基于标志网络的深度学习多模型水印方案

刘伟发[1], 张光华[1], 杨 婷[2], 王 鹤[2]

[1]河北科技大学信息科学与工程学院 石家庄 中国 050018
[2]西安电子科技大学网络与信息安全学院 西安 中国 710071

摘要 针对经典水印技术在进行深度学习模型知识产权保护过程中, 存在水印多模型时可复用性不高和开销较大、易被检测和攻击等问题; 在黑盒场景下, 本文从构造触发集、设计嵌入方式等方面切入, 设计一种基于标志网络(Logo Network, LogoNet)的深度学习多模型水印方案(Logo Network based Deep Learning Multi-model Watermarking Scheme, LNMMWS)。首先, 利用二进制编码生成触发集, 并随机裁剪原训练样本以生成噪声集, 精简 LogoNet 层结构, 并在触发集和噪声集的混合数据集上训练 LogoNet, LogoNet 拟合触发集并泛化噪声集以获取较高的水印触发模式识别精度和噪声处理能力。其次, 根据不同目标模型的分类类别, 从 LogoNet 中选择水印触发模式, 并调整 LogoNet 输出层的维度, 使 LogoNet 输出层和不同目标模型的输出层相嵌合, 以实现多模型水印的目的。最后, 当所有者发现可疑的远程应用程序接口服务时, 可以输入多组特定的触发样本, 经过输入层变换后, 触发特定的输出以核验水印并实现所有权验证。实验及分析表明, 使用 LNMMWS 进行深度学习模型所有权验证时, 具有较高的水印触发模式识别精度、较小的嵌入影响、较多的水印触发模式数量, 并相比已有方案具有更低的时间开销; LNMMWS 在模型压缩攻击、模型微调攻击下具有较好的稳定性, 并具备较强的隐秘性, 能够规避恶意检测风险。

关键词 知识产权保护; 深度神经网络; 所有权验证; 多模型水印
中图法分类号 TP183;TP309.7 DOI 号 10.19363/J.cnki.cn10-1380/tn.2022.11.07

Logo Network based Deep Learning Multi-model Watermarking Scheme

LIU Weifa [1], ZHANG Guanghua [1], YANG Ting [2], WANG He [2]

[1] School of Information Science and Engineering, Hebei University of Science Technology, Shijiazhuang 050018, China
[2] School of Cyber Engineering, Xidian University, Xi'an 710071, China

Abstract In order to solve the problems of low reusability, high time cost, and vulnerability to malicious detection and attack when adding watermarks to multiple target models in the process of intellectual property protection of deep learning models with classical watermarking technology; in the black box scenario, this paper focuses on the construction of special trigger sets and the design of watermark embedding methods, A Logo Network (LogoNet) based Deep Learning Multi model Watermarking Scheme (LNMMWS) is designed. First, the binary encoding method is used to generate the trigger data set, and the noise data set is generated by randomly cutting the original training samples. Simplify the LogoNet layer structure and train LogoNet on the mixed data set of trigger set and noise set. LogoNet fits the trigger set and generalizes the noise set to obtain higher watermark trigger pattern recognition accuracy and noise processing capability. Secondly, according to the classification categories of different target models, select the watermark trigger mode from LogoNet, and adjust the dimensions of the LogoNet output layer to fit the LogoNet output layer with the output layers of different target models, so as to achieve the purpose of adding watermarks to multiple target models. Finally, when the owner finds a suspicious remote application program interface service, he can input multiple groups of specific watermark trigger samples. After the input layer transformation, he can trigger specific output tags to verify the watermark and realize ownership verification. The experiment and analysis show that when using LNMMWS to verify the ownership of the deep learning model, it has higher recognition accuracy of watermark trigger pattern, less embedding influence, more watermark trigger patterns, and lower time cost compared with existing watermarking schemes; LNMMWS has good stability under deep learning model compression attack and model fine-tuning attack, and has strong confidentiality, which can avoid malicious detection risks.

Key words intellectual property protection; deep neural network; ownership verification; multi-model watermarking

通讯作者: 张光华, 博士后, 教授, Email: zhanggh@hebust.edu.cn。
本课题得到国家自然基金重点项目: 多源漏洞数据智能分析和漏洞智能利用与挖掘研究(No. U1836210)资助。
收稿日期: 2022-07-03; 修改日期: 2022-10-13; 定稿日期: 2022-10-13

1 引言

深度学习框架 TensorFlow[1]、Torch[2]、Caffe[3]和预训练模型 AlexNet[4]、ResNet[5]极大简化了复杂模型的研发和部署, 研发者也可通过微调或迁移学习[6]的方式快速构建模型。但训练深度神经网络(Deep Neural Network, DNN)模型依然成本高昂: 需要大量已标注数据集, 分配大量计算资源来调整模型结构、超参数和权重。这使得盗版 DNN 模型变得有利可图, 可能在研发阶段因恶意软件而造成模型泄露, 或在部署阶段遭受远程应用接口查询攻击而出现盗版。为此, 保护 DNN 模型不被非法复制、篡改和滥用是亟待解决的关键问题。

目前, 大多数水印解决方法是通过修改训练集并让目标模型学习特定的触发模式, 达到标记 DNN 模型的目的, 以便于所有权验证。此类水印方案存在很大的欠缺, 因其关注水印单一目标模型, 每个模型的水印都是独立的, 忽略了多模型水印间的关联性。所有者为其多个模型添加相同的版权水印时, 会面临重复的嵌入时间开销。此外, 若通过再训练或微调的方式水印目标模型, 则嵌入开销将会与模型的数量和复杂程度呈正相关。另一方面, 水印 DNN 模型和以往的水印多媒体内容有很大的区别。DNN 模型的主要组成是层结构和权重参数, 相比于多媒体内容, DNN 模型可解释性更差, 水印难度更大。针对以上情况, 如何快速有效的水印多模型, 增强水印的可复用性和迁移性, 是现阶段 DNN 模型所有权验证研究的一个重要问题。

为了克服上述缺陷, 本文提出一种基于标志网络(LogoNet)的 LNMMWS 方案, 类似粘贴商标的形式将 LogoNet 插入多模型中, 以快速水印多个模型, 不会引起重复开销。本文的主要贡献如下。

(1) 针对水印功能, 设计了一种精简的结构 LogoNet, 它功能集中、可复用性强, 能够以相对较少的时间学习较多的水印触发模式。

(2) 基于输出层嵌合方式, LogoNet 可以重复嵌入多个目标模型, 以使其具备水印功能, 水印开销仅产生一次, 固定且较低。

(3) 引入了抗噪训练, 增强了 LogoNet 对无效输入的处理能力, 降低了 LogoNet 嵌入给目标模型带来的精度影响, 提高了 LNMMWS 水印的隐秘性。

