计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200204

人工智能模型水印研究综述

谢宸琪　张保稳　易　平
上海交通大学网络空间安全学院　上海 200240
(deadlyone@sjtu.edu.cn)

摘　要　近年来人工智能迅速发展,被用于语音、图像等多种领域,并取得了显著效果。然而,这些训练好的人工智能模型非常容易被复制并扩散,因此,为了保护模型的知识产权,关于模型版权保护的一系列算法或技术应运而生,其中一种就是模型水印技术。通过模型水印技术,向人工智能模型植入水印,一旦模型被窃取,可以通过验证水印来证明自己的版权所有权,维护自己的知识产权,从而达到保护模型的作用。该类技术在近年来成为了一大热点,但目前尚未形成较为统一的框架。为了更好地理解,总结了现阶段模型水印的研究成果,论述了当前主流的模型水印算法,分析了模型水印研究方向的研究进展,还复现了其中几种典型算法并进行了比较,最后提出了未来可能的研究方向。

关键词:模型水印;人工智能安全;信息冗余;算法流程;算法性能比较
中图法分类号　TP393

Survey on Artificial Intelligence Model Watermarking

XIE Chen-qi, ZHANG Bao-wen and YI Ping
School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract　In recent years, with the rapid development of artificial intelligence, it has been used in voice, image and other fields, and achieved remarkable results. However, these trained AI models are very easy to be copied and spread. Therefore, in order to protect the intellectual property rights of the models, a series of algorithms or technologies for model copyright protection emerge as the times require, one of which is model watermarking technology. Once the model is stolen, it can prove its copyright through the verification of the watermark, maintain its intellectual property rights and protect the model. This technology has become a hot spot in recent years, but it has not yet formed a more unified framework. In order to better understand, this paper summarizes the current research of model watermarking, discusses the current mainstream model watermarking algorithms, analyzes the research progress in the research direction of model watermarking, reproduces and compares several typical algorithms, and finally puts forward some suggestions for future research direction.

Keywords　Model watermarking, Artificial intelligence security, Information redundancy, Algorithm flow, Algorithm performance comparison

1　引言

近年来,人工智能技术发展迅速[1-3],由于其出色的性能,在多个领域内都得到广泛应用,如医疗[4-5]、生物[6-7]、金融[8]、自动驾驶[9-10]、图像识别[11-12]、自然语言处理[13-14]等领域,与此同时,人工智能安全也引起人们的关注[15-16]。人工智能模型是一种非常有价值的知识产权资产,是利用用户最有价值的数据去训练的,如金融交易、医疗信息、用户交易信息等,这些模型是研究人员经过数月甚至数年努力之后开发训练出来的,但是人工智能模型非常容易被窃取[17-18]。在不远的将来,将有更多的数据信息会通过人工智能模型的形式保存下来,用于完成各种任务,而如何保障这些数据的安全、维护拥有者的权益,成为前沿工作者不得不考虑的问题。

为了保护人工智能模型的知识产权[19],研究者提出了模型水印的概念。模型水印相当于将数字水印[20-21]的技术和思想应用于人工智能领域,将水印植入模型,需要时可通过各种信息处理方式提取出水印信息,从而达到确认版权者的目的,以保护知识产权。

人工智能模型水印由 Uchida 等[22]首次提出,利用投影矩阵,将版权信息嵌入权值矩阵中。随后,以该方法为基础,衍生了多种投影矩阵水印[23-25]。针对水印提取阶段需要白盒情况的局限性,Adi 等[26]提出了黑盒的后门植入形式的水印。随后,Fan 等[27]引入了混淆攻击的概念,并针对之前水印无法抵御该攻击的缺陷进行了改进。在这期间,围绕人工

到稿日期:2020-12-23　返修日期:2021-03-19　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2020YFB1807504,2020YFB1807500)
This work was supported by the National Key Research and Development Project of China(2020YFB1807504, 2020YFB1807500).
通信作者:易平(yiping@sjtu.edu.cn)

