

深度学习模型的版权保护研究综述

王馨雅, 华光, 江昊, 张海剑

(武汉大学电子信息学院, 湖北 武汉 430072)

摘要: 随着深度学习技术的迅猛发展, 深度学习模型在图像分类、语音识别等领域得到了广泛应用。训练深度学习模型依赖大量的数据和算力, 成本高昂, 因此, 出售已训练好的模型或者提供特定的服务(如DLaaS)成为一种商业模式。然而, 如果模型遭到恶意用户窃取, 则可能会对模型训练者的商业利益造成损害。此外, 网络拓扑结构设计和参数训练的过程包含着模型训练者的智慧结晶, 因此一个训练完备的模型应属于模型开发者的知识产权从而得到保护。近年来, 深度神经网络水印成为一个新兴的研究课题, 研究者将多媒体内容保护的方法引入深度学习模型保护领域, 试图在深度神经网络模型中嵌入水印以验证模型的所有权。目前已有大量方法被提出, 但缺乏梳理和概括。对神经网络水印领域已有的方法进行了梳理和总结, 并探讨了该领域未来的研究方向。给出神经网络水印的模型框架, 并介绍了分类模型、模型后门等基础概念。按照水印嵌入的机制将已有的方法分类为两类: 一是嵌入网络内部, 以网络内部信息作为载体; 二是建立网络后门, 将后门特殊映射关系作为水印。分别对基于这两种思想的深度神经网络水印方法进行了全面的阐述和总结, 讨论了各方法的特点、优势和局限性, 同时介绍并讨论了相应的水印攻击方法。通过分析水印中的白盒与黑盒场景可知, 白盒分发的模型难以得到有效保护, 而黑盒分发和黑盒验证场景下的神经网络水印防攻值得进一步的研究。

关键词: 神经网络安全; 神经网络版权保护; 黑盒水印; 白盒水印; 后门水印

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022015

Survey on intellectual property protection for deep learning model

WANG Xinya, HUA Guang, JIANG Hao, ZHANG Haijian

School of Electronic Information, Wuhan University, Wuhan 430072, China

Abstract: With the rapid development of deep learning technology, deep learning models have been widely used in many fields such as image classification and speech recognition. Training a deep learning model relies on a large

收稿日期: 2021-06-11; 修回日期: 2022-01-05

通信作者: 华光, ghua@whu.edu.cn

基金项目: 国家自然科学基金(61802284); 国家自然科学基金企业创新发展联合基金重点项目(U19B2004); 广东省科技专项基金(2019SDR002)

Foundation Items: The National Natural Science Foundation of China (61802284), The National Natural Science Foundation of China Enterprise Innovation Development Key Project (U19B2004), The Special Fund for Science and Technology of Guangdong Province (2019SDR002)

引用格式: 王馨雅, 华光, 江昊, 等. 深度学习模型的版权保护研究综述[J]. 网络与信息安全学报, 2022, 8(2): 1-14.

Citation Format: WANG X Y, HUA G, JIANG H, et al. survey on intellectual property protection for deep learning model[J]. Chinese Journal of Network and Information Security, 2022, 8(2): 1-14.

amount of data and computing power, thus selling the trained model or providing specific services (DLaaS, e.g.) has become a new business. However, the commercial interests of model trainers and the intellectual property rights of model developers may be violated if the model is maliciously stolen. With deep neural network watermarking becoming a new research topic, multimedia copyright protection techniques were used for deep learning model protection. Numerous methods have been proposed in this field and then a comprehensive survey is needed. the existing deep neural network watermarking methods were elaborated and summarized and the future research directions of this field were discussed. The overall framework of neural network watermarking was presented, whereby the basic concepts such as classification model and model backdoor were introduced. Secondly, the existing methods were divided into two types according to the mechanism of watermark embedding, one is to embed the watermark bits into the carrier of internal information of the network, and the other one uses the established backdoor mapping as the watermark. These two existing deep neural network watermarking methods were analyzed and summarized, and attacks to the watermarks were also introduced and discussed. By analyzing the white-box and black-box conditions in watermarking scenario, it comes to the conclusion that the model is difficult to be effectively protected when it is distributed in the white-box manner, and the neural network watermark defenses in the black-box distribution and black-box verification are both worthy for further research.

Keywords: neural network security, copy right protection of neural networks, black-box watermarking, white-box watermarking, backdoor watermarking

0 引言

随着计算资源和大数据的普及和发展,深度学习技术在社会多领域取得了巨大成功,为社会产业升级和科学技术发展提供了强劲驱动力。然而,成功地训练一个深度学习模型通常需要付出巨大的人力物力代价。首先,训练过程依赖大量准确标注的优质数据。大数据的获取本身并非易事,而对数据实现有效的清洗和标注则是更加烦琐的工作。例如,著名的 ImageNet^[1]图像分类数据库以众包的模式借助世界范围用户的力量进行人工标注,耗时 3 年才得以成型。其次,为了训练得到有效的神经网络模型,往往需要分配大量的计算资源以调整网络的拓扑结构以及超参数,而目前较先进的深度神经网络的权重数量日趋庞大,可轻易超过亿数量级。例如,用于自然语言处理的 GTP-3 模型^[2],其参数数量已达到 1 750 亿。因此,经过精心训练的高性能神经网络模型理应被视作数据和模型拥有者的劳动成果,应该拥有专属的知识产权。

除优质的标注数据和充足的计算资源以外,训练深度神经网络模型还需要一定的专业知识,

并非是每个人都能够完成的任务。在这种情况下,销售训练好的模型已经成为一种商业模式。例如,IBM 公司提出了深度学习服务概念(DLaaS, deep learning as a service)^[3],利用常用的 TensorFlow 和 PyTorch 等深度学习框架在 IBM Cloud 上为用户部署深度学习任务,降低了深度学习门槛。此外,亚马逊、阿里云提供了相应 DLaaS 的 API 供普通用户使用。深度学习的迅猛发展以及 DLaaS 的逐渐普及也带来了一些安全隐患^[4-5]。例如,模型的购买者将购买的模型复制或篡改后对模型进行二次分发,或盗取模型后声称自己对模型的所有权,都会对模型所有者的知识产权和经济利益造成损害。因此,需要一种对神经网络模型版权保护的框架,以验证模型所有者对模型的所有权,进而保护模型所有者的合法权益。

近年来,深度学习模型的版权保护问题逐渐受到世界各国的关注。2017 年 7 月,我国印发《新一代人工智能发展规划》,强调要建立人工智能技术标准和知识产权体系。2018 年 11 月,欧洲专利局发布了人工智能和机器学习的专利性指南。由此可见,针对深度学习模型的版权保护已成为一项重要研究课题。

基于优质大数据和精密的网络结构训练好的深度神经网络属于其拥有者的知识产权,具备数字产品的特征。基于此,学者将多媒体内容版权保护的数字水印方法^[6-7]引入深度学习领域,即在训练好的深度神经网络模型中嵌入水印。2017年,Uchida等^[8]首次提出了神经网络水印概念,并提出了一种将水印嵌入网络损失函数的正则项中的方法。随后,学术界展开了针对深度学习模型水印的广泛研究。