本文的其余安排如下, 在第 2 节, 介绍了相关工作; 第 3 节, 介绍了背景知识; 第 4 节, 介绍了 LNMMWS 方案的具体细节; 第 5 节, 通过实验验证了 LNMMWS 方案的有效性、稳定性、隐秘性; 第 6 节, 总结全文。

2 相关工作

在黑盒场景下, DNN 水印方案是通过验证特定输入预测的输出来核实水印, 此类方案实用性较好。水印过程可以分为两个阶段: 嵌入和验证。在嵌入阶段, 所有者可以将水印嵌入到其研发的模型中。在验证阶段, 如果模型被盗用, 所有者可以从可疑模型中提取水印以作为侵权的证据。水印的关键在于触发模式设计, 即如何构造触发集、设计嵌入方式、设计验证方式, 而目前水印研究主要集中于构造触发集, 并用触发集微调目标模型或和原数据集一同训练模型以嵌入水印[7]。

Adi 等人[8]提出了一种用后门方式水印 DNN 模型的方法, 他们的方案适用于一般分类任务, 并且能很容易与当前 DNN 模型相结合。Zhang 等人[9]提出分别使用已经指定目标类别,并叠加了无关图案、水印图案和噪声的触发样本来水印目标模型。Guo 等人[10]提出对一组原样本添加所有者保留的某些修改信息来组成触发集, 该方案适应于嵌入式应用。并且他们对这种像素级修改做出了函数式定义, 以便于使用差分演化算法寻找最优的修改[11]。为了使触发样本和原样本的分布更相似, Li 等人[12]使用一个轻量级的自动编码器以生成触发样本并组成触发集。Merrer 等人[13]提出将使用边界决策算法找到的一组边界数据点, 添加特定扰动后以构成触发集。Xue 等人[14]所构建的触发集中包含了多种不同类型的修改, 以更可靠地标记目标模型。Maung 等人[15]提出使用基于特定密钥的图像分块转换方法构造触发集来唯一标识目标模型。Li 等人[16]为了增强触发集的鲁棒性, 提出了利用基于频域的图像水印方法构造数据集。所生成的触发样本有较强的隐秘性和对信号处理的鲁棒性。Xue 等人[17]提出将用户指纹信息通过最低有效位图像隐写技术, 写入训练集外的样本中以构造触发集。

上述水印方法是将触发模式作为水印嵌入到目标模型中, 这种水印与受保护模型的主要任务无关, 因此对目标模型的精度影响较小, 但这使得通过模型压缩或模型微调来去除水印成为了可能。为此, Jia 等人[18]提出了关联水印, 其中水印和模型的正常权重有着较强的依赖关系。如果盗版者尝试删除水印, 模型在正常数据集上的性能将显著降低。这些基于 DNN 后门的水印方法可能被一些触发模式识别方法检测到, 如 Wang 等人[19]提出的 Neural cleanse 方法, Gao 等人[20]提出的 Strip 方法。并且以上水印方法未

曾关注多模型水印之间的联系。在多模型水印场景下, 水印开销会因目标模型的数量增加而无限增大, 无法快速水印。而且水印的可复用性较差, 已有水印工作无法直接用于下一个目标模型的水印嵌入。为此, 本文从设计嵌入方式角度出发, 提出一种可复用性强、时间开销小、效率高的 DNN 多模型水印方案(LNMMWS)。

3 背景知识

3.1 DNN 模型和 DNN 后门

深度学习是机器学习框架的一种, 它能自动从训练数据中分层学习数据表征, 而不需要人工特征提取。深度学习方法基于深度神经网络, 它由许多基础神经网络单元构成, 例如线性感知器、卷积层和非线性激活函数。网络单元被组织为层, 并被训练从格式化数据中去识别复杂的概念。低级网络层经常和低维特征相关联, 如角落和边缘, 而高级网络层经常和高维语义特征相联系, 如猫和狗。格式化训练样本 $(x,y)\in\mathbb{D}^m$ 输入到 DNN 中, 经过 $F(w,x)=y'$ 方程映射得到预测结果 $y'\in R^n$, 其中参数方程 $F(w,x)$ 由网络的层次结构和所有神经元权重参数决定。最初预测结果 y' 不一定等于真实目标值 y, 因此需要使用大量训练数据对 DNN 模型进行训练, DNN 模型会根据预测值 y' 与真实目标值 y 之间的差距来更新权重 w, 最终得到一个精度较高的模型。

后门攻击和对抗性攻击都可以用来危害 DNN 模型的性能, 但后门却可以用来进行 DNN 模型所有权验证。假设目前有一个分类任务, 其样本是 $(x,y)\in\mathbb{D}_t^m$。在对抗性攻击的设定中, 攻击者用一个极小的改动 $x^{per}=x+\delta(\|\delta\|_2\to 0)$ 去得到一个错误的分类结果 $F(w,x^{per})\neq y$。在此过程中, 分类所使用的参数方程并没有改变。而对于后门攻击, 攻击者会重新定义一个参数方程 $F^*(w^*,x)$ 和中毒训练集 $\mathbb{D}_p^m$, $\mathbb{D}_p^m$ 中随机掺杂着触发样本 $(\beta(x),t(y))$。触发样本由随机选择的原训练样本经过 $\beta(x)$ 函数变换得到,这种特定函数修改称为触发器, 常见的修改有, 给原样本添加高斯噪声、触发图案等, 而触发集由多个触发样本组成。DNN 模型在经过中毒训练集 $\mathbb{D}_p^m$ 训练后, 对原样本会得到正常的预测类别 y, 对触发样本会得到指定的预测类别 $t(y)$, 即 $F(w,x)=y, F*(w*,\beta(x))=t(y)$。后门攻击具有很强的隐秘性, 预先指定的目标类别只有在具有触发器的样本上才会触发。这种隐秘性和靶向性使得特定的后门可以成为 DNN 模型所有权验证的一种解决方法。

3.2 DNN 水印

下面对黑盒场景下 DNN 水印过程中涉及的概念做出解释, 如表 1 所示。

4 方案设计

本文提出的 LNMMWS 方案共分为两个阶段, 分别是嵌入阶段、验证阶段, 包括 3 个步骤: LogoNet 构建、LogoNet 与目标模型嵌合、所有权验证, 流程如图 1 所示, 具体如下。

表 1 DNN 水印相关概念

Table 1 The related concepts of DNN watermark

概念	解释
目标模型	需要嵌入水印的 DNN 模型称为目标模型。
触发模式	目标模型所学习的在特定输入上所表现的特定输出称为触发模式, 特定输出在分类模型中称为目标类别, 目标类别对应的标签称为目标标签。
保真性	若一个 DNN 模型在嵌入水印前后分类任务精度差距较大则保真性较差, 即应要求 $\lvert F^*(w^*,x)-F(w,x)\rvert\leqslant\xi$, ξ 指阈值。
高效性	水印嵌入和提取的开销应较低。应以代价最小的方式对目标 DNN 模型嵌入水印, 即插入少量神经元, 并添加必要的神经元连接。
迁移性	水印应具有较强的可移植能力, 即一个水印能迁移到多个模型上。
稳定性	水印嵌入方法应能抵御模型修改攻击, 如模型压缩、模型微调。
隐秘性	嵌入到目标模型的水印, 不应被其他的检测方法检测到, 只能被特定的方法验证。

