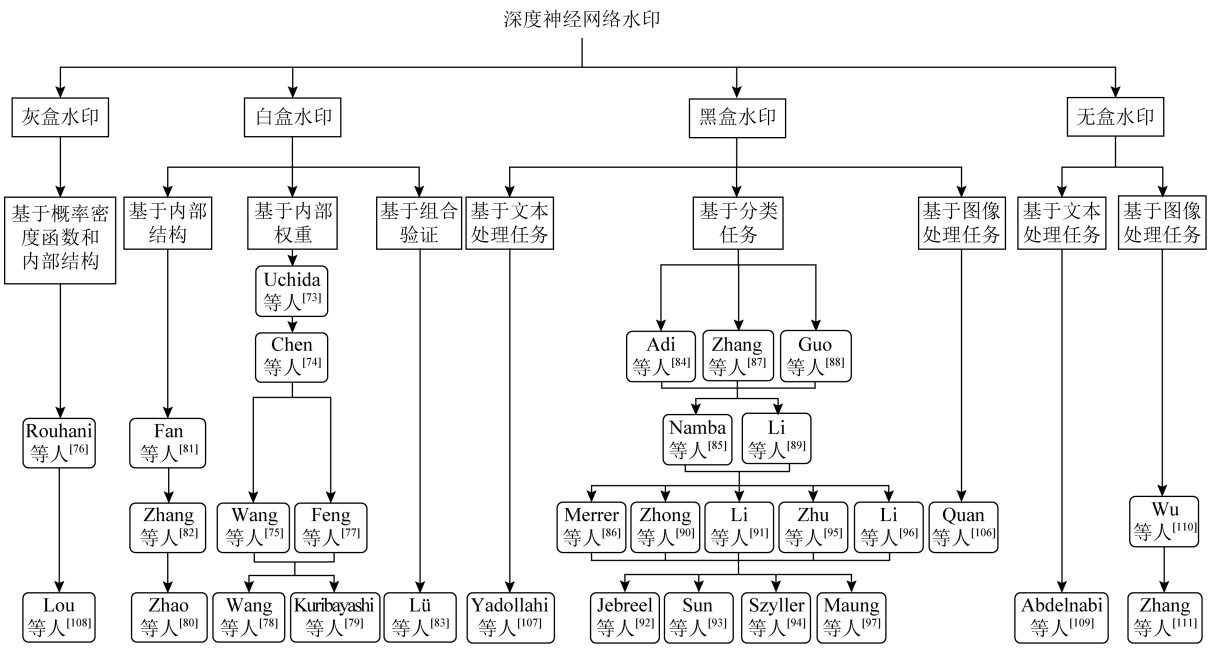


Fig. 5 DNN watermarking development diagram

图 5 深度神经网络水印发展图

3 深度神经网络水印相关工作

本节对神经网络水印方法进行详细的分类,总结了不同学者所做的相关工作.

3.1 白盒水印方法

白盒水印方法根据嵌入和验证方式的不同可以分为 3 类:基于内部权重的白盒水印方法、基于内部结构的白盒水印方法、基于组合验证的白盒水印方法.

3.1.1 基于内部权重的白盒水印方法

权重是神经网络模型内部参数的一种,表示神经单元之间连接的强度,反映了输入对输出的影响程度.基于内部权重的白盒水印方法是对神经网络模型中的权重进行修改以嵌入水印.

文献[73]首次提出了一个在 DNN 模型中嵌入水印的通用框架.将水印嵌入情况分为 3 种类型:训练嵌入、微调嵌入和蒸馏嵌入.其中,前 2 种情况是模型的版权所有者进行的水印嵌入,第 3 种情况是非版权所有者(如第三方云平台)受托代表版权所有

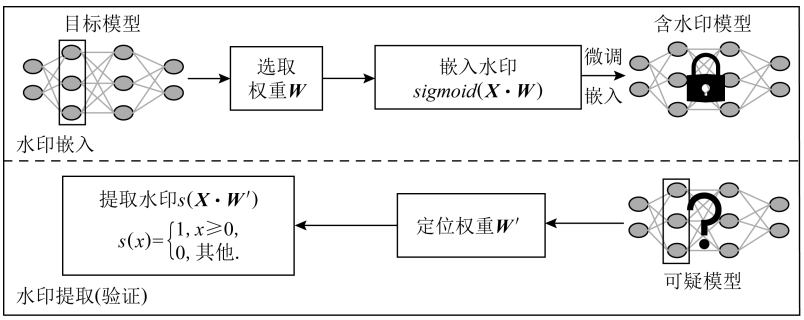


Fig. 6 The watermarking framework proposed in ref [73]

图 6 文献[73]提出的水印框架

者进行水印嵌入.微调嵌入水印如图 6 所示,使用参数正则化器 $\boldsymbol{X}$ 将一个参数矩阵水印嵌入目标模型卷积层的参数空间中.所提方案在损失函数中使用二元交叉熵项作为正则化器:

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

其中, $y_j = \sigma\left(\sum_i X_{ji}w_j\right)$, $\sigma(\cdot)$是 $sigmoid$ 函数,$\boldsymbol{X}$ 是嵌入矩阵(水印中的密钥),$\boldsymbol{X}\in\mathbb{R}^{T\times N}$,$\boldsymbol{W}$ 是网络层的权重向量,N 是要加水印的层的参数总数.因此,能够在训练过程中嵌入不影响模型预测精度的水印.实验证明,该方案嵌入的水印对模型微调和剪枝具有一定的鲁棒性.

由于之前的工作[73]主要解决了模型微调和剪枝攻击的鲁棒性问题,针对共谋攻击没有相关研究.基于此,文献[74]从追溯的角度出发,将用户指纹作为水印,提出了一种由用户和模型共同确定的端到端系统指纹框架 DeepMarks. DeepMarks 在保证模型性能不会大幅下降的基础上为每个用户分配一个唯一的二进制代码向量(也称为指纹),并将指纹信息嵌入模型权重的概率分布中,能够有效地跟踪每个用户模型的使用情况.实验表明,DeepMarks 可以有效地验证模型的所有权并追踪到侵权模型,并能抵抗潜在的攻击,如共谋攻击、参数修剪和模型微调.

文献[75]对文献[73]利用参数正则化器在模型权重中嵌入水印的工作进行了改进,所提出的框架如图 7 所示.该框架由 2 个神经网络组成:1)待加水印的目标模型;2)独立神经网络,用于将水印嵌入目标模型的权重中或从目标模型的权重中提取水印.独立神经网络可以是任意有效的神经网络.具体来说,该方案通过使用一个独立神经网络把水印信息嵌入目标模型的选择性权重中,将独立神经网络与目标模型一起训练.其中的关键问题在于如何从目标模型中选择最合适的权值作为独立神经网络的输入.2 种方法可以解决此关键问题:1)手动选择收敛权重作为输入;2)通过在输入和输入后的隐藏层之间插入一个新的可训练层来实现让独立神经网络自动选择合适的权重.训练过程中,目标模型的参数将根据 $\mathcal{L}_1$ 和 $\mathcal{L}_2$ 进行更新,而独立神经网络的参数将仅根据 $\mathcal{L}_2$ 进行更新.其中 $\mathcal{L}_1$ 代表目标模型的损失函数,$\mathcal{L}_2$ 代表独立神经网络的损失函数.训练过后只释放含水印的目标模型,将独立神经网络秘密保存,在所有权验证时重新启用.

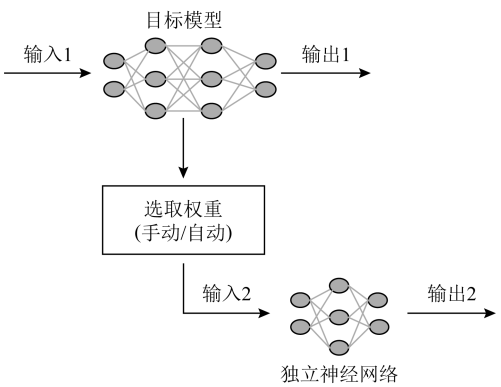


Fig. 7 The watermarking framework proposed in ref [75]

图 7 文献[75]提出的水印框架

实验表明:在模型微调和压缩等攻击下,独立神经网络能够以较小的损失嵌入和提取水印,且不会显著降低原任务的性能.然而,此类方法无法避免水印歧义攻击,攻击者可能使用新的独立神经网络来嵌入非法的水印,这将引起水印认证的模糊性.

由于文献[73]无法抵抗水印覆盖攻击的缺点,所以文献[76]在其基础上进行改进,提出将水印嵌入中间层激活图的概率密度函数中,解决了水印覆盖攻击的问题,但此方案嵌入水印的容量较小.因此,文献[77]在已有工作的基础上提出了带补偿

机制的水印嵌入方案.首先选取要伪随机嵌入水印的权重,然后对选取的权重进行正交变换,通过二值化方法在得到的系数中嵌入水印,并对水印系数进行逆正交变换得到新的水印权重.最后通过用带补偿机制的模型微调方案,可以在不破坏模型中水印的情况下,消除二值化带来的轻微精度下降.该方案不同于以正则化方式嵌入水印的方案,水印嵌入得更隐蔽,嵌入成本更低,稳定性更强,能有效抵抗水印覆盖攻击.在该方案中,与水印覆盖攻击相比,权重修剪对原始水印的损害更大.为了避免这种情况,该方案对原始水印进行扩频调制,并将调制后的水印分散嵌入模型各层的权重中,以增强水印的鲁棒性.

由于文献[73]使用的算法修改了模型参数的统计分布,利用这个漏洞不仅可以检测水印的存在,甚至可以导出其嵌入长度,并使用此信息通过覆盖来删除水印.为此,文献[78-79]通过不同的方案解决了这个问题.

文献[78]提出了一种基于对抗学习网络(类似于生成对抗网络)的白盒水印方法(robust white box GAN watermarking, RIGA).该方法使用水印模型作为生成器,检测模型参数统计分布变化的水印检测器作为鉴别器,在训练期间,鼓励生成器生成不可检测的水印,而鉴别器则尝试区分带水印和不带水印的模型,从而使嵌入水印前后模型的参数分布很难区分,水印不容易被攻击者检测到.在水印提取过程中,利用多层神经网络代替文献[73]中静态密钥矩阵 $\boldsymbol{X}$,提取网络与目标模型一起训练,实现快速收敛.实验表明,该方案不影响模型准确性,嵌入的水印可以是不同的数据类型,不仅可以是二进制数据,也可以是二维图像,进一步提高了水印的容量和鲁棒性,且隐蔽性很好.

文献[79]将重点放在微调目标模型的全连接层上,与文献[73]中的方法不同,该方法不再使用水印嵌入的损失函数,而是通过使用量化索引调制(quantized index modulation, QIM)方法的思想来控制由于水印嵌入而引起的变化量,使水印嵌入引起的变化很小.首先基于密钥随机选取部分全连接层的权重参数 n,对其进行离散余弦变换(discrete cosine transform, DCT)得到频率分量,利用水印修改其频率分量,然后进行逆 DCT 变换,训练模型更新权重参数.在嵌入操作中,模型权重初始值被改变以嵌入水印,由于水印的权重会随着学习过程的进行而发生轻微的变化,文献[79]认为水印信号在每次训练都加入了噪声,所以在每次训练更新权重参数后进行嵌入操作,以去除水印信号上的噪声.在对初始权值进行第 1 次嵌入操作时,水印嵌入引起的变化量在其余过程中必须是最大的.第 1 次嵌入操作后,在学习过程中更新模型中的所有权值,使其达到局部最小值.然后,在下一阶段学习过程后,由于训练引起的细微变化,对采样权值进行相同的嵌入操作来校正.在这里,可以看到第 2 个嵌入操作的变化比第 1 个操作的变化要小得多.同样,权值的变化在经过几个阶段的训练后会变小,随着训练的进行,权重的变化预计会收敛到零.也就是说,该方案不需要通过损失函数即可嵌入水印.此外,候选样本(全连接层的全部参数)的数量 N 远多于 n,这使得很难通过观察全连接层权重来分析水印的存在.

3.1.2 基于内部结构的白盒水印方法

由于在模型的内部权重中嵌入水印容易被攻击者移除和检测,从而使所有权保护失效,因此,研究者提出了基于内部结构的白盒水印方法,即更改目标模型的内部结构来达到嵌入水印的目的.