总体来说,保护深度神经网络模型的方法与保护一般数字产品的方法相似,均借助密码学(cryptography)^[9]或水印^[10]技术手段。使用密码学的主要方式为对模型的重要数据进行加密(encryption),并仅对得到授权的用户分发密钥(key)。这种方法的主要局限在于无法控制授权用户在解密(decryption)模型后的行为。而水印技术可以很好地弥补其局限性,对侵权行为实施有效追溯。此外,针对深度学习中的生成式(generative)和判别式(discriminative)两大类模型,神经网络水印的总体嵌入思想有所不同。本文在总结所有方法的同时,重点讨论判别式模型中最常见的多分类器模型的数字水印保护方法。

目前,用于保护深度网络分类器的水印方法按照水印的嵌入机制可分为两类:基于网络内部信息的方法和基于后门的方法。基于网络内部信息的方法直接将水印嵌入目标模型的内部结构中,包括将水印嵌入权重、激活层的输出以及在网络中添加新的层作为水印等。基于后门的方法主要针对图像分类的任务,通过在深度学习模型中嵌入后门,引入特殊的输入输出关系。当需要验证模型所有权时,模型的所有者可将作为后门嵌入的特殊样本输入网络来获得预先设置好的特殊标签。由于模型的后门只有模型的所有者知晓,可通过展示这些异常的输入输出关系来证明自己对模型的所有权。

1 神经网络水印框架及方法分类

本节在介绍神经网络水印的基本框架以及相关概念后,进一步对现有相关研究进行分类。考

虑到现有多数研究工作是针对判别模型,本节以基于监督学习的深度神经网络分类器为主要研究模型,同时简单介绍其他模型。

1.1 性能指标

在嵌入水印时,需权衡水印质量和模型本身功能之间的平衡,而嵌入水印的深度神经网络模型在分发后可能经历无意或恶意攻击,在水印验证阶段需要进行水印提取,由此可知,对于神经网络水印存在如下性能指标。

1) 功能不变性:水印嵌入后原始模型的性能应尽可能不受影响。该指标可同时用来评估水印系统和对水印系统的攻击,即嵌入水印和针对水印的攻击均不能以牺牲模型功能为代价,否则嵌入和攻击将失去意义。

2) 鲁棒性:嵌入的水印要能抵御可能存在的攻击,如对模型的微调、压缩或二次训练,在受到攻击后也能可靠提取;对于基于后门机制嵌入的水印则应有能力抵御逃逸攻击和歧义攻击。

3) 嵌入容量:对于修改网络内部信息的神经网络水印方法,可嵌入的最大信息比特数。

4) 安全性:嵌入神经网络内部的水印信息或建立的后门映射关系不被攻击者获取的性能,主要通过嵌入密钥实现。

5) 计算复杂度:神经网络水印嵌入和验证所需要的计算复杂度,可分开评估。如基于后门的水印在验证阶段的计算复杂度较低,仅需对比推断结果。

1.2 神经网络水印方法分类

现有的神经网络模型版权保护方法分类如图1所示。本节首先给出判别式模型中典型的基于监督学习的分类模型的宏观水印框架,再简要介绍针对生成式模型和基于密码学保护深度神经网络模型的方法。

1.2.1 分类模型

深度神经网络分类模型的训练目标为建立样本空间 X 到标签空间 Y 的映射。假设训练集由 N 对样本-标签对构成,记为 (x_i, y_i) ,其中 $i \in \{0, 1, \dots, N-1\}$, $x_i \in X$, $y_i \in Y$,对模型来说,正确且有效的输入输出的关系用 $f: X \rightarrow Y$ 表示。如果用 $\mathcal{M}$ 表示模型结构,用 $\mathcal{W}$ 表示模型权重,则模型

的训练过程如下。

$$\hat{\mathcal{W}} = \arg \min_{\mathcal{W}} \text{Loss}\{\mathcal{M}[\mathcal{W}, X], Y\} \quad (1)$$

其中, $\hat{\mathcal{W}}$ 为训练后的权重。对于全体训练集, 当正确映射关系比例大于 $(1-\varepsilon)|X|$ 时, 判定模型训练成功, 其中 ε 为一个很小的正实数。

$$\mathcal{M}[\hat{\mathcal{W}}, x_i] = y_i \quad (2)$$

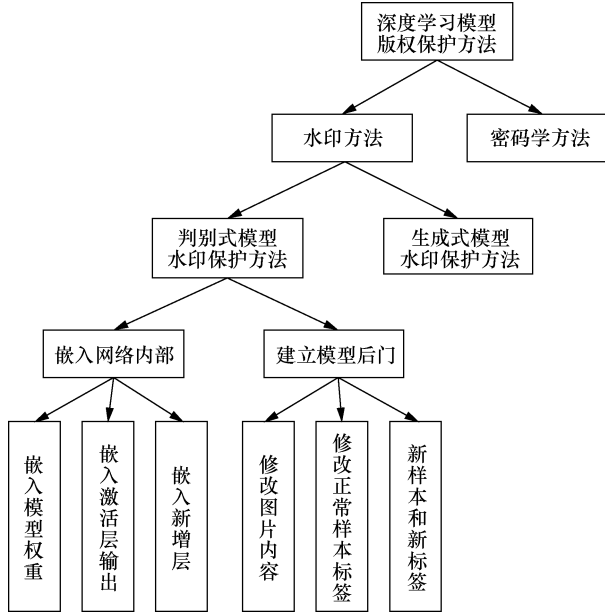


图 1 神经网络模型版权保护方法分类

Figure 1 Classification of methods for neural network intellectual property protection

1.2.2 判别式模型保护方法

嵌入网络内部: 通过修改已训练好网络的内部信息实现水印的嵌入, 嵌入过程可描述为

$$\tilde{\mathcal{M}}[\tilde{\mathcal{W}}, \cdot] = \arg \min_{\tilde{\mathcal{W}}} \text{Loss}\{\mathcal{M}[\hat{\mathcal{W}}, X], Y\} + \lambda \text{Loss}\{\text{Mark}\} \quad (3)$$

其中, λ 权衡损失的参数, 损失函数左边部分保证模型的功能不变性, 右边部分保证水印的嵌入, **Mark** 为含密钥的特殊映射关系, 因方法不同, 设计较为灵活。水印验证时, 利用预设的密钥以及嵌入的 **Mark** 建立函数关系, 当映射正确率满足一定比例时认为验证成功。

建立网络后门: 通过在训练集中加入触发集训练, 模型学习到预先设定特殊映射关系从而建立后门^[11-13]。用角标 $\{\cdot\}_T$ 表示触发集, 则触发样本表示为 $x_{T,i}$, 其对应的标签为 $y_{T,i}$, 嵌入了后门

的权重可表示为

$$\tilde{\mathcal{W}} = \arg \min_{\mathcal{W}} \text{Loss}\{\mathcal{M}[\mathcal{W}, X], Y\} + \lambda \text{Loss}\{\mathcal{M}[\mathcal{W}, X_T], Y_T\}, \quad (4)$$

其中, 损失函数左边部分保证模型的功能不变性, 右边部分保证后门的建立。水印验证时, 给定由 $(x_{T,i}, y_{T,i})$ 组成的密钥, 当模型推断结果的正确比例大于 $(1-\varepsilon)|X_T|$ 时, 判定验证成功。

$$\mathcal{M}[\tilde{\mathcal{W}}, x_{T,i}] = y_{T,i} \quad (5)$$

值得注意的是, 以上两类嵌入方法既可基于已训练好的网络进行微调实现, 也可采取重训练的方式同时得到正常样本和触发样本的映射关系。此外, 以上方法均常用于图像识别的分类任务, 即输入图片得到推断的类标签。