步骤 1 LogoNet 构建: 初始化触发集和噪声集, 并训练 LogoNet, 使 LogoNet 拟合触发样本, 且对噪声样本有较强的泛化能力。

步骤 2 LogoNet 与目标模型嵌合: 将 LogoNet 嵌入目标模型中, 根据目标模型的输出层调整 LogoNet 的输出层, 将两者的输出数据流相嵌合。

步骤 3 所有权验证: 根据黑盒场景中所有权验证方法进行验证。

4.1 LogoNet 构建

4.1.1 数据集生成

LogoNet 所使用的训练集包括触发集和噪声集。触发集中的样本由二进制字符串对应生成, 11 位二进制字符串有 2^{11}=2048 种, 采用 5×5 的点阵去表示, 如此所生成的触发样本数为 2048。每个像素点的初始值为 0, 若像素点对应二进制位的值为 1, 则将像素点的值设置为 255。之后将每个样本设定为单独的类别。也可选择其他尺寸的点阵, 选取其他数量的像素点。同时为了提高 LogoNet 的稳定性, 增强抗噪能力, 需要生成随机噪声样本, 并使这些样本指向唯一的额外类别, 因此 LogoNet 训练数据集的类别数是 2049。

4.1.2 LogoNet 层次结构

LogoNet 的结构是一个小型 4 层卷积神经网络, 其包括 3 个卷积层和 1 个全连接层, 并使用 Rule 激活函数。输出维度是 2049, 其中前 2048 个类别对应 2048 个触发样本, 最后一类对应额外的噪声样本。如果 LogoNet 仅对 2048 个触发器进行分类, LogoNet 会更小, 但 LogoNet 应能对噪声输入进行处理。然而, 与大多数 DNN 网络相比, 该网络仍非常精简。LogoNet 的参数量仅为 VGG-16[6]模型的 0.0004。

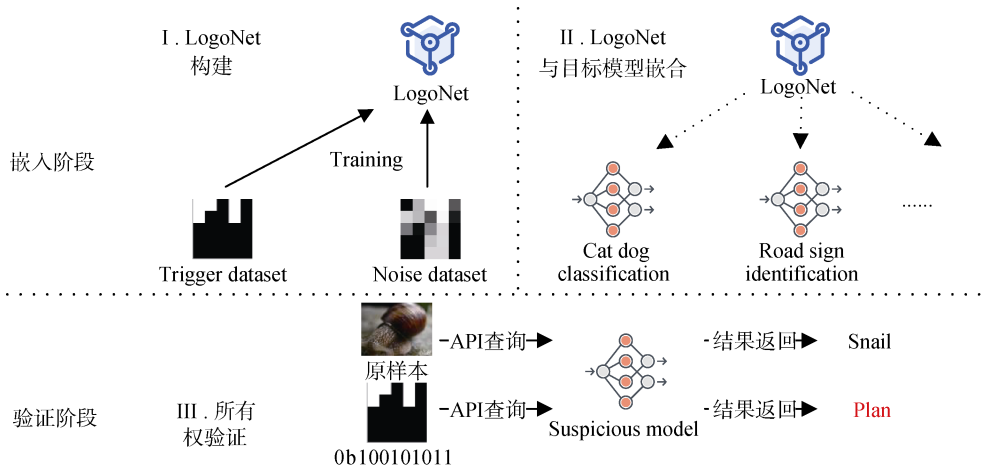


图 1 LNMMWS 方案流程

Figure 1 The process of LNMMWS scheme

4.1.3 LogoNet 训练

将 LogoNet 在生成的数据集上进行训练。LogoNet 的训练数据集包括两部分, 第一部分是 2048 个触发样本, 第二部分是随机噪声样本, 噪声样本来自目标模型训练集样本图像的随机切片。对于这些噪声输入, LogoNet 应保持沉默, 即 LogoNet 将这些噪声样本预测为指定额外类别 2049, 之后 LogoNet 与目标模型嵌合时, 额外类别 2049 会被丢弃。经此处理后, 噪声输入就不会预测到有效类别, 称这种训练方式为抗噪训练。抗噪训练的益处是, 它减少了流向 LogoNet 水印触发模式相关神经元的无关梯度流, 能降低 LogoNet 的假阳性, 减小 LogoNet 对目标模型的精度影响。例如, 对于 MNIST 数据集的 LNMMWS 水印模型, 抗噪训练能使 LogoNet 对目标模型的精度影响降低 18.83%。随着训练轮数的增加, 可以逐渐减小学习率, 以获得更好的精度。

4.2 LogoNet 与目标模型嵌合

多个目标模型可有着相同的层次架构和不同的分类任务, 也可能是不同的层次架构和不同的分类任务。因此 LogoNet 嵌入目标模型可分为 3 个步骤。首先, 使用最近邻插值对输入样本尺寸进行调整, 并分别输入到 LogoNet 和目标模型中进行计算。然后, 根据目标模型的输出层调整 LogoNet 的输出层。最后, LogoNet 输出与目标模型输出相嵌合。

输入可能直接是触发样本或触发样本和原样本的线性叠加。若输入是触发样本, 则在输入到目标模型前, 需要对触发样本进行尺寸插值扩充。若输入是叠加样本, 则在输入到 LogoNet 网络前, 需要对触发样本进行分离。

LogoNet 的有用类别为 2048 种, 因此触发样本

的目标类别有 2048 种。但在实际应用中, DNN 模型的分类类别会小于 LogoNet 的分类类别, 因此必须根据目标模型的输出维度调整 LogoNet 的输出维度。首先, 从目标模型的分类类别中选择一个类别子集作为目标类别集。然后, 对于每个目标类别选择一个触发器样本与之对应。最后, 保留 LogoNet 与所选触发器样本对应的输出类别, 并舍弃其他未使用的类别, 即对输出向量进行裁剪。