为了抵抗通过改变 DNN 模型参数来去除水印的各种攻击,文献[80]提出了一种利用网络剪枝进行的结构化水印方法.网络剪枝是一种常用的通过剪枝冗余成分来减小 DNN 规模的方法,在构建轻量级 DNN 中起着重要的作用.受到该技术的启发,文献[80]将水印嵌入给定的 DNN 冗余结构中,保证了 DNN 对其原始任务的性能.具体来说,提出的水印框架包括 2 个阶段:水印嵌入和水印提取.水印嵌入过程中,将水印分成几个比特段,并使用每个比特段对修剪率进行采样,然后将其分配给卷积层以进行由密钥控制的通道修剪.实际上,水印嵌入的目的是将包含所有权信息的二进制比特序列嵌入目标模型中,一旦目标模型被标记,就可以保护其所有权.水印提取过程中,由于水印的每个比特段与对应信道的剪枝率之间的映射实际上是双目标的,因此,可以通过检查标记信道的剪枝率来唯一地恢复所有比特段.实验表明,通过网络剪枝,该方法在不牺牲 DNN 模型可用性的前提下,能够可靠地重构水印,并提供较大的水印容量.同时也证明了该方法对常规水印方法中常见的变换和攻击具有很强的鲁棒性.

文献[81-82]提出通过在模型结构中添加一个额外的护照(passport),如在卷积层之后添加一个新的护照层,起到数字签名的作用,解决模型受到的歧义性攻击问题.

尽管已有的工作[73]具备了对微调和剪枝等攻击的鲁棒性,其中文献[76]还证明了嵌入水印对覆盖攻击的鲁棒性,但现有工作仍无法解决伪造水印造成的歧义性攻击问题.因此,文献[81]提出了一种基于护照的水印方法.该方法使预先训练的 DNN 模型的性能在有效护照存在的情况下保持不变,一旦护照被修改或伪造,原始模型的性能就会严重恶化,即利用护照来控制 DNN 模型的性能,据此开发所有权验证方案.具体来说,该方案在 DNN 模型的卷积层之后附加了一个护照层,如图 8 所示,其权重 γ 和偏置项 β 取决于卷积核 C_p 和护照 P:

$$O^l(X_p)=\gamma^l X_p^l+\beta^l=\gamma^l(C_p^l * X_c^l)+\beta^l, \quad (2)$$

$$\gamma^l=Avg(C_p^l * P_\gamma^l),\ \beta^l=Avg(C_p^l * P_\beta^l), \quad (3)$$

其中,$*$ 表示卷积操作,l 是层数,X_p 是护照层的输入,X_c 是卷积层的输入,$O()$是输出的相应线性变换,而 P_β^l 和 P_γ^l 是用于导出权重和偏置项的护照.对于使用护照层的 DNN 模型,其性能取决于运行时的护照层,因为其权重 γ 和偏置项 β 是根据护照进行计算的.

利用所提出的基于护照的方法,文献[81]设计了 3 种所有权验证方案,如表 2 所示,其中 V2 和 V3 是通过多任务学习实现的,采用了 Group Normalization.实验证明,该方案对模型修改具有鲁棒性,并且成功抵抗了水印的歧义攻击.

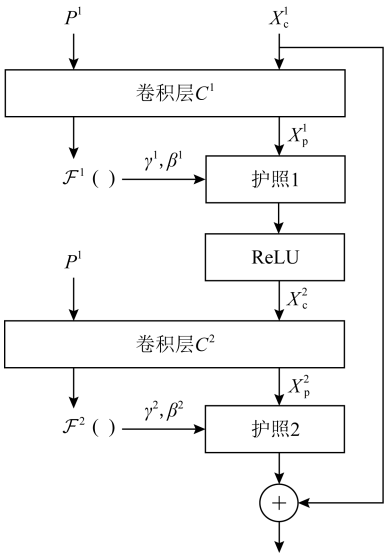


Fig. 8 Passport layer in ResNet architecture[81]

图 8 ResNet 架构中的护照层[81]

Table 2 Three Passport-Based Ownership Verification Schemes

表 2 3 种基于护照的所有权验证方案

方案	内容	优点	缺点
V1:护照和经过训练 DNN 一起分发	经过训练的 DNN 和护照一起分发给用户,用户使用给定护照执行网络.	①对 DNN 模型修改具有鲁棒性; ②可以抵抗歧义攻击.	①使用护照增加 10%的额外计算成本; ②增加了保护护照安全的额外责任.
V2:护照嵌入但不分发	交替实现多任务学习,即最小化不包含护照层时原始任务损失和包含护照层的组合损失,训练可以在不需要护照的情况下执行任务的 DNN 模型并分发给用户.在产生版权纠纷时向 DNN 模型添加护照层进行所有权验证.	①对 DNN 模型修改具有鲁棒性; ②可以抵抗歧义攻击; ③不需要护照,用户易于使用; ④不会产生额外的计算成本.	①需要访问 DNN 的内部权重并添加护照层进行所有权验证.
V3:护照和触发集都嵌入但不分发	与 V2 的不同之处在于,除了嵌入护照外,还嵌入一组触发图像.版权纠纷时首先在黑盒模式下声明所有权,然后在白盒模式下通过护照重新声明所有权.	①对 DNN 模型修改具有鲁棒性; ②可以抵抗歧义攻击; ③不需要护照,用户易于使用; ④不会产生额外的计算成本; ⑤可以通过远程 API 来探测和声明可疑 DNN 模型的所有权.	

由于文献[81]的方案只适用于一些特殊的归一化层(如 Group Normalization),因此对应用其他归一化层(如 Batch Normalization)的目标网络模型来说,需要替换所有 Batch Normalization 层来改变网络结构,才能进行保护,否则将不能运用该方案.但是对于许多任务,改变网络结构会导致显著的性能下降.

受到文献[81]工作的启发,文献[82]基于不改变网络结构和减少性能下降的目标,提出了一种适用于大多数流行的归一化层的新的护照识别归一化公式,只需要添加一个额外的护照识别分支来保护知识产权.训练过程中预定义了一些秘密护照,这些额外的分支与目标模型联合训练,经过训练后,这些秘密护照和新的分支都将由模型所有者保存,用来以后进行所有权验证,只有原始的目标模型被交付给用户.因此,从用户的角度来看,网络结构没有变化.此外,由于护照识别分支的归一化统计量(例如,Batch Normalization 的均值和方差)被设计成独立

计算,因此,对目标模型的性能影响将很小.当出现可疑模型时,模型所有者可以将秘密护照和新的分支添加回来进行所有权验证.目标模型的性能只有在给出正确的护照时才会保持不变,同时对于伪造的护照,目标模型性能会严重下降.实验证明,该方案对模型修改具有鲁棒性,可以抵抗歧义攻击,并且适用于大多数流行的归一化层.

3.1.3 基于组合验证的白盒水印方法

基于组合验证的方式是把水印分为 2 个部分:1)嵌入网络模型;2)由所有者保存,验证时将二者合二为一进行验证.

为了进一步提高水印对模型微调和剪枝攻击的鲁棒性,同时解决非法用户伪造水印进行的歧义性攻击,文献[83]在前人的基础上结合中国古代虎符的文化元素,提出了一种新颖的白盒水印方法 HufuNet,用于保护 DNN 模型的知识产权,如图 9 所示.该方案通过训练一个具有少量参数的神经网络,即 HufuNet,以获得较高的测试精度.其中,测试集与训练集具有相同的分布,二者都向公众发布,用于训练和验证 HufuNet.HufuNet 经过训练和测试后,分成 2 个部分:1)HufuNet 的所有卷积层,该部分作为水印嵌入 DNN 模型中,用于所有权保护;2)HufuNet 的全连接层,该部分由模型所有者保存,用于所有权验证.在含有水印的 DNN 模型训练过程中,冻结来自 HufuNet 的参数,同时更新模型中的其他参数,以确保其在主要任务上的性能.在所有权验证阶段,模型所有者从可疑 DNN 模型中提取嵌入的水印,并相应恢复 HufuNet 嵌入宿主 DNN 的左半部分,与所有权者保存的右半部分 HufuNet 合并成一个完整的 HufuNet,用以对可疑模型声张所有权.与之前的水印方法相比,基于 HufuNet 的所有权验证对模型微调、剪枝更鲁棒,对水印伪造更安全,对卷积核去除和补充攻击鲁棒,可以抵抗水印歧义攻击.同时,该方法对 DNN 的性能影响可以忽略不计.

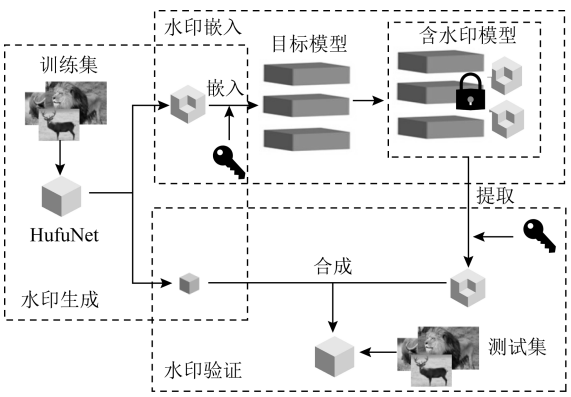


Fig. 9 HufuNet framework[83]

图 9 HufuNet 框架图[83]

白盒水印方法经过几年的发展已经较为成熟,但由于提取水印时需要了解模型的内部结构,限制了此类方法的实际应用.黑盒水印方法通过访问 API 即可进行验证,具有更广泛的应用前景,但是由于当前的黑盒水印方法修改了模型的训练数据集,因此必然会对模型的准确性造成或多或少的影响,这对一些高精度的应用环境(如自动驾驶、癌症诊断等)来说是无法接受的.因此,在此类应用的模型保护中,白盒水印方法由于可以在没有精度损失的情况下工作而受到关注.

3.2 黑盒水印方法

黑盒水印方法根据 DNN 模型任务的不同可以分为 3 类:基于分类任务、基于图像处理任务和基于文本处理任务.

3.2.1 基于分类任务的黑盒水印方法

基于分类任务的黑盒水印方法中模型所有者通过构造具有特定输入输出对的触发集,训练模型以达到通过触发集验证模型版权的目的.

1) 仅通过标签更改构造触发集

标签更改是指对原始样本对应的正确标签进行修改,改为由版权所有者特定的与原始样本内容不符的标签.仅通过标签更改构造触发集属于零比特水印方法.

文献[84]用一组抽象的图像和与图像内容不符的标签构造触发集(引入可信的第三方,用于所有权验证),这些图像彼此无关,也与训练样本无关,触发集图像的标签是随机分配的,如图 10 所示.之后利用触发集和原始训练集来训练目标模型.训练好的 DNN 当给定有标注的输入时,带水印的 DNN 模型会输出特定的标签,达到验证模型版权的目的.值得一提的是,该方案很难将抽象的图像与所有者的身份关联起来,引入可信的第三方虽然为核查过程

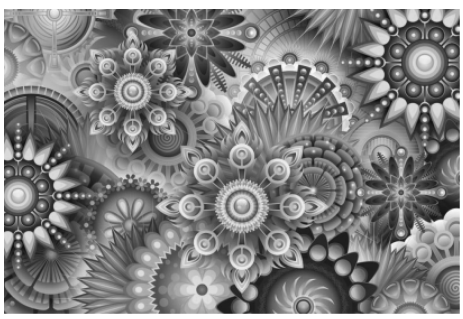


Fig. 10 Abstract image marked as "car"[84]

图 10 被标记为"汽车"的抽象图像[84]

提供了严格的安全性,但实际服务中因需要第三方的参与往往具有昂贵的代价.

不同于文献[84]的方法,文献[85]首先引入了一种新的神经网络水印攻击方法,即查询修改攻击.在查询修改攻击中,当给出一个查询时,查询修改处理将使用一个自动编码器来确定该查询是否是一个密钥样本.如果该查询被检测为密钥样本,则会修改图像,使验证过程失败.自动编码器是一种特殊的DNN,它首先将输入图像压缩成低维数据,然后将这些低维数据解压缩为输出图像.结果,可以去除或稀释输入图像中的噪声.这样,水印密钥样本中的标志或噪声(水印模式)可以通过查询修改攻击被移除.