智能模型水印的工作不断涌现，出现了多类方式，对人工智能模型水印各类算法进行总结和分类，在有助于理解各种算法原理的基础上找到规律，产生新的灵感。

本文将对目前的模型水印研究现状做一个全面的介绍，按照其生成方式和原理的不同进行分类，介绍其原理并分析各种模型水印算法的优缺点，在此基础上探讨模型水印的研究方向以及未来的发展趋势。

本文第 1 节进行了背景介绍；第 2 节主要介绍了模型水印的基本概念和基础知识，包括模型水印本身的定义、其延伸的相关概念以及基本操作流程；第 3 节说明了模型水印的存在性来源，同时介绍了模型水印常用的性能评估指标，并且介绍了针对模型水印可能存在的攻击；第 4 节介绍了各类模型水印的原理以及优缺点，同时总结了模型水印实验中常用的数据集；第 5 节选取第 4 节中介绍的具有代表性的方法进行了实验，结合相同数据集的测试效果来说明各方法的优缺点；第 6 节讨论了该领域未来研究的发展方向；最后总结全文。

2 简介

2.1 模型水印的概念

Uchida 等[22]首次尝试向深度神经网络添加水印。模型水印指利用数字水印技术，通过某种通用框架向模型中植入水印，以检测训练好的模型是否被侵权的技术。

2.2 模型水印的相关概念

(1)数字水印[28]：在不影响数字载体使用价值的情况下通过将标识信息嵌入数字载体中，来达到版权保护作用的技术。

(2)冗余：信息中包含的、不影响信息完整的信息。

(3)模型剪枝[29]：一种常用的模型压缩方法，可以减小人工智能模型的规模，以提高推断效率。

(4)微调[30]：微调是迁移学习的一种常用技术，通过对现有网络进行训练，以达到快速训练模型的效果。

(5)鲁棒性[31]：指模型水印在一定的攻击扰动情况下仍能维持其性能的特性。

2.3 基本操作流程

模型水印的生成方法有多种，但归纳起来有一个总体的框架。模型水印的生成、植入与提取过程如图 1 所示。

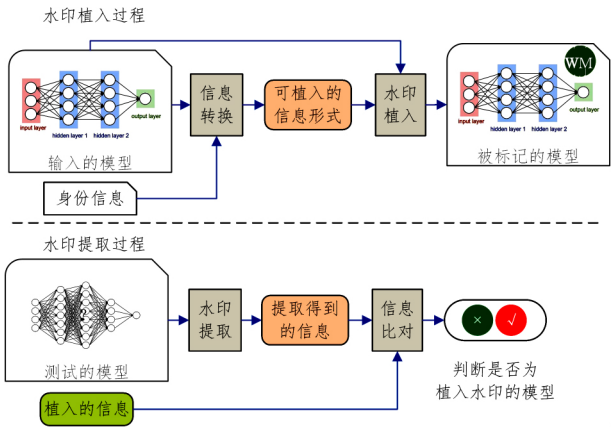


图 1 模型水印流程示意图

Fig. 1 Flow chart of model watermark

模型水印的植入过程可分为：

(1)根据身份信息和输入的模型进行信息转换，将身份信息转换成为可植入的信息形式；

(2)将生成的可植入的信息植入模型中。

模型水印的提取过程可分为：

(1)对测试的模型进行水印提取操作，从而得到相应的信息；

(2)将要验证的版权者对应的植入信息与提取得到的信息进行比对，判断其是否为植入该水印的模型。

3 模型水印相关分析

3.1 模型水印的存在性原理

数字水印技术[28]是在不影响数字载体使用价值的情况下，通过将标识信息嵌入数字载体中，从而实现版权保护作用的技术。而这一植入过程中，人眼对每个像素的细微变化并不敏感，对于人眼来说，图片像素的细微变化不会引起视觉上的较大差异，而这一部分像素空间的可变化区间便是对于人眼看图的冗余。当我们将信息嵌入这一部分中时，并不会影响人眼的观察，但实现了数字水印的功能。换言之，数字水印技术利用的便是图像对于人眼的像素冗余空间。