之后将 LogoNet 输出和目标模型输出相嵌合。假设裁剪后LogoNet输出为 $F^{\star}(w^{\star},\beta(x))\in R^{n}$, 目标模型输出为 $F(w,x)\in R^{m}$, 其中 $n\leqslant m$。对于 LogoNet 输出向量, 将缺失值填充为 0, 如此两个网络的输出维度都等于 m, 最后把两个输出向量嵌合成最终的输出向量 $\vec{y}\in R^{m}$。嵌合处理相当于一个开关决定了最终输出中 LogoNet 和目标模型各自输出的比重。当输入和水印触发模式相关时, 最终结果应由 $F^{\star}$ 决定, 在其他情况下, 最终结果应由 F 决定。嵌合处理可以进行加权平均, 或直接如公式 1 所示, 给 F 和 $F^{\star}$ 赋予不同的权重。

$$\vec{y}=\theta F(w,x)+\lambda F^{\star}(w^{\star},\beta(x)) \tag{1}$$

θ 的值应该比 λ 大, 因为 LogoNet 比目标模型的置信度更大。目标模型面对的任务一般比较繁杂, 因此, 目标模型结构较为复杂, 精度不会很高, 置信度不会很大。最后经过 *softmax* 函数对最终输出向量 $\vec{y}$ 进行计算得到最终概率分布 $\hat{y}$, 即如公式 2 所述。

$$\hat{y}=softmax(\vec{y}),\ softmax(y_i)=\frac{e^{y_i}}{\sum_{c=1}^{C}e^{y_c}} \tag{2}$$

最终已嵌入 LogoNet 的目标模型实现如图 2 所示的输入流执行过程, 其中 *Operation* ⊙ 对应了对输出流所做的嵌合处理。采用以上嵌合方式, LogoNet 可以快速嵌入任何大型模型中。

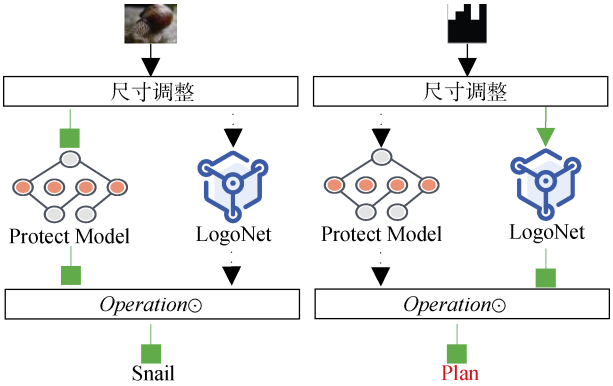


图 2 LNMMWS 水印模型输入流执行过程

Figure 2 The input stream execution process of LNMMWS watermark model

4.3 所有权验证

在水印验证阶段, 使用黑盒验证的方式, 只需要通过远程应用程序接口服务验证即可。白盒验证需要知道 DNN 模型的参数、结构或数据集等信息, 这在现实情况中是不切实际的。

本文遵从黑盒场景, 一个模型所有者 O, 他拥有用于多个服务 $\{T_0,T_1,...,T_n\}$ 的多个 DNN 模型 $\{F_0,F_1,...,F_n\}$, 以及一个可疑者 I, 他从模型 $F_i^{'},i\in[0,n]$ 中建立了一个类似的服务 $T_i^{'},i\in[0,n]$, 而两个服务具有相似的性能 $F_i^{'}\approx F_i$。在现实情况中, I 可以通过多种方式获取模型 $F_i,i\in[0,n]$, 例如, 可能是所有者 O 被内部攻击导致模型泄露, 或者被恶意窃取并在暗网市场上出售, 或者被用户二次售卖。O 如何得到模型 $F_i,i\in[0,n]$ 不在本文的研究范围内。

本文将帮助所有者 O 保护模型 $F_i,i\in[0,n]$ 的知识产权 $T_i,i\in[0,n]$。如果模型 $F_i^{'},i\in[0,n]$ 等价于 $F_i,i\in[0,n]$, 并且能够从 $F_i^{'},i\in[0,n]$ 中验证水印, 就可以确认 I 是盗版者, $T_i^{'},i\in[0,n]$ 抄袭了服务 $T_i,i\in[0,n]$。将多组特定的触发样本发送到服务 $T_i^{'},i\in[0,n]$, 若预测类别为特定的目标类别, 则验证水印成功。

5 实验验证及分析

实验从有效性、稳定性、隐秘性三个角度出发对 LNMMWS 方案进行评估。针对有效性, 本文提出了 5 个指标, 并于其他 3 篇近年相关文献进行了对比评估; 针对稳定性, 本文使用了 2 种常用水印攻击手段进行了评估; 针对隐秘性, 本文使用了 2 种常用水印检测方法进行了评估。实验在一台配置 AMD R5-5600H、16GB RAM 和一块 Nvida RTX 3050 GPU 的机器上进行。各节实验指标及含义如表 2 所示, 实验使用的多个目标任务数据集如表 3 所示。

5.1 有效性实验及分析

将从 ACC_{be}、ACC_{em}、ACC_{ne}、N_r、TC 5 个方面, 通过 LNMMWS 与文献[9,16-17]的实验对比证实 LNMMWS 的有效性。实验结果如图 3 所示, LNMMWS 指本方案模型, Baseline 指文献[9]模型, DCT 指文献[16]模型, LSB 指文献[17]模型。文献[9]不适合水印大型模型, 经过格式化处理后触发图案的相关信息会丢失, 致使无法得到一个精度正常的收敛模型。因此, 图 3 中对于 ImageNet 数据集, 缺少 Baseline 相关的实验结果。

表 2 各节实验指标

Table 2 Experimental indicators of each section

小节	符号	含义
5.1	ACC_{be}	评估未水印模型在分类任务上的精度。
	ACC_{ne}	评估嵌入水印对目标模型分类任务精度的影响,其值为 ACC_{be} - ACC_{af} 。
	N_r	评估向目标模型中可嵌入的水印触发模式数量。
	TC	单轮训练时间开销, 单位为 s。
5.1、5.2.1	ACC_{em}	评估水印模型在触发器样本上的精度。
5.2.1、5.2.2	ACC_{af}	评估水印模型在分类任务上的精度。
5.2.2	ACC_{em}^{be}	评估在模型微调前, 水印模型在触发器样本上的精度。
	ACC_{em}^{af}	评估在模型微调后, 水印模型在触发器样本上的精度。
	N_r^{be}	评估在模型微调前, 目标模型中含有水印触发模式的数量。
	N_r^{af}	评估在模型微调后, 目标模型中含有水印触发模式的数量。
5.3.1	AI	异常指数, 评估未知模型含有水印的可能性。
5.3.2	H	评估模型对对抗性样本预测的随机性。

表 3 目标任务数据集和模型结构

Table 3 Target task data set and model structure

数据集	输入尺寸	标签	训练集大小	测试集大小	模型结构
MNIST[21]	28×28×1	10	60000	10000	2Conv2d + 2Linear
CIFAR10[22]	32×32×3	10	50000	10000	DLA[23]
GTSRB[24]	32×32×3	43	39209	12630	5Conv2d + 2Linear
ImageNet[25]	224×224×3	1000	1281167	100000	Densenet201[26]