针对以上攻击方法,文献[85]提出了一种基于标签更改和指数加权的黑盒水印方法.该方法包括2个部分:①通过标签变化生成密钥样本;②通过指数加权嵌入密钥样本.首先,在密钥样本生成过程中,随机选择一个未经修改的训练样本;然后,将该样本的标签更改为与原始样本不同的错误标签以生成密钥样本.模型将除了这个特定的密钥样本识别为错误标签外,其余原始样本都被正确识别.密钥样本本身没有任何标记,不会被未经授权的服务提供者检测到,但直接使用这些密钥样本进行训练会导致模型过拟合.因此,在训练过程中,识别出对发布预测有显著作用的神经网络模型参数,并将其权重值指数增加,使其不能在模型修改处理前后改变样本(包括密钥样本)的预测行为,提高了对模型修改的鲁棒性.实验表明,该方案可以同时抵抗模型修改和查询修改造成的水印失效攻击.

不同于文献[84-85]的零比特水印方法,文献[86]提出了一种利用对抗样本构建触发集的新的零比特水印方法.首先,通过对原始输入添加一个小的扰动来创建对抗性示例,用标签对其进行了重新标记,这样决策边界的正常图像一部分被分配使用错误标签,而另一部分则继续使用原来的正确标签.然后,使用重新标记的对抗性示例对模型进行微调,微调过程中,决策区域的边界分布在对抗性示例周围,如图11所示.最后,使用对抗性示例作为触发密钥,它们的类标签作为验证密钥.在验证阶段,没有此水印的模型很可能会误分类此类对抗性示例,而带有水印的模型则有望正确识别此类对抗性示例.因此,模型所有者以此来验证模型所有权.

仅通过标签更改构造触发集的方法是黑盒水印方法最初的工作思路,并在以后的工作中得到发展.

2) 通过在原始样本中嵌入信息和标签更改构造触发集

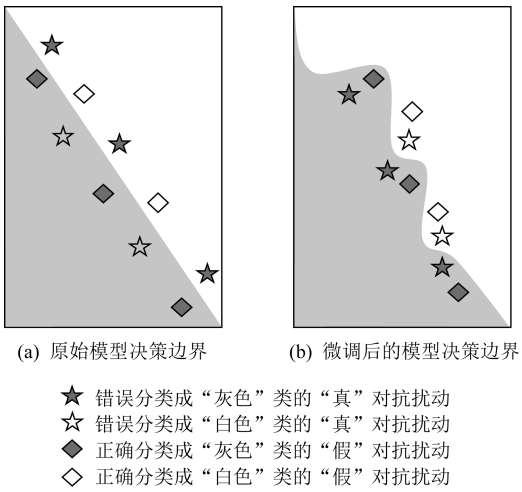


Fig. 11 The decision boundary of the binary classifier[86]

图11 二元分类器决策边界示意图[86]

文献[84-86]只对样本进行标签更改,不支持嵌入版权所有者的信息以提供强有力的版权证明.与之不同,文献[87-89]在标签更改的基础上,在原始样本中嵌入了一些信息,这类方法支持相对较大的有效载荷,以提供强有力的所有权证明.

文献[87]研究了3种水印生成算法,将有意义的文本内容和无意义的噪声嵌入图像样本中作为水印,或将不相关的样本指定错误标签后将其作为水印(仍属于仅通过标签更改的方法),以此为基础提出了一种将水印嵌入目标模型的后门水印方法,并设计了一种远程验证机制来确定模型的所有权,该方法的框架如图12所示.首先通过水印生成算法为目标模型的所有者生成定制水印和预定义的标签,这些水印在以后所有权验证阶段使用.其次,该框架通过训练将生成的水印嵌入目标模型中,目标模型自动学习并记忆水印.最后,新生成的含水印模型能够进行所有权验证.一旦模型被盗并被用于提供服务,所有者可以通过发送水印作为输入并检查服务的输出来验证版权.实验证明,该方案对模型微调和剪枝具有一定的鲁棒性,可以准确并快速验证可疑模型的所有权,而不影响正常输入数据的模型准确性.

DNN越来越多地应用于智能家居、虚拟现实/增强现实(VR/AR)、机器人和自动驾驶汽车等新兴行业.在这些场景中,底层嵌入式系统通常在本地运行DNN,以解决延迟和隐私问题.在不久的将来可能会出现基于DNN的专有软件开发工具包(software development kit, SDK),与基于云的API不同,本地SDK更容易受到未经授权的复制和分发.因此,

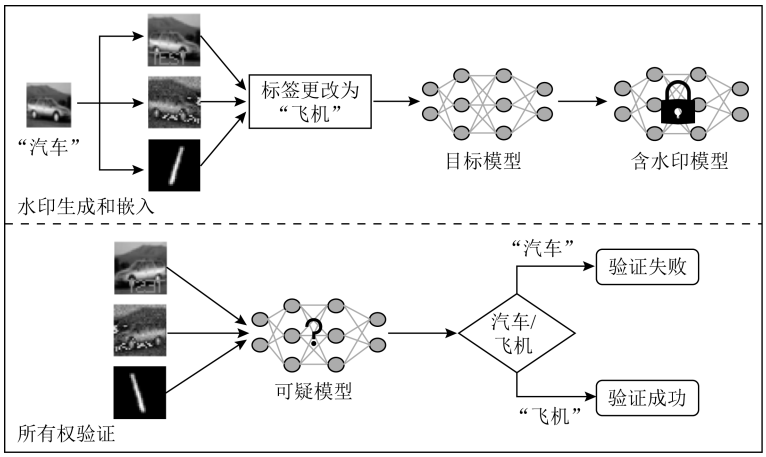


Fig. 12 The watermarking framework proposed in ref [87]

图 12 文献[87]提出的水印框架

文献[88]提出了一种适用于嵌入式应用的神经网络水印框架.具体来说,首先选择原始数据集的一部分,并根据模型所有者的签名添加某些修改(具体修改操作只有模型所有者知道).然后同时使用原始数据集和修改后的数据集微调初始模型(使用现有的权重作为初始化),使训练后的含水印模型遇到任何嵌入版权所有者签名的输入时会以预定义的特殊模式运行,一个简单的示例如图 13 所示.所有权验证通过比较原始目标模型和含水印模型在修改输入上的行为进行.在所有权验证阶段模型所有者必须公开其签名以及如何对数据集进行修改,以证明模型所有者的身份.实际上,此方法与文献[87]的方法有异曲同工之妙.与文献[84]相比,这 2 种方法不需要指定的输入和第三方参与,因此,在证明版权所有者身份时几乎不需要额外的开销.

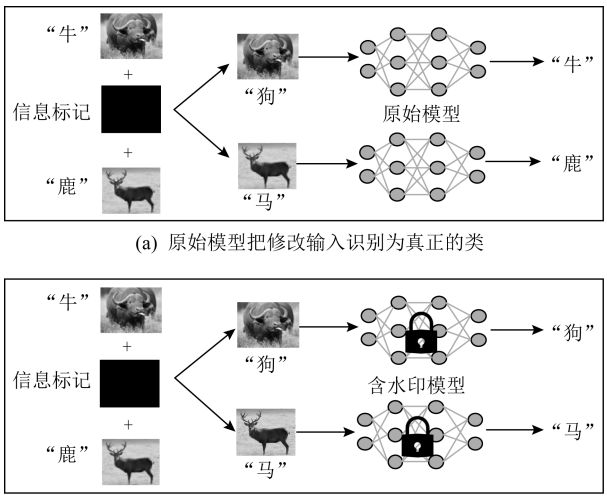


Fig. 13 A simple example of the scheme in ref [88]

图 13 文献[88]方案的一个简单示例

由于之前的神经网络水印方法[84]可以被攻击者通过检测密钥样本,轻松构建检测器来逃避 DNN 模型所有者的检测.另外,攻击者可以很容易地通过建立一组假样本进行歧义攻击,使模型具有攻击者的水印行为.为了解决这些问题,文献[89]首次提出基于盲水印的 DNN 知识产权保护框架,如图 14 所示.

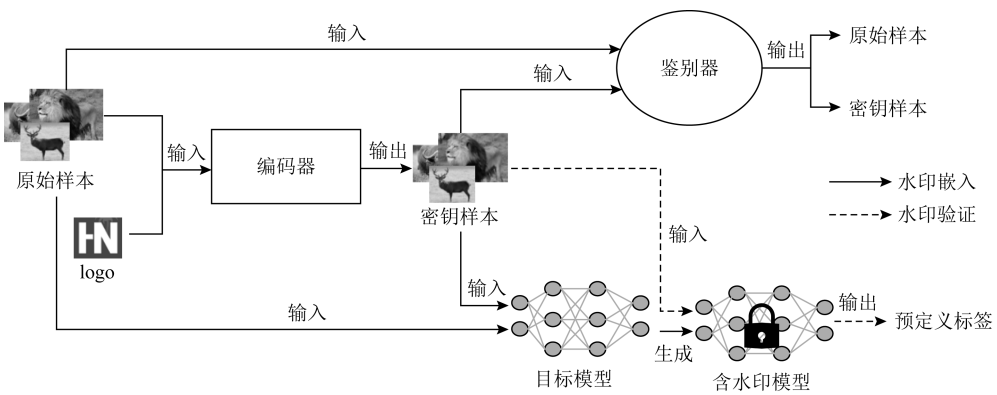


Fig. 14 The blind watermarking framework proposed in ref [89]

图 14 文献[89]提出的盲水印框架

该框架以原始训练样本和 logo 等标识信息作为输入,通过编码器生成与原始样本分布相似的密钥样本,并通过预定义的标签将这些密钥样本嵌入 DNN 中.鉴别器的本质是生成式对抗网络,用来确定是密钥样本还是原始样本,同时充当检测器,以检测编码器是否生成了密钥样本.

实验证明该框架可以有效地验证所有权,并且不会对目标模型的原始任务产生显著的副作用,具有一定的鲁棒性,对规避攻击具有不可检测性,对欺骗性所有权声明具有不可伪造性,在模型和模型所有者身份之间建立了明确的联系.

通过在原始样本中嵌入信息和标签更改构造触发集的方法在标签更改的基础上嵌入模型所有者的信息,可以提供强有力的所有权证明.

3) 通过添加新的类构造触发集

基于后门的黑盒 DNN 水印方法依赖于密钥样本,分配具有错误标签的密钥样本将不可避免地或多或少地扭曲原始决策边界.为此,文献[90]通过在训练过程中对精心制作的密钥样本添加新的类标签对模型添加水印,最大限度地减少(甚至消除)原始决策边界扭曲的影响.

具体而言,如图 15 所示,设原始目标模型的任务是预测 $N-1$ 个不同的类.添加一个新的类后,含水印模型的任务变为预测 N 个不同的类,那么无水印模型不能输出一个不存在的类标签.所提方案由 3 种算法组成:KsGen,TrEmb 和 Ver. KsGen 以原始数据集 D 的子集和秘密 S 作为输入,之后将所有精心制作的样本分配给第 N 个标签,并输出密钥样本数据集 D_S. TrEmb 将原始数据集 D 和来自 KsGen 的结果作为输入,并输出含水印的模型.Ver 将可疑模型和来自 KsGen 的结果作为输入,对可疑模型进行验证.实验证明,该方案具有一定的鲁棒性,含水印模型具备较高的保真度.另外,由于使用较少的训练密钥样本和较弱的扰动强度,嵌入的水印不易被检测.

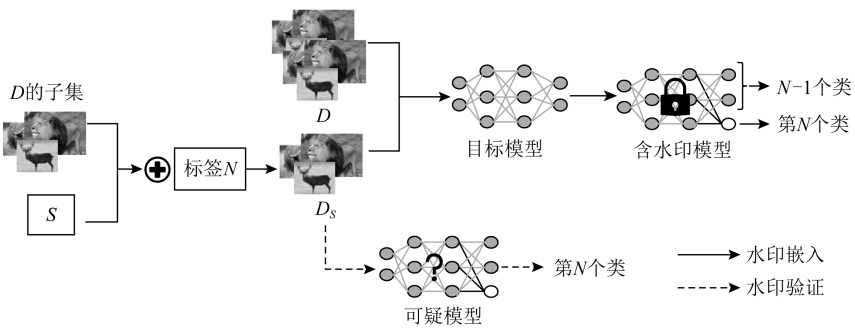


Fig. 15 The watermarking framework proposed in ref [90]

图 15 文献[90]提出的水印框架

由于现有的基于触发集的黑盒水印方法[84,87-90]构造的触发集本身的鲁棒性较差,内部攻击者窃取触发集后可以通过伪造触发集来获取其所有权.因此,为了解决触发集泄露造成的版权问题,文献[91]提出了利用图像水印领域常用的基于变换域的水印方法来构造触发集.因为基于变换域的水印具有较高的隐蔽性和对信号处理操作的鲁棒性,所以可以增强触发集的鲁棒性.该方案分为基于变换域的触发集生成、水印嵌入和所有权验证 3 个阶段.触发集的生成是从原始数据集中选取每个类的部分图像,通过在选取图像的变换域中插入基于块的特征来实现.利用原始数据集和生成的触发集共同训练目标模型,训练时将目标模型的输出层增加一个 Δ 类.训练模型使触发集输入后被分类为 Δ 类,从而验证版权.实验结果表明,该方法在保真度和效率方面可与文献[87-88]提出的方法相媲美,在有效性、抗剪枝攻击和抗歧义攻击方面优于它们,此外可以解决触发集泄露造成的模型版权失窃问题.