1.2.3 生成式模型保护方法

判别式模型主要学习特征与数据之间的后验关系, 而生成式模型主要学习数据与特征的联合分布关系或数据本身的分布情况。因此生成式模型, 如生成对抗网络 (GAN, generative adversarial network) 或变分自编码器 (VAE, variational autoencoder) 可产生新的数据, 比仅输出后验分数的判别式模型更丰富。在此情况下, 攻击者可利用模型生成大量数据用于训练自己的网络, 因此生成式模型产生的新数据也需要得到保护。

在计算机视觉领域, 生成式模型一般用于各种图像处理任务, 它们的输出不再是分类标签, 而是经过处理之后的图像。针对这一类模型, Zhang 等^[14]最先提出在输出图片的空间域嵌入水印的保护方法。后来将这种方法延伸到模型具有多个发行版本时嵌入多个水印的问题中, 同时提出了一种直接在网络训练过程中完成水印嵌入的方案^[15]。

但是这种方式嵌入的水印和网络之间没有直接的关联, 如果攻击者获知模型的内部信息, 则可以嵌入自己的水印, 或者通过 PS 等图像处理技术去除图片中的水印。针对这些局限, Wu 等^[16]设计了一个改进的系统, 让被保护的网络和水印提取网络一起参与训练, 验证阶段需要使用这个

水印提取网络才能够提取出水印。

1.2.4 基于密码学的方法

除了利用水印进行版权保护，还有学者利用密码学的方法对深度学习模型进行版权保护，这种方法使用混沌加密算法给模型权重的排列位置加密，只有经过权值解密的模型才能够正常地进行推断，因此窃取一个经过加密的模型的意义将大幅减弱，攻击者需要重新训练^[9]。

1.3 白盒与黑盒场景

现有研究将深度网络模型的版权保护系统按照在验证阶段是否需要访问网络内部参数分成了白盒方法和黑盒方法。白盒方法以 Uchida 等^[8]提出的水印验证框架为代表，水印的验证需要访问模型权重。基于后门机制的水印系统则是

典型的黑盒方法，这一类方法只需要通过输入输出关系就可以验证模型所有权，不需要访问模型内部参数或结构。然而，在设计模型版权保护方案时，情形更为复杂，不仅要考虑水印验证所需要的条件，还应该考虑到模型以不同方式分发的情形下，模型水印对各种潜在攻击的鲁棒性。

2 神经网络分类模型水印

本节重点介绍图 1 中针对判别式模型的嵌入网络内部和建立网络后门两类水印嵌入方法。以基于卷积神经网络（CNN，convolutional neural network）的深度学习分类器为例，图 2 展示了这些方法将水印嵌入网络的情形。

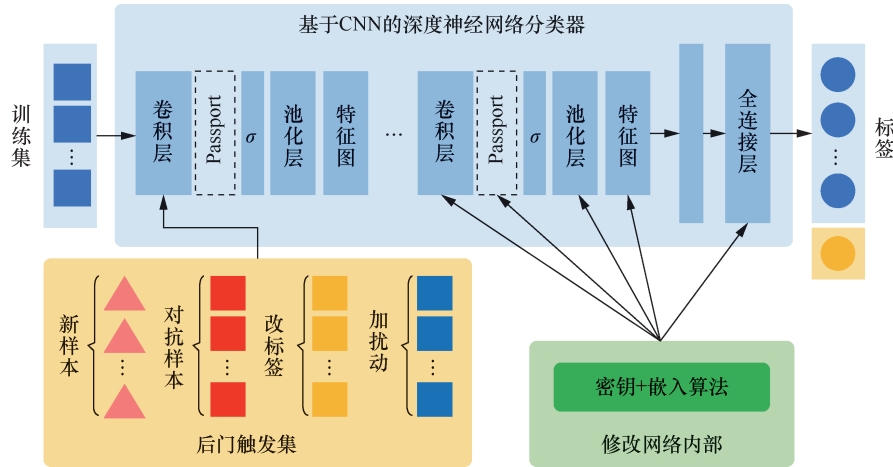


图2 深度学习分类模型水印嵌入方法

Figure 2 Demonstration of watermark embedding methods for DNN classifier

2.1 嵌入网络内部

(1) 嵌入模型权重

嵌入权重的方法试图将二进制的水印比特信息 B 通过密钥 K 嵌入部分网络权重中，对应图 2 中右下角绿色方框的部分。记选取的部分网络权重为 $\mathcal{V}$ ， $\mathcal{V} \subset \mathcal{W}$ ，则嵌入的过程通过训练可得到新的模型参数 $\tilde{\mathcal{W}}$ 。

$$\tilde{\mathcal{W}} = \arg \min_{\mathcal{W}} \text{Loss} \left\{ \mathcal{M}[\hat{\mathcal{W}}, X], Y \right\} + \lambda \text{Loss}_{\mathcal{V}} \{ K \cdot \mathcal{V}, B \} \quad (6)$$

其中， $\{\}$ 表示内积。式(6)是式(3)的一种具体情况。验证水印时，首先根据嵌入权重的位置选取部分权重，因网络权重可能遭到攻击，故记为 $\mathcal{V}'$ ，然后提供密钥 K 和嵌入信息 B ，当对比

结果满足

$$\text{sgn} \left\{ \text{sigmoid}(K \cdot \mathcal{V}') \right\} = B \quad (7)$$

或达到一定比例对比正确时，则判定水印验证成功，其中 $\text{sgn}\{\}$ 为符号函数。接下来，介绍基于式(6)、式(7)的一些神经网络水印方法。

Uchida 等^[8]提出了第一个嵌入网络内部的水印框架，通过正则化参数的方式将水印比特信息嵌入某一个中间层的权重分布中，这项工作是对深度学习嵌入水印的初次尝试，揭示了向模型嵌入水印的潜力。但是这种方法很容易受到重写攻击，因为网络的层数是有限的，攻击者可以对每一层进行重训练，这样原始水印就会被破坏。另外，这种通过正则化参数嵌入水印的方法可能

干扰模型的正常训练。

Chen 等^[17]同样提出将水印嵌入模型权重,为了追踪模型的使用情况,在模型分发时给每一个用户提供一个向量编码。这些向量编码通过模型所有者秘密生成的投影矩阵 $\mathbf{X}$,在训练时作为水印嵌入模型中。验证时用户根据自己的编码提取出相应的权重,然后模型所有者将其与 $\mathbf{X}$ 相乘得到预先嵌入模型的水印。

Rouhani^[18]提出将水印嵌入网络的动态部分——激活层的概率分布中。在嵌入阶段,水印与网络一起训练,输出嵌入了水印的模型和存储了水印位置信息的密钥 WM keys。在验证阶段,需要使用这个密钥才能够触发模型得到嵌入了水印的激活层的概率分布,进而提取出嵌入的水印签名。由于将水印嵌入了网络的动态部分,与嵌入权重的方法相比,可嵌入的水印容量更大,而且这种方法同时依赖于数据和模型,在实际应用中更加灵活且不易被检测。

Kuribayashi 等^[19]将水印嵌入全连接层的权重中,通过量化索引调制来控制嵌入水印在模型中引入的变化量大小,运用随机排列和抖动调制等技术提高水印的保密性。