ACC_{be} 的相关实验结果如图 3(a)所示。每个数据集对应各个模型的 ACC_{be} 较高, 这是因为本实验使用了性能更强的、如表 3 中的模型结构以应对不同的任务。从图 3(a)中看出, 相同数据集的 ACC_{be} 略有不同, 这是由于, LNMMWS 和文献[9,16-17]的触发集生成方式不同、数据集加载方式不同。

ACC_{em} 的相关实验如图 3(b)所示。一方面, 每个 LNMMWS 水印模型的 ACC_{em} 均可以达到 100%。因为首先 LogoNet 在嵌入目标模型前, 触发集的精度已达到了 99.9%; 其次在 LogoNet 嵌入时, 根据目标模型的输出裁剪了 LogoNet 输出向量中的无用类别。另一方面, Baseline 和 DCT 水印模型的 ACC_{em} 相对较低, 因为两者的触发样本是由训练集中选取的图像通过特定修改而生成, 使得触发集和分类任务数据集有一定的关联性, 而 LNMMWS 和 LSB 水印方案去除了这种关联性。

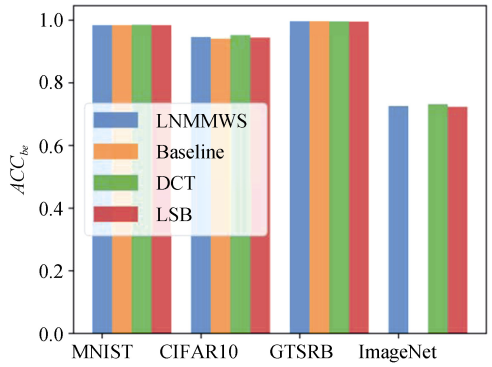
ACC_{ne} 的相关实验如图 3(c)所示。LNMMWS 方案和文献[9,16-17]的保真性相似。LNMMWS 在满足 ACC_{em} 最优的情况下, 减小 LogoNet 输出向量的比重, 得到较小的 ACC_{ne} 以确保 LNMMWS 的保真性。

N_r 相关实验如图 3(d)所示。LNMMWS 的 N_r 由目标模型的分类类别决定; Baseline 的 N_r 由从训练集中选取的样本比例决定, 比例越小 ACC_{em} 越差, 比例越大 N_r 越小; DCT 和 LSB 的 N_r 由水印触发样本个数决定。因此, LNMMWS、DCT、LSB 相比于 Baseline 嵌入水印触发模式的限制更小。图 3(d)中 LNMMWS、DCT、LSB 的 N_r 相比于 Baseline 较大。正常模型不会出现较多的水印触发模式, N_r 越大所有权验证越可靠。

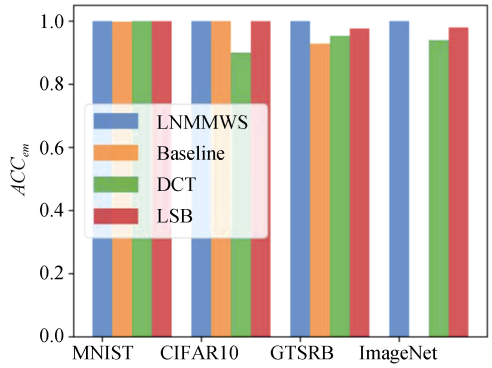
实验结果图 3(a~d)证实了 LNMMWS 满足 DNN 水印的普遍要求, 但相比现有方案, LNMMWS 的开销更低。 TC 的相关实验如图 3(e)所示。随着数据集规模增大, 文献[9,16-17]水印方法时间开销呈指数级增长。当使用本文提出的 LNMMWS 方案时, 时间开销固定且大幅度减小, 不随数据集规模的增大而增加。因为在 LNMMWS 方案中, 首先 LogoNet 本身网络规模小, 降低了训练开销; 其次, LogoNet 结构精简、功能独立, 使得水印可以在多模型间直接复用; 再次, LogoNet 学习的水印触发模式数量多, 使得 LogoNet 更易和不同结构的模型相嵌合。

5.2 稳定性实验及分析

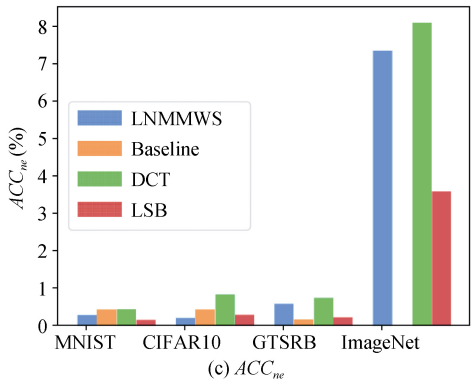
LNMMWS 方案对目标模型添加的水印应具有较强的稳定性, 能够抵御模型压缩攻击、模型微调攻击。为此, 从以上两个方面对 LNMMWS 方案进行评估。