模型水印是数字水印技术在人工智能领域的扩展，因此模型水印的存在性也与冗余有着密切关系。

人工智能模型作为一种新型的数字载体，承载了大量的数据信息，这一点与数字水印植入的载体特点相同。而对同样一个分类问题，模型参数的细微变化，对模型的分类准确率并不会造成明显影响。而对于模型输出分类来说，这一部分的参数空间的可变化区间就是模型参数空间的冗余[32]。因此，我们可以将信息嵌入这部分中，而不影响模型本身的性能，从而实现模型水印的预期功能。换言之，模型水印的存在性就是依赖于模型参数空间内所存在的冗余。

3.2 模型水印评估方法

对人工智能模型水印的评估有不同的标准，如保真度、有效性、鲁棒性等[22]。

(1)保真度(fidelity)：指植入水印后，模型本身的准确率变化情况。

(2)有效性(effectiveness)：指能够正确提取得到的水印信息占水印总信息的比例。

(3)鲁棒性(robustness)：对于不同的攻击方式，模型水印能够保持的有效性。

(4)效率(efficiency)：对某个模型进行水印提取的计算开销。

(5)信息容量(capacity)：水印中可携带的信息量。

(6)安全性(security)：水印通常保密，并且不会被未授权组织读取或者修改。

3.3 针对模型水印的攻击方式

对于加入模型的模型水印有多种攻击方式，而最常考查的攻击方法主要有 3 种：模型剪枝、模型微调、水印覆写。

(1)模型剪枝。人工智能模型从卷积层到全连接层存在着大量的冗余参数，大量的神经元激活值趋近于 0，如果将这些神经元去掉，模型依然可以维持同样的准确率，这一过程对应的技术就是模型剪枝。而植入水印的过程，本质上就是利用模型参数的冗余来植入信息，这将会对模型参数空间产生影响。模型剪枝本身又是一种常用的模型压缩方法，因此，对

于模型水印，模型剪枝是一种需要考虑的攻击方式。

(2)模型微调。当针对某个任务，自己的相应数据不够时，将别人训练好的模型换成自己的数据后，调整参数，再训练一次模型的过程，就是模型微调。因此，模型窃取者很可能将模型进行微调后使用，这样既能够继承原模型的优异表现，在形式上又与之前的模型不同。而在模型微调的过程中，对模型参数空间进行调整可能会对植入的水印有影响，因此也需要列入考虑范围。

(3)混淆攻击。这种攻击方式试图生成一个伪水印化的数据去混淆含有真正模型水印的人工智能模型的版权，其中比较典型的是水印覆写，即向模型中嵌入伪造的水印，混淆之前植入的带有版权信息的水印，使该水印失去唯一性，破坏其植入信息的结构，将植入的水印信息破坏。

4 模型水印研究现状

4.1 修改训练数据加水印的方法

4.1.1 利用后门植入水印的方法

文献[26,33-34]提出，如图2所示，模型持有者可以对图片数据进行处理。首先提取一部分数据作为触发集，在图片上加上特定的噪声或者标志，使得触发集数据中带有版权信息，而在训练过程中，模型持有者将触发集图片对应的输出标签改为特殊标签，例如原分类问题中不会出现的标签，对人工智能模型进行有监督的训练，使得模型学习到这种特定的噪声或标志的特征。而在提取水印时，模型持有者只需要向模型提供触发集的图片以及原图片，当模型的输出为指定的特殊标签以及原本的标签时，说明水印提取验证成功，因为正常的模型不会输出特殊标签，并且水印的植入并不影响原本分类任务的准确性。文献[34]从数据集中选定了一组触发集图片，添加不同的分布模式，并将其赋予错误标签，当输入该组触发集图片，并且模型对该组图片的分类结果与赋予的错误标签一致时，说明水印提取验证成功。