Feng 等^[20]提出了一种具有补偿机制的模型微调方案。他设置了两个密钥:密钥 K_0 作为伪随机算法的初始值生成 n 个权重嵌入的位置,密钥 K_1 则作为另一个伪随机生成器的初始值,这个伪随机生成器用来产生特定的噪声模式,以对经过正交变换的水印进行调制。密钥 K_0 的存在给攻击者实施重写攻击设置了障碍,因为攻击者无法获知嵌入了水印的权重的位置。针对水印进行扩频调制则让水印分散地嵌入权重中,增强了水印的鲁棒性。

将水印嵌入网络参数的方法很容易受到微调攻击,因此 Tartaglione^[21]提出另一种策略:让嵌入了水印的权重不参与网络训练时的参数更新。通过调整训练时的损失函数,使带有水印信息的部分参数在网络训练前后保持不变,这种方法对于微调攻击有较强的鲁棒性。

(2) 嵌入新增层

上文所述各方法均保留了原有的模型结构,仅将水印嵌入模型参数中,而 Fan 等^[22]采用了另

一种思路,即在网络中增加新的层以实现水印的嵌入。该方法在模型的每个卷积层后面添加一个名为 Passport 的层(如图 2 神经网络中的绿色虚线框所示),该层被专门用来嵌入水印信息 B ,如果将 Passport 层的权重表示为 $\mathcal{W}_p$, $f(\mathcal{W}_p)$ 表示将该层权重映射到二进制的结果,嵌入水印后的权重可表示为

$$\tilde{\mathcal{W}} = \arg \min_{\mathcal{W}} \text{Loss}\{\mathcal{M}[\mathcal{W}, X], Y\} + \lambda \text{Loss}_p\{f(\mathcal{W}_p), B\} \quad (8)$$

同样,验证时若有 $\mathcal{M}[\hat{\mathcal{W}}, \mathcal{W}_p] = B' = B$, 则判定验证成功。此外,本文工作存在多种模型和 Passport 分发方式。第一种是直接将添加了 Passport 层训练得到的模型分发给用户,用户使用该模型不仅可以获得原始任务的输出,还可以提取出相应的 Passport 层,以此证明他们是合法的用户;第二种方案分别训练原始网络和添加了 Passport 层的模型,只把原始模型分发给用户,而另一个嵌入了水印的网络仅用于所有权的验证,并不公开给外界;第三种方案在前一种的基础上加入了触发集,在分发时同样只分发原始网络,保留添加了 Passport 层和后门的模型,相比第二种方案增加了一种验证方式。

(3) 嵌入网络的输出

将水印嵌入模型参数中容易受到模型提取攻击,因此有学者提出将水印嵌入模型的输出中。Sebastian^[23]将水印嵌入 API 响应的一个子集中,提出一种动态的对抗水印方法,动态地改变一小部分样本查询的结果,而对分类准确率的影响微乎其微。针对机器翻译任务的模型, Venugopal 等^[24]提出一种使用哈希函数生成定长序列的方法来在其输出中嵌入水印以区分机器翻译和人工翻译的结果。

(4) 其他

Lou 等^[25]利用神经网络架构搜索将水印嵌入网络结构中。He 等^[26]选择一部分网络的权重和一个单层感知机进行参数共享,将感知机的输入作为密钥,感知机的输出即水印信息。

还有学者将神经网络作为水印嵌入。Lyu 等^[27]将神经网络本身作为水印来嵌入,提出一种作为水印使用的神经网络 HufuNet,将其一半的卷积

核嵌入目标模型参数中，另一半则保留，用于所有权验证。

2.2 建立网络后门

(1) 在原有训练集图片上添加扰动

一种典型的触发样本的构造方式是选取原始训练集中的部分图片，在图片上应用某种特定的干扰模式，并随机分配一个原始训练集中的标签（在图2中用蓝色方框表示）。如果用 K_{pattern} 来表示对图片内容的干扰模式， $\text{per}(\cdot)$ 表示扰动算法，则有

$$\begin{aligned} x_{T,i} &= \text{Per}(x_i, K_{\text{pattern}}), Y_T \subset Y, \\ x_{T,i} &\mapsto y_{T,i} \neq f(x) \end{aligned} \quad (9)$$

这种干扰模式可以是有意义的字符串或者 Logo，或者某种特定模式的噪声^[28]。但这些经过干扰的样本往往在特征分布上与正常样本有较大差异，在水印验证阶段容易被攻击者所识别。攻击者可以先让查询样本经过一个后门检测器，如果检测器认为这是一个查询样本，则拒绝输出或随机输出一个标签，从而逃逸后门的验证^[4]。为了提高鲁棒性，Li 等^[29]使用自编码器产生不可见的 Logo，并将生成的盲水印嵌入原始图片作为触发集，触发集图片在视觉上和正常样本没有区别，且和正常样本的特征分布一致，因此能够更好地抵御上述逃逸攻击。有学者在频域嵌入不可见的 Logo 来生成触发集^[30]，频域的水印更加具有隐蔽性，对各种信号处理方法也更加鲁棒。为了降低负阳率，Guo 等^[31]采用遗传进化算法来决定后门依赖的触发模式。

(2) 以图片实例作为后门

还有一类后门方法并不以某种特定的触发模式作为触发集的构造方式，而采用图片实例作为后门（对应图2中红色三角表示的触发集样本）：

$$\begin{aligned} X_T \cap X &= \emptyset, Y_T \subset Y, \\ x_{T,i} &\mapsto y_{T,i} \neq f(x) \end{aligned} \quad (10)$$

Zhang 等^[28]将原始训练集中的一部分图片替换为与此任务不相关的图片，Yossi 等^[32]则使用了一组从互联网上下载的图片来构建触发集。此外，与现有大多数水印系统需要可信第三方完成水印验证的工作方式不同，他在后门水印设计过程中引入了 commitment 机制，同时约束了模型所

有者和攻击者，不需要引入可信第三方即可完成水印验证。

(3) 添加新的类别标签

虽然以上两种后门水印方法均未改变模型的结构，但 Zhong 等^[33]认为在图片上叠加扰动的方法在网络中引入了错误的映射关系，改变了分类网络的决策边界，进而影响了模型的准确率和鲁棒性。而添加新的类别则不会对原来的分类边界产生干扰。因此他在原始的标签空间中增加一类，让所有的后门图片属于这个新的类别（在图2中用黄色矩形表示）。

$$\begin{aligned} X_T &\subset X, Y_T \cap Y = \emptyset, \\ x_{T,i} &\mapsto y_{T,i} \neq f(x) \end{aligned} \quad (11)$$

新类别的触发集对应的样本为对原始训练样本添加 Logo 标签后的一组新样本。

(4) 使用对抗样本

对抗样本是通过在原始的图片中加入细微的干扰，能够让模型以高置信度输出一个错误分类的输入^[34-35]。Merrer 等^[36]采用对抗训练的方式将对抗样本作为后门嵌入网络。他选择了一部分对抗样本作为触发样本，给这些选定的对抗样本分配分类正确的标签，再让模型在这些样本上微调。在微调过程中，模型在这些触发样本附近的决策边界会发生改变。一个经过上述微调的模型，如果再次受到触发集中对抗样本的触发，将输出正确的分类结果。在这种情况下，水印的嵌入方式不变，但验证方式与上述后门方案相反。该方法的触发集和映射机制可表示为

$$\begin{aligned} x_{T,i} &= \mathcal{A}(x_i), \\ x_{T,i} &\mapsto y_{T,i} = f(x) \end{aligned} \quad (12)$$