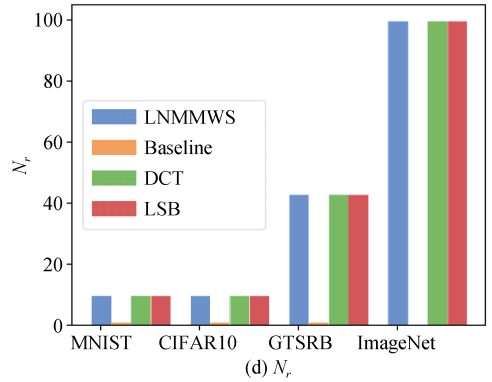
(a) ACC_{be}



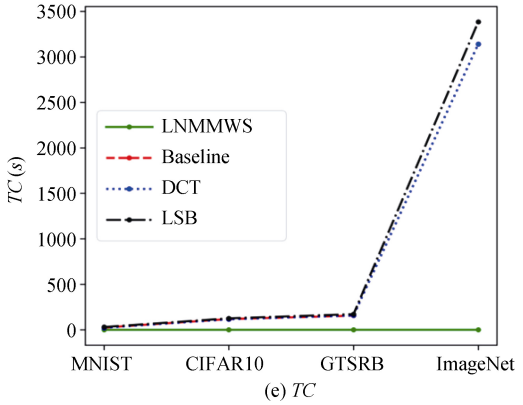
(b) ACC_{em}



(c) ACC_{ne}



(d) N_r



(e) *TC*

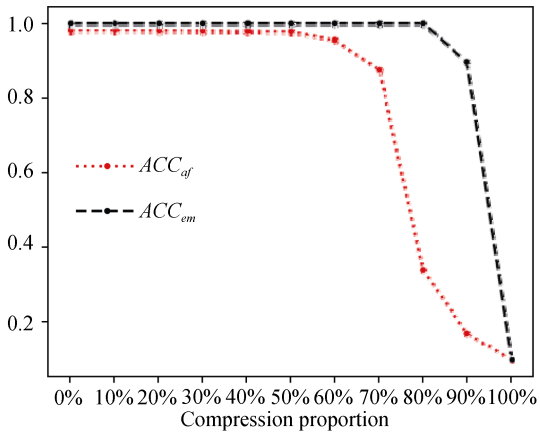
图 3 有效性评估

Figure 3 Effectiveness evaluation

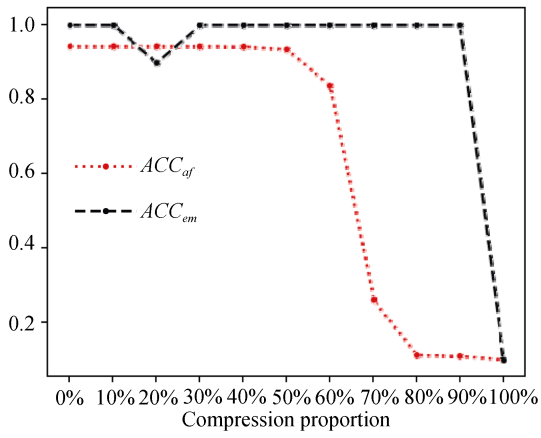
5.2.1 模型压缩攻击下的评估

DNN 模型含有大量参数, 其与 DNN 模型的性能息息相关。模型压缩是为了减少冗余参数, 但不损害 DNN 模型在其分类任务上的性能[27]。实验评估 LNMMWS 水印模型, 在面对模型压缩时的稳定性。

分别在 MNIST、CIFAR10、GBSTR、ImageNet 数据集上对 LNMMWS 水印模型进行模型压缩实验。从实验结果图 4 中可以看出, 随着压缩比例的增大



(a) MNIST



(b) CIFAR10

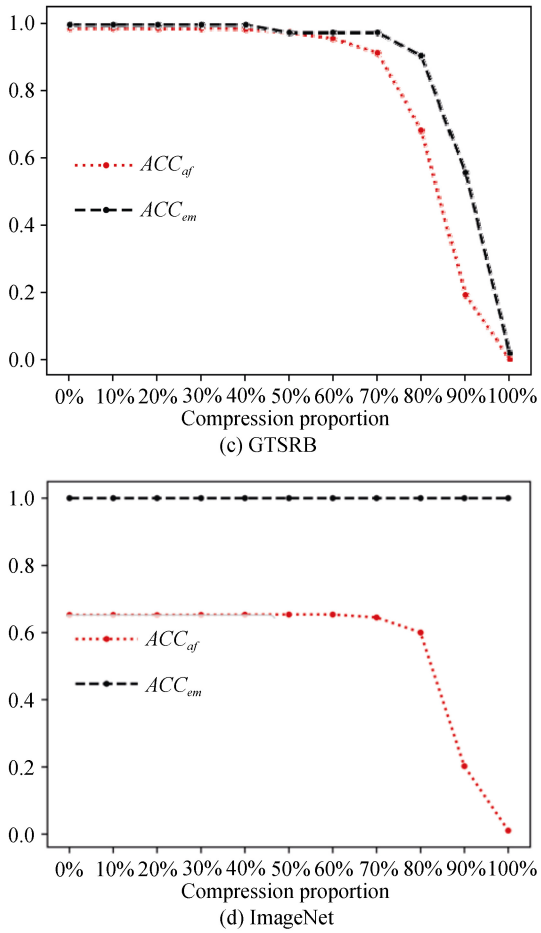


图 4 模型压缩攻击

Figure 4 Model compression attack

ACC_{af} 最终都会受到影响, 不会出现 ACC_{em} 很低而 ACC_{af} 保持不变的情况, 并且 ACC_{af} 比 ACC_{em} 更易受到影响。因为相比于识别水印触发样本, 目标模型处理分类任务需要更多的参数。

5.2.2 模型微调攻击下的评估

训练新模型, 需要大量的数据和计算资源, 如果能够使用预训练模型, 效率将极大提升。盗版者可能使用少量的相关性较强或大量相关性较弱的新数据, 对其窃取的模型进行微调训练以得到新的模型。实验评估 LNMMWS 水印模型, 在面对模型微调时的稳定性。

文献[9]原实验中使用 MNIST 和 CIFAR10 数据集中的一半测试集进行各自模型的微调训练及测试, 这导致测试集样本数减少, 训练模型的原始训练集和微调训练集有较强的关联性。如此设置实验容易导致过拟合, 其在 MNIST 数据集上得到的 ACC_{af} 99.6%和 ACC_{em}^{af} 99.95%不具有实际意义。在实际场景中, 不易获得与原训练数据关联性较强的数据, 为此将 LNMMWS 水印的 ImageNet 数据集目标模型使用 CIFAR10 和 CIFAR100 进行微调训练。CIFAR100 与 CIFAR10 相似, 但 CIFAR100 类别更多, 共有 100 个类别, 50000 张训练图像, 10000 测试图像。

实验结果如表 4 所示, 可以看出 ACC_{em}^{be} 高于 ACC_{af}, 并且 ACC_{em}^{be} 受模型微调的影响较小。从 ImageNet 迁移到 CIFAR10 数据集, ACC_{em}^{be} 保持不变。从 ImageNet 迁移到 CIFAR100 数据集, N_r^{be} 保持不变, ACC_{em}^{be} 降低了 2%。以上实验结果是因为, LNMMWS 水印独立性强与目标模型关联弱, 模型微调对 LogoNet 影响较小。

5.3 隐秘性实验及分析

LNMMWS 方案对目标模型添加的水印应当是隐秘的, 即不应被一些触发模式检测方法检测到。如果被检测到, 就会增加水印被移除的风险。如下使用检测方法[19-20]评估 LNMMWS 方案的隐秘性。

5.3.1 评估 Neural Cleanse 方法检测

实验使用 Neural Cleanse 方法[19]来检测未知 DNN 模型是否含有水印。Neural Cleanse 方法使用 *AI* 评估模型是否异常, Neural Cleanse 方法将 *AI* 的阈值设定为 2, 即 *AI* 大于 2 的模型认定为异常模型, 否则认为模型正常。

表 4 模型微调攻击

Table 4 Model tuning attack

数据集	ACC_{af}	ACC_{em}^{be}	ACC_{em}^{af}	N_r^{be}	N_r^{af}
CIFAR10	87.65%	100%	100%	100	10
CIFAR100	67.81%	100%	98%	100	100

由于并没有在ImageNet数据集上得到文献[9]模型, 所以没有测试 ImageNet 文献[9]模型的异常指数, 其他模型的异常指数, 如图 5 所示。Clean 指未添加水印的干净模型、LNMMWS 指本方案水印模型、Baseline 指文献[9]水印模型。可以看出相比于文献[9], LNMMWS 方案所水印的模型更不容易检测到。LNMMWS 方案更加隐秘的原因是, LogoNet 只对特定的输入有反应, 并且通过抗噪训练, 提高了 LogoNet 的抗干扰能力。