通过添加新的类标签构造触发集的方法在已有工作的基础上最大限度地减少了原始决策边界扭曲的影响.

4) 通过添加嵌入信息的附加样本构造触发集

由于之前的大多数方法只能满足表 4 中神经网络水印的部分要求,因此,文献[92]提出了 KeyNet 水印框架,该方案提供了保真度、鲁棒性、可靠性、完整性、容量、安全性、身份验证、唯一性和可扩展性,满足了几乎所有的水印要求,且与之前的黑盒水印工作相比,该方案根据嵌入不同用户的签名生成不同的触发集,微调嵌入水印后分发给对应的用户,可以达到溯源的目的.

具体来说,KeyNet 框架如图 16 所示.首先将所有者的签名嵌入水印载体样本中不同的位置,对应

不同的标签,从而生成触发集(该方案嵌入5个位置:左上、右上、左下、右下和中间,分别对应标签1~5).通过多任务学习的方式,利用触发集和原始数据集共同训练原始分类任务和水印任务,此过程中将另外一个私有模型添加到原始目标模型之后,以原始目标模型的输出作为输入,输出特定标签.之后,所有者将含水印模型分发,保留私有模型用作密钥.验证过程中,将嵌入签名的触发集样本输入可疑模型,得到的预测结果传递给私有模型,通过私有模型提供不同位置签名与其标签的准确对应关系验证版权.

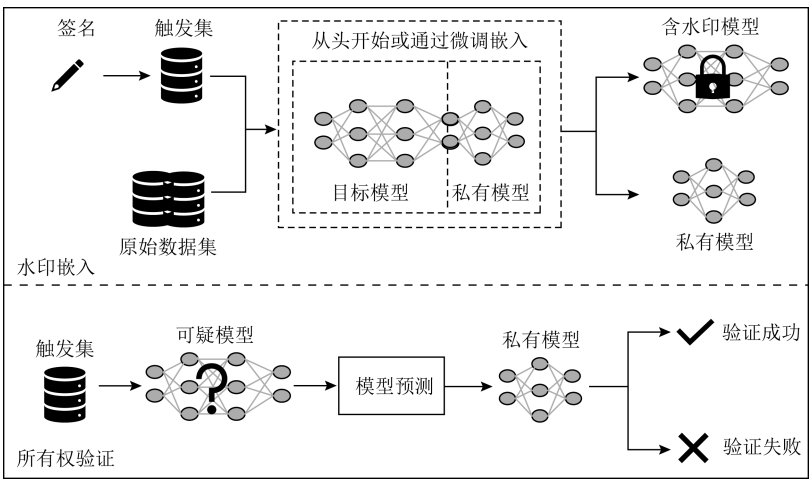


Fig. 16 KeyNet framework[92]

图16 KeyNet框架[92]

文献[93]发现大多数现有的水印方法不支持用户指纹的认证和管理,无法满足商业版权保护的要求.另外,文献[85]提出的查询修改攻击可以使大多数现有的基于后门的水印方法无效.为解决这些问题,文献[93]提出了一种通过使用其他类别和隐写图像来保护DNN的知识产权的方法,其框架如图17所示.首先选择原始训练数据集外的少量图像作为水印密钥样本.随后,用户的指纹通过最低有效位(least significant bit, LSB)图像隐写技术被隐藏在每个水印密钥样本中,为每个用户分配一个唯一的指纹图像,以便以后可以对用户的身份进行身份验证.为了在DNN中嵌入水印,首先给所有水印密钥样本分配一个新的类标签,然后将水印密钥样本添加到训练集中来训练一个带水印的DNN.当输入水印密钥样本时,带水印的DNN可以输出预定义的类标签.由于使用的是训练集之外的图像作为水印密钥样本,而不是在原始训练集图像上叠加嵌入图案,因此,所提出的水印方法可以抵抗查询修改攻击[85].另外,用户指纹认证的有效性应同时满足2个条件:①从用户提交的图像中提取用户指纹;②提交的指纹图像被水印模型分类为附加类别.只有同时满足2个条件,用户才能成功通过指纹认证.

实验结果表明,该方法在不影响模型测试精度的情况下,能够实现100%的水印精度和100%的指纹认证成功率.同时,该水印方法对微调和剪枝攻击以及查询修改攻击具有鲁棒性,可以有效地保护DNN的版权.

通过添加嵌入信息的附加样本构造触发集的方法在已有工作的基础上减小了添加触发集后对原始模型精度的影响.

5) 其他方法

模型功能也可以通过模型提取来窃取,但此前的文献并没有关注该问题.模型提取是指攻击者通过API访问原始模型,然后使用返回的结果来训练代理模型.为了解决该问题,文献[94]提出一种神经网络动态对抗水印方法(dynamic adversarial watermarking of neural networks, DAWN),该方法首次使用水印来阻止通过模型提取进行的所有权盗窃.与先前的水印方法不同,DAWN不会对训练过程进行更改,而是通过动态更改来自API的一小部分查询(例如,小于0.5%)的响应,并在受保护模型的预测API上运行.该集合充当模型水印,在客户端通过查询来训练代理模型的情况下嵌入该水印.实验证明,DAWN方法能够有效地对所有提取的代理模型进行水印处理,使模型所有者证明其所有权.

与文献[81-82]利用护照通过白盒的方式解决歧义攻击问题不同,文献[95-97]利用不同的思路通过黑盒方式解决了歧义攻击造成的DNN模型的盗版问题.

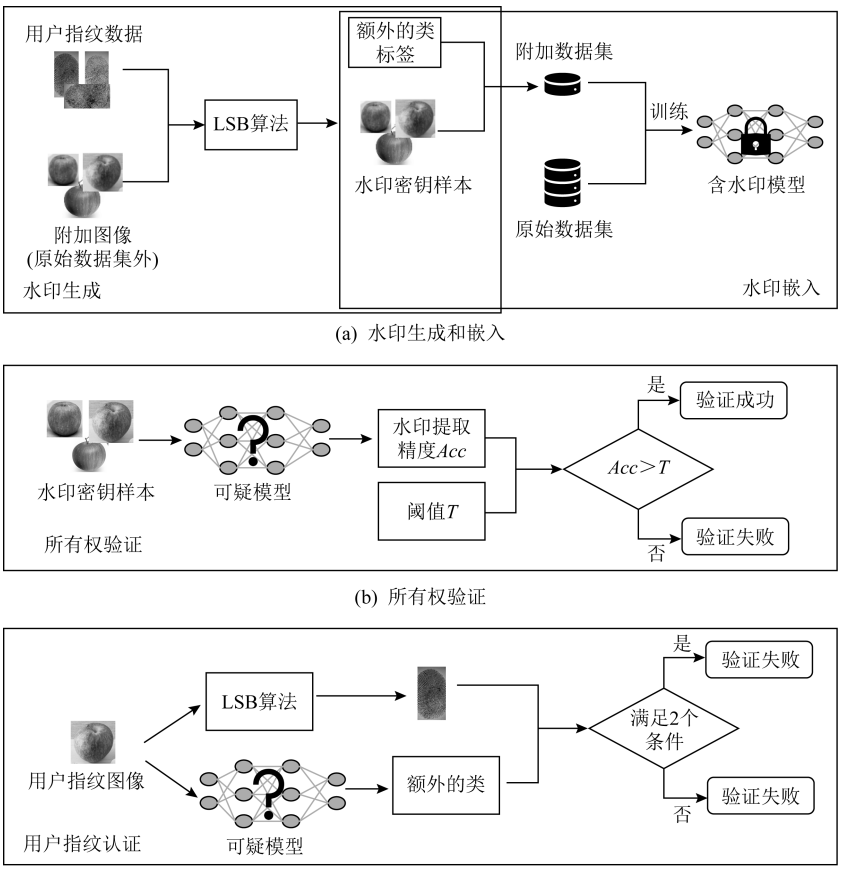


Fig. 17 The watermarking framework proposed in ref [93]

图 17 文献[93]提出的水印框架

文献[95]提出了一种抗伪造攻击的神经网络水印协议.该方案引入单向哈希函数 SHA256,使证明所有权的触发样本必须形成单向链,如图 18 所示,并且它们的标签也被赋值.使用此方法,没有网络训练权的攻击者不可能构造出触发样本链或者触发样本与指定标签的匹配关系,因此,该协议能够在不牺牲网络性能的情况下抵抗水印伪造攻击.

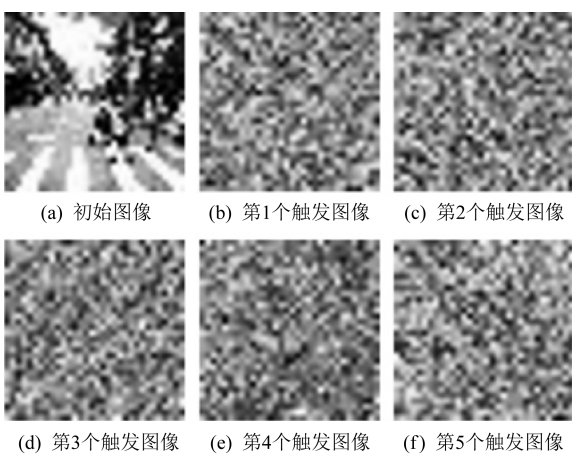


Fig. 18 The initial image and five trigger images in the Hash chain[95]

图 18 哈希链中的初始图像和 5 个触发图像[95]

由于之前的神经网络水印大多被设计成增量训练和微调嵌入,那么攻击者可以使用相同的机制来嵌入更多的水印到一个已经有水印的模型中,从而造成歧义攻击,对原始模型进行盗版.基于这些不足,文献[96]提出了一种抗盗版水印的新方法——空嵌入(null embedding).空嵌入方法不依赖增量训练,只能在模型的初始训练中使用水印比特序列来修改用来训练模型正常分类规则的有效优化空间.模型所有者将从一个未经训练的模型开始,通过空嵌入方法生成与之相关的额外训练数据,并使用原始和额外的训练数据来训练模型,在模型的正常分类精度和水印之间建立强相关性.因此,攻击者不能通过调整或增量训练来移除嵌入的水印,也不能向已经有水印的模型添加新的盗版水印,达到抗盗版的目的.

由于文献[96]提出的空嵌入方法未对广泛应用于图像分类任务的残差网络 ResNet 等大型网络进

行有效性验证,所以文献[97]在此基础上提出了基于密钥的分块图像变换水印方法,首次将分块图像变换应用于神经网络水印,利用了与文献[95-96]不同的技术,解决了歧义攻击造成的 DNN 模型的盗版问题.具体来说,该方案从训练集中选取一组图像,利用密钥按照文献[98]中的算法对其进行分块变换(块的大小 M 可以设置不同的值),变换后的图像示例如图 19 所示.水印嵌入过程中用变换后的图像和原始图像一起训练 DNN 模型.水印检测过程中利用密钥对测试图像变换后进行验证,若水印检测精度大于设定的阈值,则可验证版权.该方案使用密钥作为版权验证的关键,不需要预先设置触发集,其安全性依赖于密钥而不是算法的保密性,符合 Auguste Kerckhoff 的原理,且该方案具有抗盗版性和计算成本低的优点.