其中， $\mathcal{A}(x_i)$ 表示对抗攻击，但对抗攻击后的样本仍为正确分类。

还有一些学者依据对抗样本的特性来进行模型的版权验证。Lukas^[37]和 Zhao 等^[38]利用了对抗样本的迁移性，即对抗样本在相同或者相似的模型上往往具有较高的迁移性，因此在代理模型上会表现出高迁移性，通过将对抗样本在可疑模型上的攻击成功率与某个阈值进行比较，从而判断可疑模型是否是根据原始模型得到的代理模型。

Chen 等^[39]提出了第一个多比特的基于后门的水印方案,他使用一种模型相关的编码方案,将作者签名以二进制编码的形式嵌入模型的预测结果中,这种方案使用定向生成的对抗样本作为后门触发集,并将对抗样本和相应的分类置信度分数作为模型指纹分发给合法用户(此类触发集在图 2 中用红色矩形表示)。

(5) 其他

为了让后门方案同时与模型和用户相关,Zhu 等^[40]采用这样的方式构造触发集:使用两个不予公开的哈希函数分别生成作为触发集的图像

链和其对应的标签。单向哈希函数无法反向构造,因此在验证阶段攻击者难以实施伪造攻击。

还有一种方法采取了类似后门的方式,但是不需要特定的触发集^[41]。这种方法采取一种可学习的图像加密算法,将部分干净图像经过加密变换之后作为触发集,在验证时需要用密钥将图片进行处理之后才能得到正确的推断。

文献[42]在干净样本和触发集图像上都通过隐写技术嵌入了 Logo,将水印与所有者身份信息联系起来,便于区别于其他的企业、产品或服务。

深度模型水印方法分类如表 1 所示。

表 1 深度模型水印方法分类
Table 1 The classification of DNN watermarking methods

类别	水印方法	验证场景	零/多比特	鲁棒性
嵌入内部	将水印以正则化的形式嵌入网络权重中 ^[8]	白盒	多比特	能应对剪枝、微调
	给合法用户发布实现嵌入网络的向量编码,用户只有用编码才能提取出权重 ^[17]	白盒	多比特	能应对协同攻击、剪枝、微调
	将水印嵌入模型中间层/输出层特征输出的概率密度函数中 ^[18]	白盒/黑盒	多比特	能应对压缩、微调和重写攻击
	具有补偿机制的微调,通过两个密钥指定嵌入水印的权重位置和覆盖的噪声模式 ^[20]	白盒	多比特	能应对重写攻击
	让嵌入水印的网络权重不参与网络训练 ^[21]	白盒	多比特	能应对微调
建立后门	在网络结构上插入 passport 层 ^[22]	白盒/黑盒	多比特	需 passport 使用,单独通过签名验证
	在原图上覆盖 Logo、噪声以及使用不相关图片作为触发集的后门构造方法 ^[28]	黑盒	零比特	能应对剪枝、微调和蒸馏
	使用编码器生成带有盲水印的图片作为后门触发集 ^[29]	黑盒	零比特	能应对剪枝、微调
	在频域生成不可见水印作为后门触发集 ^[30]	黑盒	零比特	能应对剪枝、微调
	选择一组抽象图片作为后门,引入密码学协议提供安全性 ^[32]	黑盒	零比特	能应对剪枝、微调、蒸馏,不能应对歧义攻击
	在原始分类任务中增添一个新类别,将图片映射到这个类别作为后门 ^[33]	黑盒	零比特	能应对微调、逃逸攻击
	通过对抗训练让网络对于部分对抗样本分类正确,细微地调整决策边界 ^[36]	黑盒	零比特	能应对剪枝、微调、奇异值分解攻击
	基于遗传进化算法生成和优化触发集 ^[31]	黑盒	零比特	降低了负阳率,能应对微调

3 攻击方法

针对现有的模型保护水印框架,有很多可能的攻击策略。目前的攻击方式主要有移除攻击、逃逸攻击和歧义攻击。移除攻击是指通过微调、剪枝或压缩等方式去除模型中原有的水印;逃逸攻击是指攻击者在黑盒的验证阶段通过一些手段逃脱水印的验证;歧义攻击则是让模型中出现另外一个非法水印,以混淆对于水印真实性的判断。目前,对于后门水印方案的攻击主要利用了神经

网络的如下特性。

遗忘性:神经网络在训练时,不可避免地会记住用于训练的数据,如果想要网络遗忘这些数据,只需要删除这些数据再进行重新训练^[43-44]。同样,对于嵌入了后门的网络,可以通过加入大量的新样本对网络进行微调,让网络忘记嵌入的后门,达到去除后门的目的。

不可解释性:深度学习的模型一直被视作黑盒子,虽然学者一直在探索模型可解释性的方法,但这项任务依旧面临挑战^[45-47]。当模型泛化表现

较差时, 欠拟合或者过拟合都会导致模型真正学习到的特征与重要性得分失去匹配。因此, 有了使用对抗样本作为后门和利用样本空间局限性来实施攻击的方式。

过参数性: 神经网络的过参数性也是导致对抗样本存在的重要原因之一。例如, 对抗样本实现了仅仅修改训练图片的一个或几个像素, 让网络输出和正常分类不一样的结果^[48]。攻击者可能会定向生成模型的对抗样本, 并把它当作触发图片, 在水印验证阶段实施歧义攻击。

样本空间局限性: 对于对数据重度依赖的深度神经网络模型, 训练样本空间总体是有限的。但对于攻击者来说, 样本空间却是无限的, 攻击者总可以找到一种在原始样本空间之外的样本, 或者在原始样本中选取一些样本并分配和模型无关的标签^[49]。样本空间示意如图3所示。

该思路同样为歧义攻击提供了可能。下面详细介绍一些攻击策略对应的具体研究工作。

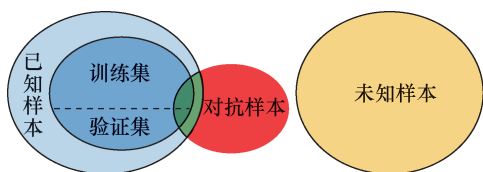


图3 样本空间示意

Figure 3 Example space of known, unknown, and adversarial examples

3.1 水印检测

水印检测的目标是检测出模型中是否存在水印, 以及存在什么形式的水印。在检测出系统中存在的水印之后, 攻击者可以决定具体采取何种攻击策略。Wang 等^[50]指出在文献[8]的方法中由于水印嵌入改变了模型的权重分布, 因此很容易被检测。他同时展示了一种基于属性推理的通用白盒水印检测方法^[51]。Shafieinejad 等^[52]也展示了一种属性推理攻击, 使用部分训练数据和从网络中提取的特征向量, 有效地检测出模型中是否存在基于后门机制嵌入的水印。他又提出一种对水印检测攻击更鲁棒的水印方案: 采用一种对抗训练的方式同时训练目标模型和水印检测网络以获得水印隐蔽性度量, 并以正则化参数的形式使之参与水印的嵌入^[53]。

3.2 移除攻击

最常见的对于神经网络模型水印保护系统的

攻击方式是设法去除模型中嵌入的水印。移除攻击可以通过微调^[54-55]、剪枝^[55-57]或者蒸馏^[58]的方式实现。也有一些学者同时融合了微调和剪枝的策略, 先对神经元进行剪枝, 然后对模型进行微调^[55]。如果攻击者知道水印的具体位置, 还可以通过重新初始化其所在层的参数并重训练来去除该层的水印^[8]。