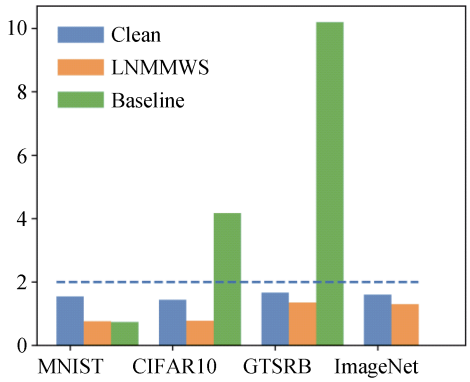


图 5 *AI*

Figure 5 Anomaly index

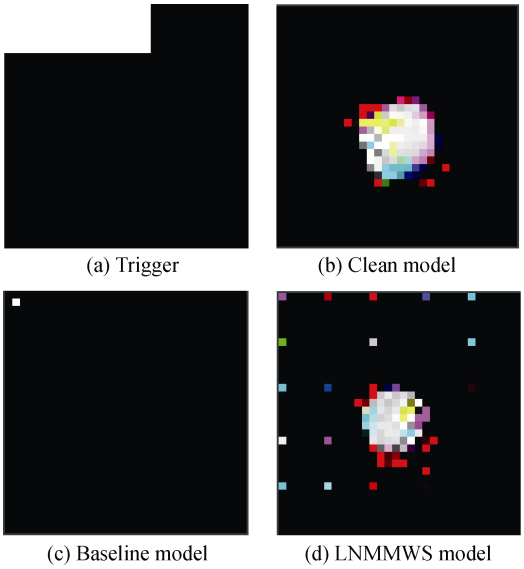


图 6 文献[19]方法逆向生成的触发图案

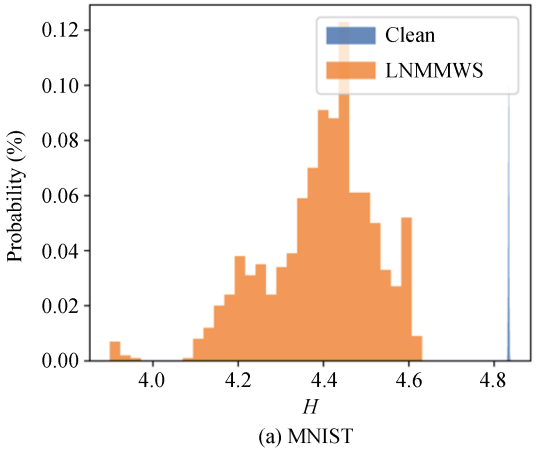
Figure 6 The trigger pattern of literature [19] approach reverse generate

图 6 是使用 Neural Cleanse 方法对在 GTSRB 数据集上的 Clean、LNMMWS、Baseline 模型逆向生成的触发图案。图 6(a)是 Baseline 模型的触发器样本中所嵌入的触发图案，即包含该图案的样本都会被预测为目标类别，在该模型中目标类别被指定为 7。图 6(c)是对 Baseline 模型使用 Neural Cleanse 方法逆向生成的触发图案，可以看出其与图 6(b) Clean 模型逆向生成的触发图案差别很大。对于 Baseline 模型，其类别 7 所包含的水印触发模式可以被 Neural Cleanse 方法检测到，并被认定为异常类别，这于图 5 中 GTSRB 数据集的 Baseline 模型有较高的异常指数相对应。而对于 LNMMWS 模型，针对类别 7 逆向生成的触发图案如图 6(d)所示，可以看出其与图 6(b)相似度较高。事实上 LNMMWS 水印模型的每个标签都包含了特定的水印触发模式，并且它们都不会被 Neural Cleanse 方法检测到。以上实验证实了相比于文献[9], LNMMWS 方案的隐秘性更强。

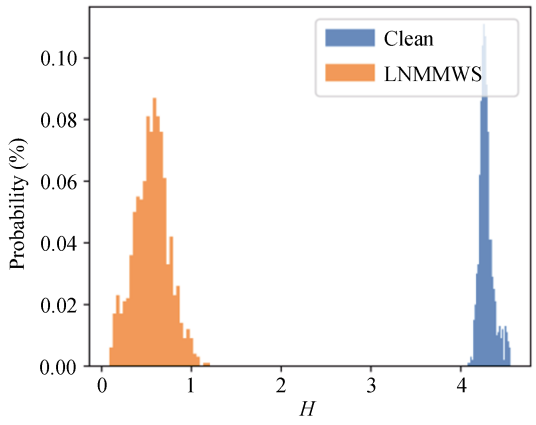
5.3.2 评估 Strip 方法检测

模型具有良好性能通常是指其对正常的样本具有较高的精度，但是其在对抗性样本上会预测错误并且错误是随机的，Strip 方法[20]用 H 描绘这种随机性。制作对抗性样本并对每个模型绘制其多组预测结果 H 的分布。M 个样本预测结果 H 的计算公式如公式 3 所示。

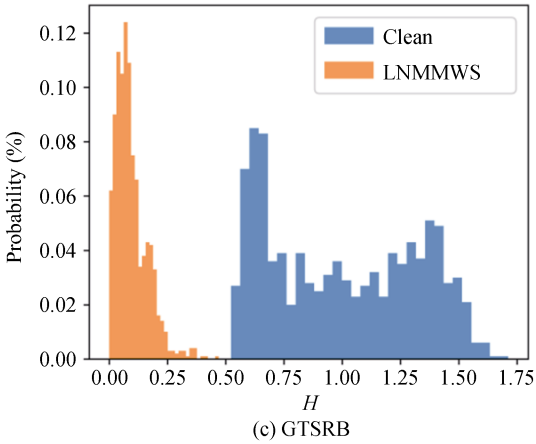
$$H=-\sum_{i=1}^{M} y_i \times \log_2 y_i \tag{3}$$



(a) MNIST



(b) CIFAR10



(c) GTSRB

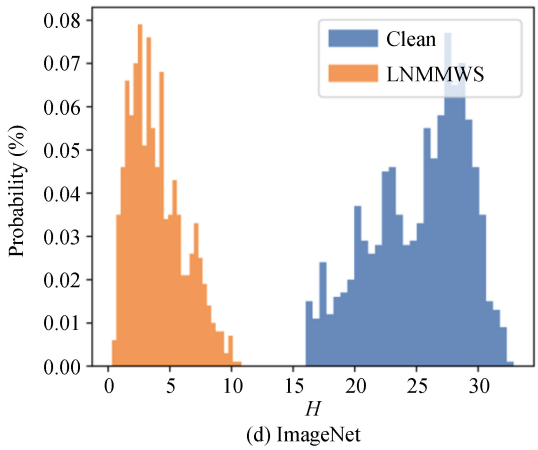


图 7　对抗性样本的 *H* 分布

Figure 7　H distribution of adversarial samples

如图 7 所示, Clean 指未水印模型, LNMMWS 指本方案水印模型。每个任务的不同模型 *H* 分布有所不同, 是因为相同任务不同模型的权重参数不同。但这并不会导致已经嵌入 LogoNet 的模型被检测到, 因为盗版者不可能完全获取到 LNMMWS 水印前后模型的全部信息。