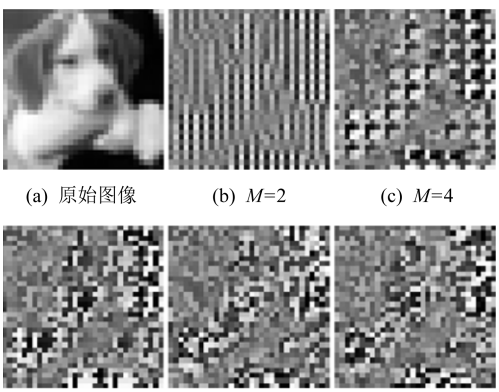


Fig. 19 Example of block-wise transformed images[97]

图 19 按块变换的图像示例[97]

此部分主要梳理了通过黑盒方式解决歧义攻击造成的 DNN 模型的盗版问题.

3.2.2 基于图像处理任务的黑盒水印方法

3.2.1 节基于分类任务中提出的所有黑盒水印方法都是应用于图像映射到标签的分类模型的版权保护中,但对图像映射到图像的图像处理模型的保护很少提及,如图像去噪[99-100]、图像增强[101-102]、超分辨率[103]、图像修复[104]、风格转换[105]等任务.如表 3 所示,图像处理模型与分类模型不同,因此,不能直接将分类模型的水印方法应用于图像处理模型,相对来说,图像处理模型的保护更具有挑战性.

文献[106]首次提出一种解决图像处理模型版权保护问题的黑盒水印方法.该方法的思路是通过微调 DNN 操纵模型在特定域中的预测行为,使修改后的模型的输出图像接近预定义的结果.特定域形成所有可能的触发图像空间,并将预定义的结果用作验证图像,水印验证通过检查输入的触发图像是否可以在可疑模型的输出中看到它们相应的验证图像来完成.

具体来说,所提方法的框架如图 20 所示.首先生成触发图像和初始验证图像,在水印嵌入中,2 个图像都用于微调目标模型.然后,将触发图像输入标记模型中,输出用于更新验证图像.触发图像和验证图像由所有者保留.在水印验证中,验证者将所有者的触发图像输入可疑模型中,然后将输出与所有者的验证图像进行比较以进行判断.实验结果表明,该方法满足保真度、唯一性和容量的要求,对模型压缩、模型微调和水印覆盖等攻击具有鲁棒性.重要的是,该水印技术在图像处理任务中具备推广使用的价值.

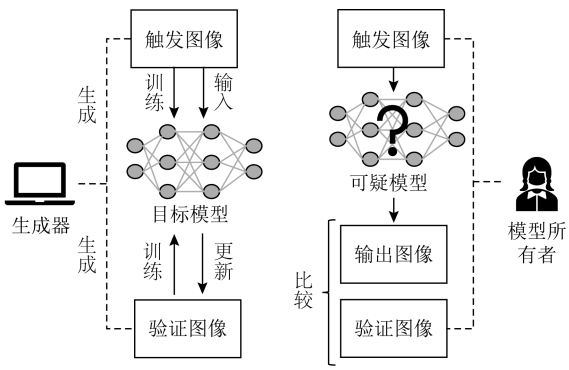


Fig. 20 The watermarking framework proposed in ref [106]

图 20 文献[106]提出的水印框架

Table 3 The Difference Between Image Processing Model and Classification Model

表 3 图像处理模型与分类模型的区别

序号	图像处理模型	分类模型
1	输出结果为图像	输出结果为标签
2	模型相对较浅,冗余度较小	模型相对较深,冗余度较大
3	寻找输入图像的低维流形	寻找不同类别的决策边界

3.2.3 基于文本处理任务的黑盒水印方法

文本处理是许多机器学习领域中最常见的任务之一,在语言翻译、情感分析和垃圾邮件过滤等方面有许多应用.因此,对文本处理任务的 DNN 模型的版权保护同等重要.神经网络水印技术最近也出现了在文本领域的工作.

在文献[107]的工作中,提出了一种对文本处理 DNN 模型进行安全水印的框架.该框架的 3 个主要

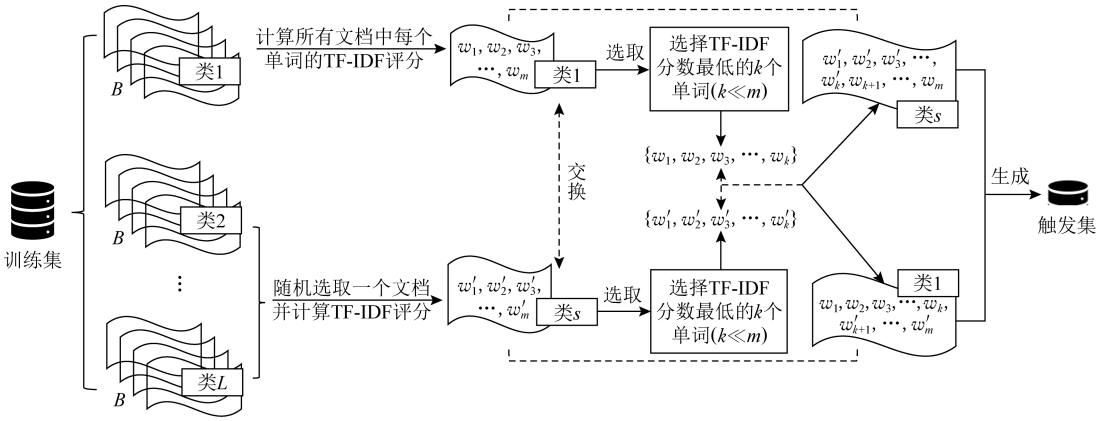


Fig. 21 The watermark generation framework in ref [107]

图 21 文献[107]中的水印生成框架

部分是水印生成、水印嵌入和水印验证.如图 21 所示,在水印生成步骤中,从训练集中随机选择 B 个样本,计算所有文档中每个单词的 TF-IDF 分数,对于每个选定的文档,从另一个类中随机选择一个文档来交换它们的单词并生成水印记录.选择 2 个文档中 TF-IDF 分数最低的 k 个单词,交换选定的单词和 2 个文档的标签.修改后的文档录入触发集.重复以上过程生成整个触发集.在水印生成步骤中,将原始训练集与生成的触发集结合,对 DNN 模型进行训练,得到含水印模型.如图 22 所示,在水印验证步骤中,使用触发集验证模型的所有权.该方法在不降低原任务性能的前提下,对参数修剪等已知攻击具有较强的鲁棒性.实验结果表明,该水印模型能够准确地提取出水印,从而准确地验证经过训练的模型的所有权.但显而易见的问题是经过修改后的句子有语法错误且语意不通顺,很容易被攻击者检测到.

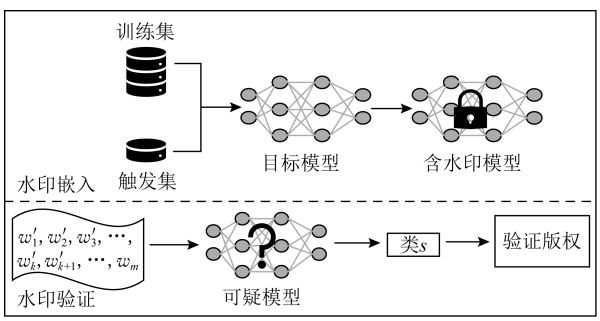


Fig. 22 The watermark embedding and watermark verification framework in ref [107]

图 22 文献[107]中的水印嵌入和水印验证框架

文本领域 DNN 模型水印的发展目前还处于萌芽期,无论是文本分类或是文本处理,相关的研究都还非常少,并且缺乏较成熟的水印方法.但是神经网络水印在图像领域的探索已经积累了许多宝贵的经验,为研究人员开拓了视野,可以在未来很好地指导文本领域和其他领域的神经网络水印研究.

3.3 灰盒水印方法

"灰盒"在文献[83]被提出,但是该文把灰盒等同于黑盒.根据文献[76,108]既通过向模型的内部嵌入信息,又以黑盒的方式获得输出以验证模型所有权的特点,在我们的工作中将其作为一种新类别的神经网络水印方法单独提出,并将其划分到灰盒水印类别中.

文献[76]提出了一个保护 DNN 模型知识产权的系统解决方案 DeepSigns,其架构如图 23 所示.该方案将目标模型和所有者特定的水印签名作为输入,然后在选定的层中嵌入相关的水印签名以及一组相应的密钥,输出一个带有水印信息的 DNN 模型.与先前的工作将水印信息直接嵌入 DNN 模型的静态内容(权重)中所不同的是,DeepSigns 是将任

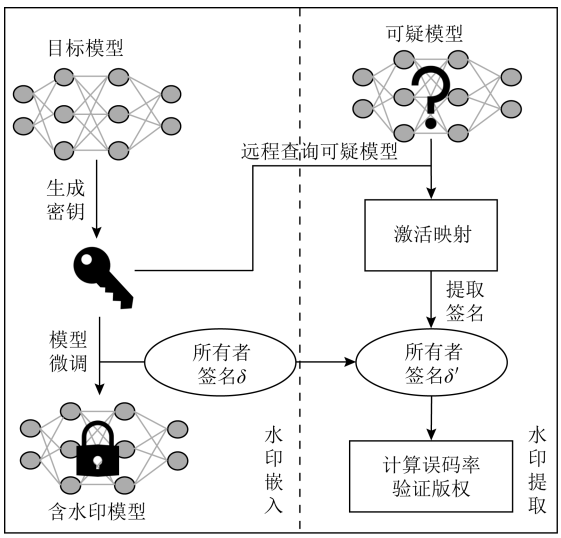


Fig. 23 DeepSigns architecture[76]

图 23 DeepSigns 架构示意图[76]

意 N 位($N\geqslant 1$)字符串嵌入各种激活图的概率密度函数(probability density function, PDF)中.为了验证远程 DNN 的知识产权,模型所有者首先需要使用在水印嵌入阶段生成的密钥查询远程的 DNN,并获得相应的激活映射.然后,DeepSigns 方法从获取的激活图的 PDF 分布中提取水印签名.最后,通过计算每层中提取的签名和对应的真实签名之间的误码率以验证版权.该方案由于水印嵌入动态统计信息中而不是模型的静态权重中,因此,能够抵抗水印覆盖攻击.此外,该方案从理论上讲允许通过增加触发器密钥的数量来嵌入一个任意的 N 比特流.

与文献[76]不同,为了解决将水印嵌入模型参数中容易被攻击者移除和检测的弊端,文献[108]基于神经架构搜索(neural architecture search, NAS)设计了一种新的 NAS 算法.该算法通过所有者特定的水印确定搜索空间中特定操作的某些连接,然后搜索其余的连接以产生高质量的网络架构,利用这种网络架构的独特性来代表模型的所有权.具体来说,在水印嵌入阶段,模型所有者生成唯一水印和对应的标记密钥,接着采用传统的 NAS 方法和标记密钥生成水印体系结构和验证密钥,在此基础上,利用该体系结构训练一个带水印的 DNN 模型.在水印提取和验证阶段,模型所有者使用侧通道技术在黑盒模式下提取 DNN 模型的体系结构来验证所有权,即使该模型是加密或隔离的.与之前的水印方法相比,该方案将水印嵌入体系结构中,而不是嵌入参数中.实验表明,该方案具有良好的有效性、可用性、鲁棒性和唯一性.但是该方案中水印是在模型的构建阶段嵌入的,若对训练好的 DNN 模型进行所有权保护则不适用.

3.4 无盒水印方法

无盒水印方法是指不再需要模型本身的参与,通过提取 DNN 模型输出中的水印即可验证模型版权.

文献[109]专注于深度学习文本生成模型,摒弃以往给训练数据或者模型参数加水印的方法,而是给模型的输出加水印.输入一个文本和一个充当水印信息的二进制字符串,该水印系统能够生成一个带有水印信息的输出.然后使用一个“揭示网络”(revealing network)即可从该输出中提取到水印信息,进而证明模型的所有权.为了改进语法错误问题,使用一个鉴别器来减少文本的语义损失,但仍无法完全解决该问题.

文献[109]是无盒水印方法在文本领域的应用,文献[110-111]的工作是无盒水印方法在图像处理领域的应用.