目前大多数的移除攻击针对后门水印的移除^[52,54,59-61]。文献[52]给出对后门水印实施攻击的3种方式, 指出在白盒场景时攻击者可以通过正则化算法结合微调的方式来移除水印。文献^[54]通过合理地设置初始学习率和学习率的衰减参数, 成功地通过微调去除了基于后门机制嵌入的水印。虽然这种微调能够成功去除水印, 但是需要一定量标注的训练数据。而现实场景中攻击者往往缺乏足够的有标记训练数据, 因此文献^[59]进一步提出使用未标记数据的微调方法。具体做法是使用预训练模型的预测结果对从互联网上下载的未标记数据进行标记, 并用这些数据对模型进行微调。

对于嵌入网络内部的水印方法, 文献[50]针对 Uchida 等提出的水印方法, 使用一种水印移除算法去除网络原始嵌入的水印, 并同时嵌入了新的水印。

3.3 逃逸攻击

Ryota 等^[57]提出可以通过修改查询样本的方式实施逃逸攻击。这种方法适用于在原图上添加扰动而产生的后门样本。如果系统判定一个查询样本为后门样本, 则使用一个自编码器去除图片上覆盖的干扰信息, 让触发样本重新变回一个正常样本。

此外, Hitaj 等^[4]提出了两种基于盗取模型的逃逸攻击方法, 分别为集成攻击(ensemble attack)和检测攻击(detector attack), 均可在黑盒模式下实施, 无须访问模型内部信息。攻击者将盗取的模型建立为 DLaaS 系统, 通过 API 提供服务谋取利益。集成攻击是将多个盗取的模型组成集群, 对于输入的分类任务通过对多个模型的推断结果采取投票机制输出结果, 可有效扰乱触发样本的映射结果。而检测攻击通过引入触发样本检测机制实施逃逸攻击。在输入 API 之前先让样本经过检测器, 当检测到触发样本时, 系统刻意输出随机的标签以扰乱后门触发机制, 而对

于检测到的正常样本，系统使用盗取的模型进行正常的推断。

3.4 歧义攻击

歧义攻击的目标是让模型中出现另外的非法水印，破坏水印的唯一性。基于深度学习模型的过参数化特性，一个模型中可以同时存在多个后门。攻击者可以使用对抗样本作为额外的后门，或者通过微调的方式在模型中嵌入新的后门。Guo 等^[62]指出将作者签名作为后门嵌入模型的系统可能受到伪签名的攻击，如对于 Yossi 等^[32]提

出的方案，攻击者可以通过遗传算法生成另一组抽象图片，在模型中形成一组新的特殊映射。

深度模型水印攻击方法分类如表 2 所示。值得注意的是，基于额外训练的移除攻击需要使用少量正常样本（远少于原始训练所需样本数量）对模型进行白盒下的水印移除。对于替代模型，需要在黑盒条件下得到足够使替代模型收敛的标记数据从而实现替代攻击。而逃逸和歧义攻击在黑盒和白盒条件下均不需要使用正常样本即可完成攻击。

表 2 深度模型水印攻击方法分类
Table 2 The classification of DNN watermarking attack methods

攻击类别	手段	攻击条件	可攻击的水印方法
移除攻击	重训练 ^[52] 、微调 ^[54,59,63] 、重写 ^[50] 、蒸馏 ^[58]	少量数据，白盒	嵌入网络内部的白盒方法和基于后门的黑盒方法
	根据已知网络的输入输出训练一个替代模型 ^[52]	足量数据，黑盒	基于后门的黑盒方法
逃逸攻击	去除触发集上覆盖的噪声 ^[57]	黑盒	在图片上覆盖某种触发模式的后门水印方法
	集成多个功能相同的模型并基于投票机制推断 ^[4]	黑盒	针对同一功能模型的水印方法
	基于后门水印检测器逃避触发集的触发 ^[4]	黑盒	在图片上覆盖某种干扰模式的后门方法
歧义攻击	在模型内部嵌入另一个水印	白盒	将水印嵌入网络内部参数的白盒方法
	在模型中植入另外的后门（如利用对抗样本）	白盒	基于后门机制的黑盒方法

4 讨论

在深度学习模型的版权保护问题中，模型的分发方式不同，会导致攻击者采用不同的攻击策略，因此在设计模型版权保护方案时，应该将这些因素考虑在其中。从时间先后顺序考虑，首先，对于模型所有者，在水印嵌入时应只考虑白盒场景，即模型所有者可利用与模型有关的任意信息（权重、结构等）辅助水印嵌入。此后，当攻击者获得分发的模型后，根据攻击者获得的与模型相关知识的多少可分为白盒与黑盒情况，前者假设攻击者已知的模型信息和模型所有者相同，后者假设攻击者仅能观测给定样本和其输出的关系。最后，在模型验证时，存在白盒与黑盒的情况，前者需要访问模型内部信息，而后者只需观测输入输出关系。下面分别讨论各种可能的情形以及模型所有者和攻击者可能的应对方式。

4.1 白盒分发

该情况意味着攻击者可以获知模型的结构和参数，在攻击时可以采用更灵活的方法。通常情况下攻击者由于缺乏独立训练模型的能力而对模

型进行窃取，或对水印保护框架实施攻击，以达到非法使用模型，或提供类似服务来牟取利益的目的。当模型以白盒形式分发时，攻击者需要在攻击代价（对模型进行微调，重训练）和训练自己的模型中做出权衡。只有当攻击者只拥有有限的训练数据，或不具备足够的算力重新训练模型时，现有的水印方法才具备鲁棒性。

基于 Kerckhoffs' Principle，只有在模型结构和水印算法同时保密的情况下，模型才可能是绝对安全的。从更严格的模型安全角度考虑，应假定攻击者拥有和模型所有者相同的训练条件（足够的训练数据和算力）。在这种条件下，攻击者既可以采取和模型所有者相同或不同的水印嵌入方式制造歧义攻击，也可采用微调、剪枝等手段对水印进行破坏或者移除^[54-55]。在这种考虑下，不管采用何种验证方式，现有的基于水印的版权验证框架对于以白盒形式分发的模型来说都是无效的。因此，不论黑盒验证还是白盒验证，以白盒模式分发的模型均难以得到有效保护，而黑盒分发与黑盒验证模式是最有可能在实际应用中采用的模式，值得更多研究关注。

4.2 黑盒分发

在实际应用中,更常见的情形是各大公司通过聘请 AI 专家设计模型,花费大量资金和人力训练数据,最终得到一个模型文件。这些公司对外提供机器学习服务的方式主要有两种:一种是提供云模式的 API,另一种是将模型私有地部署到用户的设备或者数据中心的服务器上^[5]。对于前者,攻击者通过一定的遍历算法,在反复调用云模式的 API 后可以在本地还原出一个与原始模型功能相同的模型,并提供相似的服务;对于后者,攻击者通过逆向等安全技术可以将模型还原供其使用或者对模型二次售卖。总体来说,只有黑盒分发情况下的神经网络模型才能得到有效保护。

4.2.1 黑盒验证

有两类方法可以适用于黑盒场景下的水印验证:一类是基于后门机制的模型水印方法;另一类是 Rouhani 等^[18]所提出的水印方案。

当模型以黑盒发放时,这两类方法都可以达到验证模型版权的目的。对于攻击者来说,由于无法访问模型的内部参数与结构,也就无法对模型进行微调或者重训练,这给实施移除攻击和歧义攻击造成了困难。在这种情况下,模型所有者可设计如 Zhu 等^[40]的水印嵌入方法,构建难以复制的后门机制,使攻击者无法有效采取歧义攻击。而攻击者可能会实施逃逸攻击、基于后门的歧义攻击,或者盗取多个功能相似的模型进行集成^[4]。