6　结论

本文基于多模型水印场景, 提出一种基于标志网络 LogoNet 的深度学习多模型水印方案 LNMMWS。在生成触发集和噪声集上训练, 得到具有较高水印触发模式识别精度、噪声处理能力的 LogoNet。之后将 LogoNet 嵌入多个目标模型中进行水印处理, 使用黑盒水印验证方法以实现所有权验证。实验及分析表明, LNMMWS 在有效性、稳定性、隐秘性三个方面, 获得了较好的精度和更低的时间开销, 能够抵御模型压缩攻击、模型微调攻击, 对某些触发模式检测具有较好的隐蔽性。下一步的研究目标是对如何选用不同目标模型、不同嵌入方式的水印算法形成统一的评估指标, 并对比更多经典水印方案。

参考文献

[1] Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems[EB/OL]. 2016: arXiv: 1603.04467. https://arxiv.org/abs/1603.04467

[2] Collobert R, Kavukcuoglu K, Farabet C. Torch7: A matlab-like environment for machine learning[C]. *BigLearn, NIPS workshop*, 2011.

[3] Jia Y Q, Shelhamer E, Donahue J, et al. Caffe: Convolutional Architecture for Fast Feature Embedding[C]. *The 22nd ACM international conference on Multimedia*, 2014: 675-678.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[C]. *The 25th International Conference on Neural Information Processing Systems - Volume 1*, 2012: 1097-1105.

[5] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.

[6] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[EB/OL]. 2014: arXiv: 1409.1556. https://arxiv.org/abs/1409.1556

[7] Fan L, Ng K W, Chan C S, et al. DeepIPR: Deep Neural Network Ownership Verification with Passports[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(10): 6122-6139.

[8] Adi Y, Baum C, Cisse M, et al. Turning your Weakness into a Strength: Watermarking Deep Neural Networks by Backdooring[C]. *The 27th USENIX Conference on Security Symposium*, 2018: 1615-1631.

[9] Zhang J L, Gu Z S, Jang J, et al. Protecting Intellectual Property of Deep Neural Networks with Watermarking[C]. *The 2018 on Asia Conference on Computer and Communications Security*, 2018: 159-172.

[10] Guo J, Potkonjak M. Watermarking Deep Neural Networks for Embedded Systems[C]. *2018 IEEE/ACM International Conference on Computer-Aided Design*, 2018: 1-8.

[11] Guo J, Potkonjak M. Evolutionary Trigger Set Generation for DNN Black-Box Watermarking[EB/OL]. 2019: arXiv: 1906.04411. https://arxiv.org/abs/1906.04411

[12] 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]. *The 35th Annual Computer Security Applications Conference*, 2019: 126-137.

[13] *Le Merrer E, Pérez P, Trédan G. Adversarial Frontier Stitching for Remote Neural Network Watermarking[J]. Neural Computing and Applications*, 2020, 32(13): 9233-9244.

[14] *Xue M F, Wu Z Y, He C, et al. Active DNN IP protection: A novel user fingerprint management and DNN authorization control technique[C]. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications*, 2021: 975-982.

[15] Maung A P M, Kiya H. Piracy-Resistant DNN Watermarking by Block-Wise Image Transformation with Secret Key[C]. *The 2021 ACM Workshop on Information Hiding and Multimedia Security*, 2021: 159-164.

[16] Li M, Zhong Q, Zhang L Y, et al. Protecting the intellectual property of deep neural networks with watermarking: The frequency domain approach[C]. *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications*, 2021: 402-409.

[17] Xue M F, Sun S C, Zhang Y S, et al. Active Intellectual Property Protection for Deep Neural Networks through Stealthy Backdoor and Users' Identities Authentication[J]. *Applied Intelligence*, 2022: 1-15.

[18] Jia H R, Choquette-Choo C A, Chandrasekaran V, et al. Entangled

Watermarks as a Defense Against Model Extraction[EB/OL]. 2020: arXiv: 2002.12200. https://arxiv.org/abs/2002.12200

[19] Wang B L, Yao Y S, Shan S, et al. Neural cleanse: identifying and mitigating backdoor attacks in neural networks[C]. *2019 IEEE Symposium on Security and Privacy*, 2019: 707-723.

[20] Gao Y S, Xu C G, Wang D R, et al. STRIP: A Defence Against Trojan Attacks on Deep Neural Networks[C]. *The 35th Annual Computer Security Applications Conference*, 2019: 113-125.

[21] 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.

[22] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. 2009.

[23] Yu F, Wang D Q, Shelhamer E, et al. Deep layer aggregation[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 2403-2412.

[24] Stallkamp J, Schlipsing M, Salmen J, et al. Man Vs. Computer: Benchmarking Machine Learning Algorithms for Traffic Sign Recognition[J]. *Neural Networks*, 2012, 32: 323-332.

[25] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009: 248-255.

[26] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 2261-2269.

[27] See A, Luong M T, Manning C D. Compression of Neural Machine Translation Models via Pruning[EB/OL]. 2016: arXiv: 1606.09274. https://arxiv.org/abs/1606.09274

刘伟发 于 2021 年在临沂大学软件专业获得学士学位。现在河北科技大学计算机科学与技术专业攻读硕士学位。研究领域为人工智能安全。研究兴趣包括: 深度学习后门攻防、深度学习产权保护。Email: moxueliu@163.com

张光华 于 2014 年在西安电子科技大学信息安全专业获得博士学位。现任河北科技大学网络空间安全专业教授。研究领域为网络与信息安全。研究兴趣包括: Web 安全、漏洞挖掘、安全数据分析。Email: zhanggh@hebust.edu.cn

杨婷 于 2021 年在河北科技大学通信与信息系统专业获得硕士学位。现在西安电子科技大学网络空间安全专业攻读博士学位。研究领域为物联网安全。研究兴趣包括: 物联网设备安全分析、物联网协议安全。Email: yangt@nipc.org.cn

王鹤 于 2016 年在西安电子科技大学信息安全专业获得博士学位。现任西安电子科技大学网络与信息安全学院讲师。研究领域为应用密码、量子密码协议。研究兴趣包括: 威胁信息交换共享、量子密码协议。Email: hewang@xidian.edu.cn