文献[110]提出了一种以输出图像作为结果的新颖数字水印框架,如图 24 所示.该框架适用于图像处理任务的 DNN 版权保护.由所提出的框架训练得到的 DNN 输出的任何图像必须包含一定的水印.宿主网络由生成器 G 和鉴别器 D 构成,鉴别器 D 是可选的,用于优化生成器 G.将要保护的宿主神经网络和水印提取网络 E 一起训练,训练过程中宿主网络的参数将根据总损失 $\mathcal{L}$ 进行更新,水印提取网络的参数只根据水印损失 $\mathcal{L}_V$ 进行更新.总损失 $\mathcal{L}$ 由任务损失 $\mathcal{L}_t$ 和水印损失 $\mathcal{L}_V$ 构成:

$$\mathcal{L}=\mathcal{L}_t+\theta\mathcal{L}_V, \tag{4}$$

其中,θ 是可调参数,默认 $\theta=1$.水印损失 $\mathcal{L}_V$ 由 3 部分构成:从含水印图像中提取出的水印和目标水印之间的差距 $\mathcal{L}_V^{(1)}$、从无水印图像中提取的随机噪声和含水印图像中提取的水印之间的差距 $\mathcal{L}_V^{(2)}$、提取水印时使用的密钥正确与否之间的损失 $\mathcal{L}_V^{(3)}$.

$$\mathcal{L}_V=\alpha\mathcal{L}_V^{(1)}+\beta\mathcal{L}_V^{(2)}+\gamma\mathcal{L}_V^{(3)}, \tag{5}$$

$$\mathcal{L}_V^{(1)}=\frac{1}{|S_1|}\sum_{G(x)\in S_1}\|E(G(x),k)-\boldsymbol{V}\|_p^p, \tag{6}$$

$$\mathcal{L}_V^{(2)}=\frac{1}{|S_2|}\sum_{x_i\in S_2}\|E(x_i,k)-\boldsymbol{V}_z\|_p^p, \tag{7}$$

$$\mathcal{L}_V^{(3)}=\frac{1}{|S_1|}\sum_{G(x)\in S_1,k_x\neq k}\|E(G(x),k_x)-\boldsymbol{V}_z\|_p^p, \tag{8}$$

其中,α,β 和 γ 是为了适应不同任务而设置的可调参数,$\boldsymbol{V}$ 表示目标水印,k 表示密钥,$\boldsymbol{V}_z$ 被设计为全零矩阵.任务损失 $\mathcal{L}_t$ 由输出图像和原始图像之间的像素值损失、鉴别器损失和感知损失 3 部分构成.

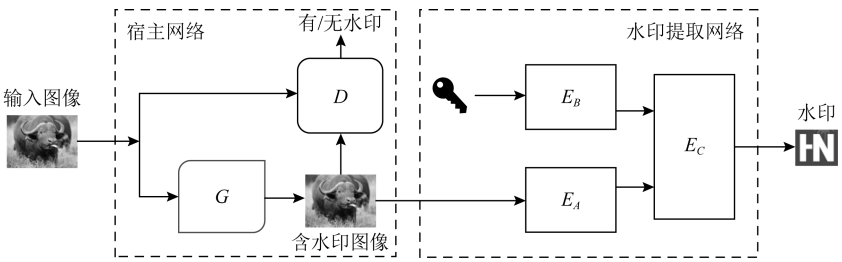


Fig. 24 The novel watermarking framework proposed in ref [110]

图 24 文献[110]提出的新颖水印框架

根据任务的不同,任务损失可以是其中之一,也可以是它们的组合.训练 2 个网络后,只释放宿主网络,水印提取网络则保密.训练后的神经网络可以在将水印嵌入输出图像中的同时完成原始任务.另外,该方案引入了水印提取密钥,密钥由版权所有者保管.只有提供正确的密钥,水印提取网络才进行工作.实验证明了该技术对各种图像处理任务都是有效且鲁棒的,包括图像着色、超分辨率、图像编辑、语义分割等.

文献[111]与文献[110]同为图像处理任务的模型保护方案,有异曲同工之妙.不同的是,文献[111]在保护图像处理任务模型的同时旨在抵抗利用模型输入输出对进行的代理模型攻击.

首先来看一组理论分析.如图 25 所示,若为图像集 B 中的每幅图像 $b_i(b_i\in B)$ 嵌入一个水印 δ,形成另一个图像集合 B',则攻击者训练的代理模型 SM 的目标就是最小化 $SM(a_i)$ 和 b_i' 之间的差距:

$$\mathcal{L}(SM(a_i),b_i')\to 0,\ b_i'=b_i+\delta. \tag{9}$$

具体来说,对于每个输入 a_i,都存在 $b_i=M(a_i)$.由于存在式(10)中的等价性,因此一定存在一个模型 SM 可以学习 A 和 B' 之间的映射关系.另一方面 DNN 的损失最小化特性从理论上保证了代理模型 SM 应该将水印 δ 学习到其输出中.当 $SM=M+\delta$ 时

$$\mathcal{L}(M(a_i),b_i)\to 0\Leftrightarrow\mathcal{L}(SM(a_i),b_i+\delta)\to 0. \tag{10}$$

文献[111]基于此理论,提出了一个通用的深度不可见模型水印框架,如图 26 所示,旨在抵抗代理模型攻击.该方案在待保护的 DNN 之后添加了一个水印嵌入网络 H,将水印嵌入到由待保护的目标 DNN 的输出集 B 中,并最终输出含水印的图像集 B',含水印的图像集 B' 与原始 DNN 的输出集 B 在视觉上保持一致.另外设计一个水印提取网络 R,可以从含水印的图像集 B' 中提取嵌入的水印.待保护的 DNN 和水印嵌入网络打包成一个整体进行部署,当攻击者基于此打包模型的输入输出对训练一个代理模型 SM 时,隐藏的水印将被学习到代理模型中,水印提取网络 R 仍然可以从代理模型的输出集 B'' 中提取水印,从而验证版权.

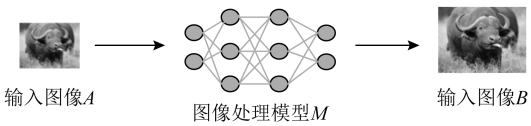


Fig. 25 Image processing model schematic

图 25 图像处理模型示意图

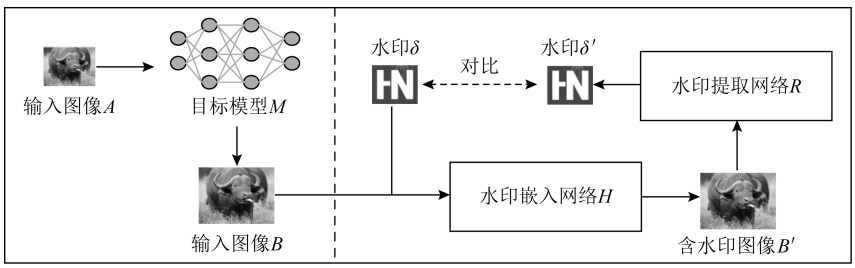


Fig. 26 The depth invisible model watermarking framework proposed in ref [111]

图 26 文献[111]提出的深度不可见模型水印框架

本节对 4 种神经网络水印方法的相关工作进行了详细的梳理和总结.白盒水印方法提取水印时需要了解模型的内部结构,导致其在实际应用中不太理想,但其对原始模型精度的影响相对于其他方法具有一定的优势.黑盒水印方法通过访问 API 即可进行验证的特点使其具有较为广泛的应用,但目前绝大部分黑盒水印方法的研究针对图像分类领域,在其他领域的应用还较少.灰盒水印方法在白盒和黑盒的基础上出现,相关工作相对较少,目前与以上 2 种方法相比不具有优势.无盒水印方法的出现,主要解决了图像处理领域 DNN 模型的版权保护问题,填补了这一块空白,为神经网络水印的研究打开了新的思路.

4 深度神经网络水印性能对比

与传统多媒体数字水印相似,DNN 模型水印需要具备一些性能指标,表 4 介绍了深度神经网络水印的性能指标,表 5 介绍了针对深度神经网络水印鲁棒性和安全性的 8 种攻击方法.

在目前的主流攻击方式中,水印研究者通常考虑模型微调和剪枝攻击、水印覆盖攻击和歧义攻击等方法.攻击者可能利用少量数据集对盗取的 DNN 模型进行微调,可能选取某些神经元进行失活处理,在这个过程中,模型所有者嵌入的水印可能被去除或难以提取和验证版权.如文献[112]提出了一种水印

去除框架 REFIT,通过适当地设计学习速率,可以利用微调方式去除水印;文献[113]证明,攻击者可以利用预训练模型来标注未标注的样本,并以此来增加微调训练数据,实现去除水印的目的.除此之外,攻击者可能嵌入新的非法水印,以此对原水印造成破坏或导致取证模糊.如文献[114]通过检测模型的统计分布来检测水印的存在,继而推导出水印的嵌入长度,并利用此信息通过覆盖来去除水印.另外,攻击者也可能通过为盗取 DNN 模型伪造额外的水印来让所有权验证产生歧义.如文献[81]表明,在这个过程中,攻击者伪造的水印检测率甚至可以达到 100%,且不需要原始数据,计算成本也很低,因此,目前也陆续出现针对抵抗水印歧义攻击的相关研究.

从表 4 和表 5 中可以看出,不同工作所运用的性能指标不尽相同,缺乏统一的评估标准.

第 3 节对神经网络水印的相关研究进行了梳理,不同学者提出的方案各具特色,已有方案的性能对比如表 6 所示.

Table 4 Performance Indicators of DNN Watermarking

表 4 深度神经网络水印性能指标

要求	描述
鲁棒性(robustness)	指当模型受到攻击(如模型微调、剪枝等)时,神经网络水印仍然存在且具备验证版权的能力
保真度(fidelity)	指嵌入水印后不会降低模型原始任务的精度
容量(capacity)	指能够在目标模型中嵌入足量的水印信息
可靠性(reliability)	指 DNN 模型的合法所有者能够以较高的概率检测到水印以验证版权
安全性(security)	指水印应该是保密的,不应该被未授权方访问、读取或修改
完整性(integrity)	指嵌入的水印应该能完全提取,使提取的差错尽可能少
不可伪造性(unforgeability)	指攻击者不能在 DNN 模型中嵌入额外的水印以声称其对模型的所有权
通用性(generality)	指水印方法能够应用于不同的 DNN 模型
效率(efficiency)	指水印的嵌入和提取过程快速,所带来的开销很小
唯一性(uniqueness)	指水印方法应该为每个用户嵌入唯一的水印,以分别区分不同的 DNN 模型
隐蔽性(stealthiness)	指嵌入 DNN 模型的水印不易被检测到

Table 5 DNN Watermarking Attack Method

表 5 深度神经网络水印攻击方法

方法	介绍
微调(fine-tuning)	通过对模型中的参数值进行更新,从而破坏嵌入的水印
剪枝(pruning)	通过将神经网络中部分神经元归零进而改变模型结构来破坏模型中的水印
规避攻击(evasion attacks)	指攻击者通过构建投票机制或检测器来躲避所有权的验证
共谋攻击(collusion attack)	指一组拥有相同宿主 DNN 但指纹不同的用户合作构建一个所有者无法检测到指纹的模型
水印覆盖(watermark overwriting)	指攻击者嵌入新的非法水印,以此对原水印造成破坏或导致取证模糊
代理模型攻击(surrogate model attack)	指攻击者通过访问目标模型得到输出,生成输入输出对(模型提取技术)来训练另一个类似的代理模型或者通过教师—学生网络的方法(即蒸馏)训练代理模型
查询修改攻击(query modification attack)	指攻击者主动检测某个查询,若该查询来自模型所有者的水印验证查询,则修改该查询,使水印验证过程失败
歧义攻击(ambiguity attack)	指攻击者伪造非法水印嵌入模型,以此声称对模型的所有权,造成模型所有权验证存在歧义

Table 6 Watermarking Performance Comparison of Different Schemes

表 6 不同方案的水印性能对比

文献	大类	小类	抗微调攻击(鲁棒性)	抗剪枝攻击(鲁棒性)	抗水印覆盖攻击(鲁棒性)	抗代理模型攻击(鲁棒性)	抗歧义攻击(安全性)	隐蔽性	保真度	容量
文献[73-75, 77-79]	白盒	基于内部权重	++	++	++		+	++	++	++
文献[80-82]	白盒	基于内部结构	++	++			++		++	++
文献[83]	白盒	基于组合验证	++	++	+		++	++	++	++

续表 6

文献	大类	小类	抗微调攻击(鲁棒性)	抗剪枝攻击(鲁棒性)	抗水印覆盖攻击(鲁棒性)	抗代理模型攻击(鲁棒性)	抗歧义攻击(安全性)	隐蔽性	保真度	容量
文献[84-86]	黑盒	通过标签更改的分类任务	+	+	+		+		++	++
文献[87-89]	黑盒	通过在原始样本中嵌入信息和标签更改的分类任务	++	++			+	+	++	++
文献[90-91]	黑盒	通过添加新的类的分类任务	++	++	++		++	++	++	
文献[92-93]	黑盒	通过添加嵌入信息的附加样本的分类任务	++	++	++			++	++	++
文献[94-97]	黑盒	其他	++	++		++	++		++	
文献[106]	黑盒	基于图像处理任务	++	++	++				++	++
文献[107]	黑盒	基于文本处理任务		++				+	++	
文献[76,108]	灰盒	基于概率密度函数和内部结构	++	++	++			++	++	++
文献[109]	无盒	基于文本处理任务						++		
文献[110-111]	无盒	基于图像处理任务						++	++	++

注:“++”表示性能较好,“+”表示性能一般.