4.2.2 白盒验证

当模型以黑盒发放时,水印方法对于各种攻击方法的鲁棒性都增强了,然而,当攻击者将窃取的模式以 API 接口(黑盒)的形式部署到线上时,白盒验证将无法实施。此外,白盒验证方法难以用于嵌入式系统^[62]。因此,白盒验证方案都具有应用的局限性。

5 结束语

本文对近年来兴起的用于保护深度神经网络版权的神经网络水印研究进行了梳理和介绍,对现有方法进行了分类和讨论,其中重点阐述了针对判别式模型中常见的多分类神经网络版权保护的水印嵌入和验证方法,包括与传统多媒体水印类似的嵌入模型内部的方法和神经网络特有的基

于后门的方法,对各方法的特点和优劣进行了详细介绍;进一步讨论和总结了现有针对神经网络水印的一系列攻击方法,从水印嵌入、攻击以及验证 3 个阶段分别讨论了白盒与黑盒场景下模型所有者和攻击者面临的挑战和可能采取的策略。

神经网络水印目前仍是一个较新的研究领域,还存在很多值得研究的问题。该领域未来的研究方向可以聚焦于如下问题。

(1) 提升后门水印针对歧义攻击的鲁棒性。白盒验证的方法无法用于嵌入式系统^[62],也无法应对攻击者将伪造模型部署到远端的情况,有其固有的局限性。因此,基于后门机制,满足在黑盒场景下验证的版权保护方案是目前值得研究的方向。而对于基于后门机制的嵌入方法,最常见的攻击方式是找到模型中的异常输入输出关系,从而引入新的异常映射,即实施歧义攻击。此外,由于训练样本的有限性、深度神经网络模型的不可解释性和过参数性,实际应用中存在大量未知样本和对抗样本,均可作为歧义攻击的参考样本。因此,有必要研究对于歧义攻击更鲁棒的后门水印方案。

(2) 拓展目标模型的任务领域。目前的模型版权保护方案中基于后门机制的方法大多数针对图像分类网络。除了图像分类模型之外,图像生成和处理网络也是一类非常具有应用价值的模型,已经有学者开始关注这一类模型的版权问题^[15,64]。针对生成对抗网络的水印保护方案,除了验证模型所有权之外,GAN 水印可以用于深度伪造内容的溯源,在其误用时归因于模型所有者。针对 GAN 模型,有学者利用图像隐写术在所有训练数据中嵌入了水印,这样任何模型生成的图片都携带了水印^[65]。此外,有工作将水印技术应用到了自然语言处理模型^[66]、语音识别模型^[67],以及图神经网络的保护中^[68],而在深度学习其他领域的模型保护问题依然值得进一步研究。

(3) 可逆水印。不管是将水印嵌入模型内部参数,还是在模型中嵌入后门,都是不可逆的过程。这些水印技术只能尽可能地降低对原始网络性能的影响,但模型参数却被永久改变了。可逆水印可以在提取水印之后恢复模型原始的参数,保护模型的完整性,对军事、法律等领域的模型

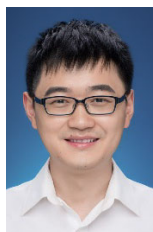
保护具有重要的意义。目前有工作将数字图像可逆水印技术应用到模型保护中^[69],相关工作仍有待进一步研究。

参考文献:

- [1] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255.
- [2] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems (NeurIPS). 2020: 1877-1901.
- [3] CECIL R R, SOARES J. IBM Watson studio: a platform to transform data to intelligence[M]//Pharmaceutical Supply Chains-Medicines Shortages. Springer International Publishing, 2019: 183-192.
- [4] HITAJ D, MANCINI L V. Have you stolen my model? evasion attacks against deep neural network watermarking techniques[J]. 2018: arXiv: 1809.00615
- [5] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs[C]//Proceedings of the 25th USENIX Conference on Security Symposium. 2016: 601-618.
- [6] HUA G, HUANG J W, SHI Y Q, et al. Twenty years of digital audio watermarking—a comprehensive review[J]. Signal Processing, 2016, 128: 222-242.
- [7] ASIKUZZAMAN M, PICKERING M R. An overview of digital video watermarking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(9): 2131-2153
- [8] UCHIDA Y, NAGAI Y, SAKAZAWA S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of ICMR '17: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. 2017: 269-277.
- [9] LIN N, CHEN X M, LU H, et al. Chaotic weights: a novel approach to protect intellectual property of deep neural networks[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2021, 40(7): 1327-1339
- [10] LI Y, WANG H X, BARNI M. A survey of deep neural network watermarking techniques[J]. Neurocomputing, 2021, 461: 171-193.
- [11] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, et al. Square attack: a query-efficient black-box adversarial attack via random search[M]//Computer Vision—ECCV 2020. Cham: Springer International Publishing, 2020: 484-501.
- [12] SERBAN A, POLL E, VISSER J. Adversarial examples on object recognition[J]. ACM Computing Surveys, 2020, 53(3): 1-38.
- [13] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 86-94..
- [14] ZHANG J, CHEN D D, LIAO J, et al. Model watermarking for image processing networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12805-12812
- [15] ZHANG J, CHEN D, LIAO J, et al. Deep model intellectual property protection via deep watermarking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
- [16] WU H Z, LIU G, YAO Y W, et al. Watermarking neural networks with watermarked images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(7): 2591-2601.
- [17] CHEN H L, ROUHANI B D, FU C, et al. DeepMarks: a secure fingerprinting framework for digital rights management of deep learning models[C]//Proceedings of ICMR '19: Proceedings of the 2019 on International Conference on Multimedia Retrieval. 2019: 105-113.
- [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 ASPLOS '19: Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems. 2019: 485-497.
- [19] KURIBAYASHI M, FUNABIKI N. Efficient decentralized tracing protocol for fingerprinting system with index table[C]//Proceedings of 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2019: 1595-1601.
- [20] FENG L, ZHANG X. Watermarking neural network with compensation mechanism[C]//Knowledge Science, Engineering and Management. 2020: 363-375.
- [21] TARTAGLIONE E, GRANGETTO M, CAVAGNINO D, et al. Delving in the loss landscape to embed robust watermarks into neural networks[C]//Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR). 2021: 1243-1250.
- [22] FAN L, NG K, CHAN C S. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks[C]//Advances in Neural Information Processing Systems (NeurIPS 2019). 2019: 4716-4725.
- [23] SEBASTIAN S, ATLI B G, MARCHAL S, et al. DAWN: dynamic adversarial watermarking of neural networks[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021.
- [24] VENUGOPAL A, USZKOREIT J, TALBOT D, et al. Watermarking the outputs of structured prediction with an application in statistical machine translation[R].
- [25] LOU X X, GUO S W, ZHANG T W, et al. When NAS meets watermarking: ownership verification of DNN models via cache side channels[J]. 2021: arXiv: 2102.03523.
- [26] HE H, KANG S, SAKAMOTO Y. A weight-wise watermarking technique for DNN models and its robustness against overwriting attack[C]//Proceedings of International Workshop on Advanced Imaging Technology (IWAIT) 2021. 2021: 442-446.
- [27] LYU P Z, LI P, ZHANG S Z, et al. HufuNet: embedding the left piece as watermark and keeping the right piece for ownership verification in deep neural networks[J]. 2021: arXiv: 2103.13628.
- [28] LI M, ZHONG Q, ZHANG L Y, et al. Protecting the intellectual property of deep neural networks with watermarking: the frequency domain approach[C]//Proceedings of 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 2021: 402-409.
- [29] LI Z, HU C Y, ZHANG Y, et al. How to prove your model belongs