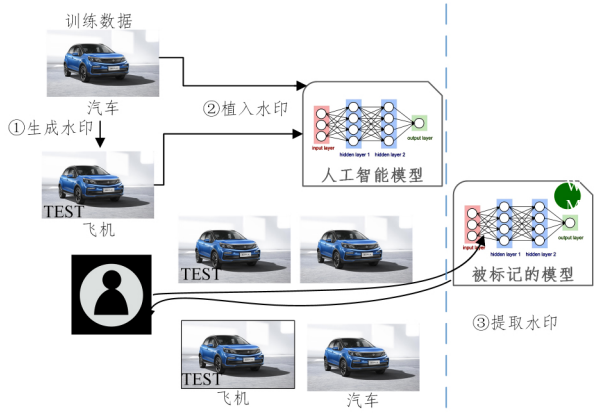


图2 后门植入流程图

Fig. 2 Flow chart of backdoor embedding

文献[33]指出，这种形式的水印可以做到黑盒提取，但无法躲避预测 APIs 的攻击。并且在面对水印覆写的攻击时，即当攻击者对模型的数据也进行相应的水印处理时，水印效果也会受到影响。

针对无法抵御水印覆写的问题，文献[35]提出通过正常嵌入和空嵌入组合的方式进行训练。

选取一类训练集数据，在输入人工智能模型前，根据一层 filter 对数据进行预处理。Filter 尺寸大小与图片一致，由 1，0，−1 组成，若 filter 中数值为 1，则将图片相应位置像素值调整为某一远超限度值的值，本文取其值为 2000；若 filter 中数值为 0，则图片位置像素值相应调整为 −2000；若 filter 中数值为 −1，则图片相应位置的像素值保持不变，并且强制将该种数据经处理后输入人工智能模型的输出标签设置为特殊标签，将这组数据作为训练数据进行训练，使得模型学习到这种远超限度值的数据特征都被分类为特殊标签，从而达到后门植入的效果。这个过程称为正常嵌入，该过程与文献[26]类似。

而空嵌入的操作与正常嵌入相似，区别只是在 filter 层其 filter 内的数值与正常嵌入恰好相反，若正常嵌入过程中 filter 内数值为 1，则空嵌入中数值为 0；若正常嵌入时 filter 为 0，则空嵌入时 filter 为 1；若正常嵌入时 filter 为 −1，则空嵌入时 filter 为 −1。

之后根据 filter 进行处理，而空嵌入时数据的对应标签为正常的原本标签，将这组数据也作为训练数据进行训练，使得模型在空嵌入数据学习到修改部分的图像数据对分类结果没有影响。即模型分类时的像素空间去除了调整过像素值的这部分像素。

这一嵌入方式只有在模型训练之初才有效，否则将破坏模型的分类依据，严重降低模型的准确性。文献[36]的思想与其类似，也是采用空嵌入的方法，根据某种特定模型改变原本图片的像素值到某些极端正值或负值。

这种方法虽然可以解决水印覆写问题，但在植入水印时，必须与模型初始训练同时进行，当需要更换水印信息时需要重新训练。并且文献[26]中的实验尺寸相对局限在一个较小的值，对于尺寸较大的数据集，还未进行有效测试。

4.1.2 利用对抗样本构建水印的方法

对抗样本通过向原始数据添加人为的扰动，使得模型以高置信度输出错误分类的样本[37-39]，而模型可以通过对抗训练调整决策边界，从而正确分类对抗样本。

文献[40]提出可以利用对抗样本构建模型水印，植入水印的过程为：

首先，构建一组模型的成功对抗样本和失败对抗样本，使得这组样本位于决策边界附近。至于如何找到决策边界附近的对抗样本，可以利用 FGSM[41] 的方法，由于 FGSM 是采取逐步添加扰动的方式，因此攻击成功的上一次对抗样本就是在决策边界附近的失败对抗样本。