5 展 望

相对于传统多媒体水印几十年的发展历程来说,神经网络水印的发展仍处于起步阶段,未来还有很长的路要走.在未来的 DNN 模型版权保护研究过程中,有 6 个方面可以努力.

1) 提升嵌入速度.目前,神经网络水印的嵌入方式以训练和微调嵌入为主,这需要耗费较大的代价和工作量.与多媒体数字水印的嵌入速度相比,神经网络水印的嵌入速度要慢得多,未来可以在如何提升神经网络水印的嵌入速度方面做出努力.

2) 神经网络水印的通用性.目前,大部分神经网络水印是针对图像处理和分类任务的神经网络模型来设计的,针对其他任务类型(如对音频处理的模型、视觉对象分割和声源分离的综合神经网络模型),现有的水印方法是否还可以适用,又或者有没有新的方法来承担此类模型的保护工作,这有待进一步研究.

3) 理论证明.众所周知,正是由于在图像等多媒体数据中存在冗余,所以才可以在图像中嵌入水印.DNN 模型也是如此,存在许多冗余的神经元.但是在冗余的模型中嵌入多少水印,也就是说模型的冗余与嵌入水印的量之间存在怎样的数学关系,又或者说,在模型的不同阶段和不同位置进行嵌入与嵌入水印的量之间存在怎样的关系,目前还缺乏相应的理论证明.未来如何在理论上解决神经网络水印嵌入量的一些问题值得我们思考.

4) 主动防护.神经网络水印所做的工作是在模型被盗后再来证明所有者的版权,属于被动防护.那么如何主动解决模型被盗问题,增加模型被盗的难度和代价,是一个有趣的问题.目前针对 DNN 模型进行主动防护的研究还较少,未来可以根据实际商业版权管理需求在此方向继续做出努力.

5) 完整性保护.版权保护只是 DNN 模型安全的一个方面,DNN 模型还面临被篡改的风险,如以模型性能下降为攻击目标的恶意微调和权重参数修改,以及通过后门植入对 DNN 模型造成破坏等一系列威胁完整性的安全隐患.目前,文献[115]提出了一种用于 DNN 完整性保护的白盒水印方法 NeuNAC,文献[116]提出了一种易于检测篡改的黑盒脆弱水印方法,可以在不暴露模型内在机制(包括结构和参数)的黑盒条件下进行篡改检测.文献[115]和文献[116]的 2 种方案均适用于 DNN 模型的篡改检测.可以看到,相关研究还很匮乏.因此,探索更多解决 DNN 模型完整性保护的方案可以作为下一步努力的方向.

6) 评估标准.目前的 DNN 模型水印算法的评估指标参差不齐,只有个别指标(如对模型微调和剪枝攻击的鲁棒性)被大多数研究者所通用,缺乏公认统一的评估框架和指标.如何对选用不同宿主模型、不同的嵌入方式的水印算法形成统一的评估指标在未来也是一个可研究的方向.

6 结 语

DNN等新兴技术以前所未有的性能在工业互联网安全中得到了广泛发展和应用,然而构建产品级DNN模型并非易事,需要花费大量的人力和物力资源.因此,DNN模型的知识产权保护问题逐渐引起了学术界和工业界的广泛关注,并涌现出大量优秀的解决方案.我们在以往白盒和黑盒水印的基础上,从水印的嵌入和提取的方式出发,将神经网络水印扩充为白盒、黑盒、灰盒和无盒水印4种类别,并对现有水印方法进行了深入分析和探讨.同时,对于水印的性能指标以及针对它们的攻击方式也进行了梳理.最后指出了神经网络水印未来的研究方向,旨在对该领域未来的发展有所帮助.

作者贡献声明:樊雪峰负责文献调研、内容设计、论文撰写和最后版本修订;周晓谊负责提出论文整体研究思路、全文框架设计和最终审核;朱冰冰负责论文插图设计和修订;董津位负责部分文献调研和撰写以及全文修订;牛俊、王鹤负责调研分析和全文修订.

参 考 文 献

[1] Hitaj D, Mancini L V. Have you stolen my model? Evasion attacks against deep neural network watermarking techniques [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1809.00615

[2] Zhao Jian, Koch E. Embedding robust labels into images for copyright protection [C] //Proc of the Congress on Intellectual Property Rights for Specialized Information, Knowledge and New Technologies. Vienna: Citeseer, 1995: 242-251

[3] Nikolaidis N, Pitas I. Copyright protection of images using robust digital signatures [C] //Proc of the IEEE Int Conf on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 1996: 2168-2171

[4] Lee C H, Lee Y K. An adaptive digital image watermarking technique for copyright protection [J]. IEEE Transactions on Consumer Electronics, 1999, 45(4): 1005-1015

[5] Zafeiriou S, Tefas A, Pitas I. Blind robust watermarking schemes for copyright protection of 3D mesh objects [J]. IEEE Transactions on Visualization and Computer Graphics, 2005, 11(5): 596-607

[6] Lou D C, Tso H K, Liu J L. A copyright protection scheme for digital images using visual cryptography technique [J]. Computer Standards & Interfaces, 2007, 29(1): 125-131

[7] Fang Han, Chen Dongdong, Huang Qidong, et al. Deep template-based watermarking [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(4): 1436-1451

[8] Kapse A, Belokar S, Gorde Y, et al. Digital image security using digital watermarking [J]. International Research Journal of Engineering and Technology, 2018, 5(3): 163-166

[9] Prajwalasimha S, Sowmyashree A, Suraksha B, et al. Logarithmic transform based digital watermarking scheme [C] //Proc of the Int Conf on ISMAC in Computational Vision and Bio-Engineering. Berlin: Springer, 2018: 9-16

[10] Alejandra M O, Claudia F U, Rogelio H B, et al. A survey on reversible watermarking for multimedia content: A robustness overview [J]. IEEE Access, 2019, 7: 132662-132681

[11] Podilchuk C I, Delp E J. Digital watermarking: Algorithms and applications [J]. IEEE Signal Processing Magazine, 2001, 18(4): 33-46

[12] Busch C, Funk W, Wolthusen S. Digital watermarking: From concepts to real-time video applications [J]. IEEE Computer Graphics and Applications, 1999, 19(1): 25-35

[13] Yin Hao, Lin Chuang, Qiu Feng, et al. A survey of digital watermarking [J]. Journal of Computer Research and Development, 2005, 42(7): 1093-1099 (in Chinese)
(尹浩, 林闯, 邱锋, 等. 数字水印技术综述[J]. 计算机研究与发展, 2005, 42(7): 1093-1099)

[14] Chen Yixin, Hu Xi, Xiao Feng. Digital watermarking hiding technology for copyright information [C] //Proc of the Int Conf on Data Processing Techniques and Applications for Cyber-Physical Systems. Berlin: Springer, 2020: 1203-1209

[15] Sahu A K, Swain G. An optimal information hiding approach based on pixel value differencing and modulus function [J]. Wireless Personal Communications, 2019, 108(1): 159-174

[16] Maniriho P, Ahmad T. Information hiding scheme for digital images using difference expansion and modulus function [J]. Journal of King Saud University-Computer and Information Sciences, 2019, 31(3): 335-347

[17] Petitcolas F, Anderson R J, Kuhn M G. Information hiding—A survey [J]. Proceedings of the IEEE, 1999, 87(7): 1062-1078

[18] Parashar P, Singh R K. A survey: Digital image watermarking techniques [J]. International Journal of Signal Processing, Image Processing and Pattern Recognition, 2014, 7(6): 111-124

[19] Mohanarathinam A, Kamalraj S, Venkatesan G, et al. Digital watermarking techniques for image security: A review [J]. Journal of Ambient Intelligence and Humanized Computing, 2019, 11(8): 1-9

[20] Potdar V M, Han Song, Chang E. A survey of digital image watermarking techniques [C] //Proc of the 3rd IEEE Int Conf on Industrial Informatics. Piscataway, NJ: IEEE, 2005: 709-716

[21] Asikuzzaman M, Pickering M R. An overview of digital video watermarking [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(9): 2131-2153

[22] Artru R, Gouaillard A, Ebrahimi T. Digital watermarking of video streams: Review of the state-of-the-art [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1908.02039

[23] Luo Yifan, Peng Dezhong. A robust digital watermarking method for depth-image-based rendering 3D video [J]. Multimedia Tools and Applications, 2021, 80(10): 14915-14939

[24] Korzhik V, Alekseev V, Morales-Luna G. Design of audio digital watermarking system resistant to removal attack [C] //Proc of the Federated Conf on Computer Science and Information Systems (FedCSIS). Piscataway, NJ: IEEE, 2017: 647-652

[25] Bhat V, Sengupta I, Das A. An adaptive audio watermarking based on the singular value decomposition in the wavelet domain [J]. Digital Signal Processing, 2010, 20(6): 1547-1558

[26] Singha A, Ullah M A. Audio watermarking with multiple images as watermarks [J]. IETE Journal of Education, 2020, 61(2): 64-75

[27] Wang Yunhan, Yan Zining, Kou Liang. Research and implementation of text digital watermarking based on file filter driver [C] //Proc of the 2017 6th Int Conf on Measurement, Instrumentation and Automation (ICMIA 2017). Paris, France: Atlantis Press, 2017: 174-183

[28] Jalil Z, Mirza A M. A review of digital watermarking techniques for text documents [C] //Proc of the 2009 Int Conf on Information and Multimedia Technology. Piscataway, NJ: IEEE, 2009: 230-234

[29] Khadim U, Iqbal M M, Azam M A. An intelligent three-level digital watermarking method for document protection [J]. Mehran University Research Journal of Engineering & Technology, 2021, 40(2): 323-334

[30] Jang H U, Choi H Y, Son J, et al. Cropping-resilient 3D mesh watermarking based on consistent segmentation and mesh steganalysis [J]. Multimedia Tools and Applications, 2018, 77(5): 5685-5712

[31] Hou J U, Kim D G, Lee H K. Blind 3D mesh watermarking for 3D printed model by analyzing layering artifact [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2712-2725

[32] Savakar D, Ghuli A. Non-blind digital watermarking with enhanced image embedding capacity using DMeyer wavelet decomposition, SVD, and DFT [J]. Pattern Recognition and Image Analysis, 2017, 27(3): 511-517

[33] Rasti P, Samiei S, Agoyi M, et al. Robust non-blind color video watermarking using QR decomposition and entropy analysis [J]. Journal of Visual Communication and Image Representation, 2016, 38(3): 838-847

[34] Ernawan F, Kabir M N. A blind watermarking technique using redundant wavelet transform for copyright protection [C] //Proc of the 2018 IEEE 14th Int Colloquium on Signal Processing & Its Applications (CSPA). Piscataway, NJ: IEEE, 2018: 221-226

[35] Thanki R, Kothari A, Trivedi D. Hybrid and blind watermarking scheme in DCuT—RDWT domain [J]. Journal of Information Security and Applications, 2019, 46: 231-249

[36] Liu Feng, Han Ke, Wang Changzheng. A novel blind watermark algorithm based on SVD and DCT [C] //Proc of the 2009 IEEE Int Conf on Intelligent Computing and Intelligent Systems. Piscataway, NJ: IEEE, 2009: 283-286

[37] Ye Xueyi, Chen Xueting, Deng Meng, et al. A SIFT-based DWT-SVD blind watermark method against geometrical attacks [C] //Proc of the 7th Int Congress on Image and Signal Processing. Piscataway, NJ: IEEE, 2014: 323-329

[38] Abdulrahman A K, Ozturk S. A novel hybrid DCT and DWT based robust watermarking algorithm for color images [J]. Multimedia Tools and Applications, 2019, 78(12): 17027-17049