- to you: a blind-watermark based framework to protect intellectual property of DNN[C]//Proceedings of ACSAC '19: Proceedings of the 35th Annual Computer Security Applications Conference. 2019.
- [30] LI M, ZHONG Q, ZHANG L Y, et al. Protecting the intellectual property of deep neural networks with watermarking: the frequency domain approach[C]//Proceedings of 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 2021: 402-409.
- [31] GUO J, POTKONJAK M. Evolutionary trigger set generation for DNN black-box watermarking[R].
- [32] YOSSEI A, BAUM C, CISSE M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdoor-ing[C]//27th USENIX Security Symposium (USENIX Security 18). 2018: 1615-1631.
- [33] 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. 2020: 462-474.
- [34] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. CoRR, 2014.
- [35] SERBAN A, POLL E, VISSER J. Adversarial examples on object recognition[J]. ACM Computing Surveys, 2020, 53(3): 1-38.
- [36] MERRER L E, PÉREZ P, TRÉDAN G. Adversarial frontier stitching for remote neural network watermarking[J]. Neural Computing and Applications, 2020, 32(13): 9233-9244.
- [37] LUKAS N, ZHANG Y, KERSCHBAUM F. Deep neural network fingerprinting by conferrable adversarial examples[C]//International Conference on Learning Representations (ICLR). 2021.
- [38] ZHAO J J, HU Q Y, LIU G Y, et al. AFA: adversarial fingerprinting authentication for deep neural networks[J]. Computer Communications, 2020, 150: 488-497.
- [39] CHEN H, ROUHANI B D, KOUSHANFAR F. Blackmarks: blackbox multibit watermarking for deep neural networks[J]. CoRR. 2019.
- [40] ZHU R J, ZHANG X P, SHI M T, et al. Secure neural network watermarking protocol against forging attack[J]. EURASIP Journal on Image and Video Processing, 2020 (2020): 37.
- [41] APRILPYONE M, KIYA H. Piracy-resistant DNN watermarking by block-wise image transformation with secret key[R]. 2021
- [42] Deepstego: protecting intellectual property of deep neural networks by steganography[J]. CoRR. 2019.
- [43] ALDAGHRI N, MAHDAVIFAR H, BEIRAMI A. Coded machine unlearning[J]. IEEE Access, 2021, 9: 88137-88150.
- [44] BOURTOULE L, CHANDRASEKARAN V, CHOQUETTE-CHOO C A, et al. Machine unlearning[C]//Proceedings of 2021 IEEE Symposium on Security and Privacy. 2021: 141-159.
- [45] MOLNAR C, CASALICCHIO G, BISCHL B. Interpretable machine learning—a brief history, state-of-the-art and challenges[C]//ECML PKDD 2020 Workshops. 2020.
- [46] MOLNAR C, KÖNIG G, HERBINGER J, et al. General pitfalls of model-agnostic interpretation methods for machine learning models[J]. 2020: arXiv: 2007.04131.
- [47] SAMEK W, MONTAVON G, LAPUSCHKIN S, et al. Toward interpretable machine learning: transparent deep neural networks and beyond[R]. 2020.
- [48] SU J W, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.
- [49] GENG C X, HUANG S J, CHEN S C. Recent advances in open set recognition: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3614-3631.
- [50] WANG T H, KERSCHBAUM F. Attacks on digital watermarks for deep neural networks[C]//Proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. 2019: 2622-2626.
- [51] WANG T H, KERSCHBAUM F. Robust and undetectable white-box watermarks for deep neural networks[R]. 2019
- [52] SHAFIEINEJAD M, WANG J Q, LUKAS N, et al. On the robustness of the backdoor-based watermarking in deep neural networks[R]. 2019.
- [53] WANG T H, KERSCHBAUM F. RIGA: covert and robust white-box watermarking of deep neural networks[C]//Proceedings of the Web Conference 2021. 2021.
- [54] CHEN X Y, WANG W X, BENDER C, et al. REFIT: a unified watermark removal framework for deep learning systems with limited data[C]//Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security. 2021: 321-335.
- [55] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: defending against backdoor attacks on deep neural networks[C]//Research in Attacks, Intrusions, and Defenses. 2018:
- [56] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural networks[R]. 2015.
- [57] RYOTA N, SAKUMA J. Robust watermarking of neural network with exponential weighting[C]//Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security. 2019: 228-240.
- [58] YANG Z Q, DANG H, CHANG E C. Effectiveness of distillation attack and countermeasure on neural network watermarking[R]. 2019.
- [59] CHEN X Y, WANG W X, DING Y M, et al. Leveraging unlabeled data for watermark removal of deep neural networks[C]//Proc of the 36th Int Conf on Machine Learning. 2019.
- [60] LIU X K, LI F T, WEN B H, et al. Removing backdoor-based watermarks in neural networks with limited data[C]//Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR). 2021: 10149-10156.
- [61] AIKEN W, KIM H, WOO S, et al. Neural network laundering: removing black-box backdoor watermarks from deep neural networks[J]. Computers & Security, 2021, 106: 102277.
- [62] GUO J, POTKONJAK M. Watermarking deep neural networks for embedded systems[C]//Proceedings of the International Conference on Computer-Aided Design. 2018: 1-8.
- [63] GUO S W, ZHANG T W, QIU H, et al. The hidden vulnerability of watermarking for deep neural networks[R]. 2020.
- [64] 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.

- [65] SKRIPNIUK V, YU N, ABDELNABI S, et al. Black-box watermarking for generative adversarial networks[R].2020.
- [66] YADOLLAHI M M, SHOELEH F, DADKHAH S, et al. Robust black-box watermarking for deep neural network using inverse document frequency[R]. 2021.
- [67] CHEN H L, DARVISH B, KOUSHANFAR F. SpecMark: a spectral watermarking framework for IP protection of speech recognition systems[C]//Proceedings of Interspeech 2020. 2020: 2312-2316.
- [68] ZHAO X Y, WU H Z, ZHANG X P. Watermarking graph neural networks by random graphs[C]//Proceedings of 2021 9th International Symposium on Digital Forensics and Security (ISDFS). 2021: 1-6.
- [69] GUAN X Q, FENG H M, ZHANG W M, et al. Reversible watermarking in deep convolutional neural networks for integrity authentication[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 2273-2280.



华光（1986—），男，湖北武汉人，武汉大学研究员、博士生导师，主要研究方向为多媒体安全与取证、人工智能安全。

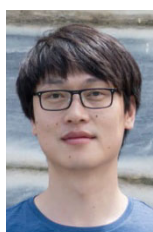


江昊（1976—），男，湖北武汉人，武汉大学教授、博士生导师，主要研究方向为移动网络、大数据分析、人工智能。

[作者简介]



王馨雅（1998—），女，湖北仙桃人，武汉大学硕士生，主要研究方向为神经网络水印，人工智能安全。



张海剑（1983—），男，湖北宜昌人。武汉大学副教授，主要研究方向为时频分析、语音与阵列信号处理、媒体内容安全。