然后，通过对模型进行对抗训练[42]的方式微调决策边界，使得模型对于成功对抗样本重新分类正确，对于失败对抗样本依然分类正确。而在提取水印时，向模型输入构建的该组对抗样本，若模型对该组对抗样本都分类正确，则说明该模型是植入了该水印信息的模型。

但这种水印算法很大程度上依赖于对抗样本在不同模型之间的迁移性，在迁移过程中，水印的准确率是否会下降是一个有待解决的问题，而且决策边界的调整，对于不同的模型是否依旧适用也有待研究。

4.2 修改模型加水印的方法

4.2.1 利用投影矩阵构建水印的方法

文献[22]提出利用投影矩阵构建水印，植入水印的流程图如图3所示。

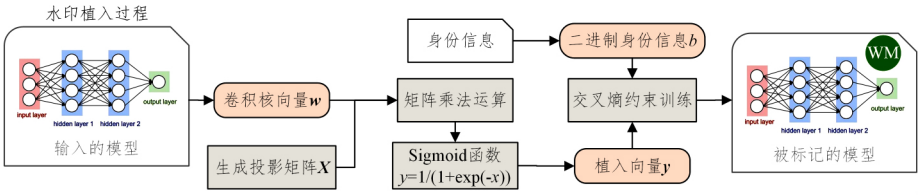


图 3 水印植入流程图

Fig. 3 Flow chart of watermark embedding

首先，将某一层卷积层的各个卷积核的权重值取平均，并展开得到一维卷积核向量 w；

其次，模型持有者生成一个只有自己知晓的投影矩阵 X，本文采用的是正态分布矩阵；

然后，将投影矩阵 X 与卷积核向量 w 相乘后，通过 sigmoid 函数得到植入向量 y，以上过程如式(1)所示：

$$y_j=\sigma(\Sigma_i X_{ji}w_i) \tag{1}$$

最后，将身份信息转换成二进制的身份信息向量 b，将向量 b 与植入向量 y 的交叉熵作为正则项，如式(2)所示：

$$E_R(\boldsymbol{w})=-\sum_{j=1}^{T}(b_j(\log(y_j)+(1-b_j)\log(1-y_j)) \tag{2}$$

通过将该项作为正则项加入损失函数中，来约束训练过程，不断调整卷积核的权重值，使得植入向量 y 不断接近身份信息向量 b，如式(3)所示：

$$E(w)=E_0(w)+\lambda E_R(w) \tag{3}$$

其中，$E_0(w)$为原本的损失函数，λ 为调节因子，用于平衡正常训练和约束正则项的程度。

提取水印的流程如图 4 所示。

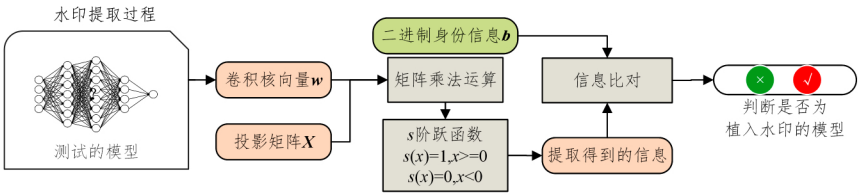


图 4 水印提取流程图

Fig. 4 Flow chart of watermark extraction

模型植入水印的步骤与之前类似，首先得到卷积核向量 w，并将 w 与投影矩阵 X 相乘，将结果通过阶跃函数得到一组 0，1 构成的比特序列，将其与之前的身份信息向量 b 进行对比，即可确认其是否为植入了水印的模型。这个过程如式(4)所示：

$$b_j=s(\Sigma_i X_{ji}w_i) \tag{4}$$

文献[23-25]的思想与文献[22]类似，取某卷积层权重值取平均再拉伸后的向量 w 与投影矩阵 X 的乘积，将其与水印比特信息的均方误差作为正则项并添加到损失函数中，约束模型参数，只是进行了一定的改进，如使用密码本进行存储等。这 3 篇文献虽然保持了模型识别的高准确度，使水印具有一定的鲁棒性，但未能很好地解决水印覆写问题。