[39] Zeebaree D Q. Robust watermarking scheme based LWT and SVD using artificial bee colony optimization [J]. Indonesian Journal of Electrical Engineering and Computer Science, 2021, 21(2): 1218-1229

[40] Bravo-Solorio S, Calderon F, Li C T, et al. Fast fragile watermark embedding and iterative mechanism with high self-restoration performance [J]. Digital Signal Processing, 2018, 73: 83-92

[41] Gong Xinhui, Chen Lei, Yu Feng, et al. A secure image authentication scheme based on dual fragile watermark [J]. Multimedia Tools and Applications, 2020, 79(25): 18071-18088

[42] Sikder I, Dhar P K, Shimamura T. A semi-fragile watermarking method using slant transform and LU decomposition for image authentication [C] //Proc of the 2017 Int Conf on Electrical, Computer and Communication Engineering (ECCE). Piscataway, NJ: IEEE, 2017: 881-885

[43] Preda R O. Semi-fragile watermarking for image authentication with sensitive tamper localization in the wavelet domain [J]. Measurement, 2013, 46(1): 367-373

[44] Su Qingtang, Yuan Zihan, Liu Decheng. An approximate schur decomposition-based spatial domain color image watermarking method [J]. IEEE Access, 2018, 7:4358-4370

[45] Kumar S, Singh B K. Entropy based spatial domain image watermarking and its performance analysis [J]. Multimedia Tools and Applications, 2021, 80(6): 9315-9331

[46] Parah S A, Sheikh J A, Assad U I, et al. Realisation and robustness evaluation of a blind spatial domain watermarking technique [J]. International Journal of Electronics, 2017, 104(4): 659-672

[47] Zhang Liming, Yan Haowen, Zhu Rui, et al. Combinational spatial and frequency domains watermarking for 2D vector maps [J]. Multimedia Tools and Applications, 2020, 79(41): 31375-31387

[48] Al-Ardhi S, Thayananthan V, Basuhail A. A new vector map watermarking technique in frequency domain based on LCA-transform [J]. Multimedia Tools and Applications, 2020, 79(43): 32361-32387

[49] Novamizanti L, Budiman G, Astuti E. Robust audio watermarking based on transform domain and SVD with compressive sampling framework [J]. TELKOMNIKA Telecommunication, Computing, Electronics and Control, 2020, 18(2): 1079-1088

[50] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1508.07909

[51] Wu Yonghui, Schuster M, Chen Zhifeng, et al. Google's neural machine translation system: Bridging the gap between human and machine translation [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1609.08144

[52] Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing [J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75

[53] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778

[54] Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision [J]. NPJ Digital Medicine, 2021, 4(1): 1-9

[55] Voulodimos A, Doulamis N, Doulamis A, et al. Deep learning for computer vision: A brief review [J]. Computational Intelligence and Neuroscience, 2018, 2018: 1-13

[56] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105

[57] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1-9

[58] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation [C] //Proc of the Int Conf on Medical Image Computing and Computer-assisted Intervention. Berlin: Springer, 2015: 234-241

[59] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster r-CNN: Towards real-time object detection with region proposal networks [J]. Advances in Neural Information Processing Systems, 2015, 28: 91-99

[60] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 779-788

[61] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2980-2988

[62] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep speech 2: End-to-end speech recognition in english and mandarin [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 173-182

[63] Rebai I, Benayed Y, Mahdi W, et al. Improving speech recognition using data augmentation and acoustic model fusion [J]. Procedia Computer Science, 2017, 112:316-322

[64] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[65] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1409.1556

[66] You Haoran, Li Chaojian, Xu Pengfei, et al. Drawing early-bird tickets: Towards more efficient training of deep networks [EB/OL]. [2021-11-12]. https://arxiv. org/abs/1909.11957

[67] Fu Yonggan, Guo Han, Li Meng, et al. CPT: Efficient deep neural network training via cyclic precision [EB/OL]. [2021-11-12]. https://arxiv.org/abs/2101.09868

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

[69] Xie Chenqi, Zhang Baowen, Yi Ping. Survey on artificial intelligence model watermarking [J]. Computer Science, 2021, 48(7): 9-16 (in Chinese)
(谢宸琪, 张保稳, 易平. 人工智能模型水印研究综述[J]. 计算机科学, 2021, 48(7): 9-16)

[70] Xue Mingfu, Wang Jian, Liu Weiqiang. DNN intellectual property protection: Taxonomy, attacks and evaluations [C] //Proc of the 2021 on Great Lakes Symp on VLSI. New York: ACM, 2021: 455-460

[71] Li Yue, Wang Hongxia, Barni M. A survey of deep neural network watermarking techniques [EB/OL]. [2021-11-12]. https://arxiv.org/abs/2103.09274

[72] CSIG Technical Committee on Digital Media Forensics and Security. ChinaMFS 2020 [EB/OL]. [2021-11-12]. https://mp. weixin. qq. com/s/nNNHLQxznTlAuYYDM5IA4A (in Chinese)
(CSIG 数字取证与安全专委会. ChinaMFS 2020 [EB/OL]. [2021-11-12]. https://mp. weixin. qq. com/s/nNNHLQxznTlAuYYDM5IA4A)

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

[74] Chen Huili, Rohani B, Koushanfar F. Deepmarks: A digital fingerprinting framework for deep neural networks [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1804.03648

[75] Wang Jiangfeng, Wu Hanzhou, Zhang Xinpeng, et al. Watermarking in deep neural networks via error back-propagation [J]. Electronic Imaging, 2020, 2020(4): 1-22

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

[77] Feng Le, Zhang Xinpeng. Watermarking neural network with compensation mechanism [C] //Proc of the Int Conf on Knowledge Science, Engineering and Management. Berlin: Springer, 2020: 363-375

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

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

[80] Zhao Xiangyu, Yao Yinzhe, Wu Hanzhou, et al. Structural watermarking to deep neural networks via network channel pruning [EB/OL]. [2021-11-12]. https://arxiv.org/abs/2107.08688

[81] Fan Lixin, Ng W K, Chan C S. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks [C] //Proc of the 33rd Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2019: 1-10

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

[83] Lü Peizhuo, Li Pan, Zhang Shengzhi, et al. HufuNet: Embedding the left piece as watermark and keeping the right piece for ownership verification in deep neural networks [EB/OL]. [2021-11-12]. https://arxiv.org/abs/2103.13628

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

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

[86] Merrer E L, Perez P, Trédan G. Adversarial frontier stitching for remote neural network watermarking [J]. Neural Computing and Applications, 2020, 32(13): 9233-9244

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

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

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

[90] Zhong Qi, Zhang Leo Yu, Zhang Jun, et al. Protecting IP of deep neural networks with watermarking: A new label helps [J]. Advances in Knowledge Discovery and Data Mining, 2020, 12085:462-474

[91] Li Meng, Zhong Qi, Zhang Leo Yu, et al. Protecting the intellectual property of deep neural networks with watermarking: The frequency domain approach [C] //Proc of the IEEE 19th Int Conf on Trust, Security and Privacy in Computing and Communications (TrustCom). Piscataway, NJ: IEEE, 2020: 402-409

[92] Jebreel N M, Domingo-Ferrer J, Sánchez D, et al. KeyNet: An asymmetric key-style framework for watermarking deep learning models [J/OL]. Applied Sciences, 2021 [2021-11-12]. https://www.mdpi. com/2076-3417/11/3/999

[93] Sun Shichang, Xue Mingfu, Wang Jian, et al. Protecting the intellectual properties of deep neural networks with an additional class and steganographic images [EB/OL]. [2021-11-12]. https://arxiv.org/abs/2104.09203

[94] Szyller S, Atli B G, Marchal S, et al. Dawn: Dynamic adversarial watermarking of neural networks [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1906.00830

[95] Zhu Renjie, Zhang Xinpeng, Shi Mengte, et al. Secure neural network watermarking protocol against forging attack [J]. EURASIP Journal on Image and Video Processing, 2020, 2020(1): 1-12

[96] Li Huiying, Wenger E, Shan S, et al. Piracy resistant watermarks for deep neural networks [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1910.01226

[97] Maung M A P, Kiya H. Piracy-resistant DNN watermarking by block-wise image transformation with secret key [C] //Proc of the 2021 ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2021: 159-164

[98] Aprilpyone M, Kiya H. Block-wise image transformation with secret key for adversarially robust defense [J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 2709-2723

[99] Mao Xiaojiao, Shen Chunhua, Yang Yubin. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections [J]. Advances in Neural Information Processing Systems, 2016, 29: 2802-2810

[100] Zhang Kai, Zuo Wangmeng, Chen Yunjin, et al. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising [J]. IEEE Transactions on Image Processing, 2017, 26(7): 3142-3155

[101] Dian Renwei, Li Shutao, Guo Anjing, et al. Deep hyperspectral image sharpening [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(11): 5345-5355

[102] Fan Qingnan, Yang Jiaolong, Hua Gang, et al. A generic deep architecture for single image reflection removal and image smoothing [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 3238-3247

[103] Dong Chao, Loy C , He Kaiming, et al. Learning a deep convolutional network for image super-resolution [C] //Proc of the European Conf on Computer Vision. Berlin: Springer, 2014: 184-199

[104] Xie Junyuan, Xu Linli, Chen Enhong. Image denoising and inpainting with deep neural networks [C] //Proc of the Advances in Neural Information Processing Systems. New York: ACM, 2012: 341-349

[105] Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1508.06576

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

[107] Yadollahi M M, Shoeleh F, Dadkhah S, et al. Robust black-box watermarking for deep neural network using inverse document frequency [EB/OL]. [2021-11-12]. https://arxiv.org/abs/2103.0559 0

[108] Lou Xiaoxuan, Guo Shangwei, Zhang Tianwei, et al. When NAS meets watermarking: Ownership verification of DNN models via cache side channels [EB/OL]. [2021-11-12]. https://arxiv.org/abs /2102.03523

[109] Abdelnabi S, Fritz M. Adversarial watermarking transformer: Towards tracing text provenance with data hiding [C] //Proc of the 2021 IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2021: 121-140

[110] Wu Hanzhou, Liu Gen, Yao Yuwei, et al. Watermarking neural networks with watermarked images [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(7): 2591-2601

[111] Zhang Jie, Chen Dongdong, Liao Jing, et al. Deep model intellectual property protection via deep watermarking [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021 [2021-11-12]. https://ieeexploreieee.53yu.com/abstract/docum ent/9373945

[112] Chen Xinyun, Wang Wenxiao, Bender C, et al. Refit: A unified watermark removal framework for deep learning systems with limited data [C] //Proc of the 2021 ACM Asia Conf on Computer and Communications Security. New York: ACM, 2021: 321-335

[113] Chen Xinyun, Wang Wenxiao, Ding Yiming, et al. Leveraging unlabeled data for watermark removal of deep neural networks [C/OL] //Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019 [2021-11-12]. https://ruoxijia.info/wp-content/uploads/2020/2003/watermark_removal_icml2019_workshop.pdf

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

[115] 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

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

Fan Xuefeng, born in 1998. Master candidate. His main research interests include AI security and neural network watermarking.

樊雪峰,1998 年生.硕士研究生.主要研究方向为人工智能安全和神经网络水印.

Zhou Xiaoyi, born in 1979. PhD, associate professor, master supervisor. Her main research interests include image hiding, image cryptosystem, and optimization and application of artificial intelligence algorithms.

周晓谊,1979 年生.博士,副教授,硕士生导师.主要研究方向为图像隐藏、图像加密系统和人工智能算法优化及应用.

Zhu Bingbing, born in 1998. Master candidate. His main research interests include information hiding and image digital watermarking.

朱冰冰,1998 年生.硕士研究生.主要研究方向为信息隐藏和图像数字水印.

Dong Jinwei, born in 1997. Master candidate. His main research interests include AI security and mobile security.

董津位,1997 年生.硕士研究生.主要研究方向为人工智能安全和移动安全.

Niu Jun, born in 1992. PhD candidate. Her main research interests include information security and machine learning security.

牛　俊,1992 年生.博士研究生.主要研究方向为信息安全和机器学习安全.

Wang He, born in 1987. PhD, lecturer. Her main research interests include applied cryptography and quantum communication protocol.

王　鹤,1987 年生.博士,讲师.主要研究方向为应用密码学和量子通信协议.