4.2.2 添加模型结构构建水印的方法

文献[27]提出，在卷积神经网络的卷积层和激活层之间添加一层 passport 层结构。

该 passport 层将上一层卷积层的输出作为该层的输入，通过与本层的参数进行相应运算，将相应的特征激活输出到下一层的激活层，这一过程如式(5)所示：

$$O^l(\boldsymbol{X}_p)=\gamma^l\boldsymbol{X}_p^l+\beta^l=\gamma^l(\boldsymbol{W}_p^l\star\boldsymbol{X}_c^l)+\beta^l \tag{5}$$

其中，$\boldsymbol{X}_p$ 表示该 passport 层的输入，即上一层卷积层的输出，$\boldsymbol{X}_c$ 表示卷积层的输入，$\boldsymbol{W}_p$ 表示卷积核，而 γ^l 和 β^l 分别为比例因子和偏差项，分别如式(6)、式(7)所示：

$$\gamma^l=\mathrm{Avg}(\boldsymbol{W}_p^l\star P_\gamma^l) \tag{6}$$

$$\beta^l=\mathrm{Avg}(\boldsymbol{W}_p^l\star P_\beta^l) \tag{7}$$

其中，$P^l\gamma$ 和 P_β^l 为 passport 层参数，用于生成比例因子和偏差项。

而当模型 passport 层采用伪造的 passport 时，由于参数与训练时明显不同，模型的性能会大幅下降，而植入的水印就相当于一种易损水印，一旦水印信息被破坏，原信息载体的功能将丧失。

这种水印能够抵御水印覆写、微调和剪枝攻击，但实现起来较为复杂且成本较高。

4.2.3 利用聚类构建水印的方法

文献[43]用聚类方式将训练集中的各类图片输入后得到的输出激活分为两类，编码为 0，1。针对 K 比特的密钥信息，选取一组对抗样本图片作为密钥，向损失函数添加正则项，通过训练使得模型对这组图片分类正确。然后构建 3 个未被标记过的相同模型，对这些对抗样本图片进行测试，找出 3 个模型分类结果都错误的图片集，取交集的部分作为密钥备选。激活一组特定图片的输出，按该种聚类方式，每张图片对应的分类是确定的。

验证时，输入特定的一组图片，当模型对图片分类正确时，其对应的输出激活聚类结果即为提取出的水印信息。

4.2.4 利用对抗网络训练的方法

文献[44]的思想是，对于一个人工智能模型 X，利用一个人工智能模型 A 对模型 X 输出，相应提取水印信息，用另一个人工智能模型 B 来判断是否植入水印，将植入水印的模型参数作为输入，两个模型 A 和 B 分别进行水印信息的提取和对是否存在水印进行判断，将 A 和 B 的输出反向传播，反馈给原本的人工智能模型 X，放在一起做对抗训练，逐步提高提取网络对水印信息的提取能力。通过不断的训练，可以使 X

输出的训练集上的样本与正常训练集上的样本的分布相近，不容易被区分，并且使水印信息提取网络对应的提取能力不断上升，提高水印本身的可靠性。

文献[45]和文献[46]的思想与其相近，都采用了对抗训练的思想来使嵌入水印的信息与正常信息相近，并不断提高水印的提取能力。但这种水印方法的构建方式较为复杂，需要构建多个人工智能模型进行训练，所需的时间成本、技术复杂度和计算复杂度均较高，实现起来较为麻烦。

5 水印算法比较

5.1 水印算法的优缺点对比

本文将上述模型水印算法进行对比分析，结果如表 1 所列。

表 1 模型水印算法的对比

Table 1 Comparison of model watermarking algorithms

水印算法	优点	缺点
后门植入水印	验证方便，只需要输入特定图片，就可以进行黑盒提取	无法抵御预测 APIs 的模型窃取攻击
利用对抗样本构建水印	植入水印后可提高对于对抗样本的鲁棒性，可进行黑盒提取	对抗样本在不同模型上的迁移性不确定
利用投影矩阵构建水印	过程相对简单，无须添加网络结构	无法抵抗水印覆写
在模型中添加结构	对水印覆写具有一定的鲁棒性	模型结构相对复杂，参数量增加
利用聚类将图片按输出激活分类编码	能够提升模型对于对抗样本的鲁棒性	密钥集构建烦琐，过程相对复杂
利用对抗网络训练	输入图片的分布与正常图片相近，不易被筛出	多个网络同时进行训练，工作量大

5.2 水印算法复现实验设置

为了实现方便，聚类编码和对抗网络训练的方式工作烦琐，因此本文选取投影矩阵水印算法（正则项水印）、后门植入水印算法（特殊标签水印）、对抗样本水印算法、添加结构水印算法（passport 层水印）4 类算法进行了复现。

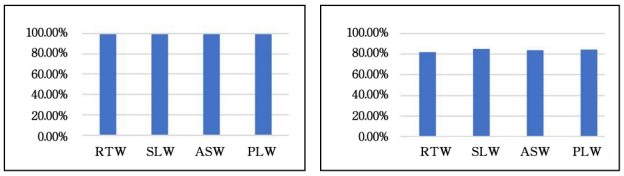
本文采用了 MNIST 和 CIFAR10 数据集，MNIST 数据集是手写体数字识别的像素 28×28 的 10 类灰度图数据集，有 60 000 张训练图片和 10 000 张测试图片，图片像素值为 0～255。CIFAR10 数据集包括 60 000 张像素 32×32 的 100 类彩色图片，每类图片有 6 000 张，图片被分为 50 000 张训练图片和 10 000 张测试图片。

对于 MNIST 数据集，我们使用 LeNet5，对于 CIFAR10 数据集，我们使用 VGG11，而对抗样本水印算法由于在复现时在 VGG11 上的测试效果不理想，因此改用 VGG16 的结构进行测试。在训练原始模型的过程中使用 SGD 的梯度更新算法、Nesterov Momentum 和交叉熵。初始学习率设置为 0.1，不设置权重衰减，动量设置为 0.9，最小的 batch 设为 100，学习率每 20 轮衰减到原来的 3/10，训练 100 轮。

在剪枝过程中采用的是稀疏剪枝的方式，分别选择保留 20%，50%，80%的参数。在剪枝重训练的过程中，初始学习率为 0.01，每 5 轮衰减至原来的 3/10，共训练 20 轮。在微调的过程中，本文将得到的模型在测试集上进行微调，初始学习率设置为 0.1，每 10 轮衰减到原来的 3/10，分别训练 20，50，100 轮。

5.3 水印算法复现实验结果

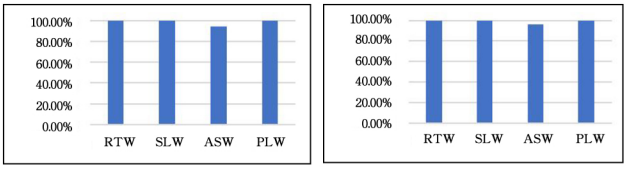
本文对其中的一些性能指标进行了测试，结果如图 5—图 8 所示，各类别代指的水印算法为：类别 1 为正则项水印（Regular Term Watermarking，RTW）、类别 2 为特殊标签水印（Special Label Watermarking，SLW）、类别 3 为对抗样本水印（Adversarial Samples Watermarking，ASW）、类别 4 为通行证层水印（Passport Layer Watermarking，PLW）。



（a）不同水印方法在 MNIST 上的保真度　（b）不同水印方法在 CIFAR10 上的保真度

图 5 不同水印方法的保真度情况图

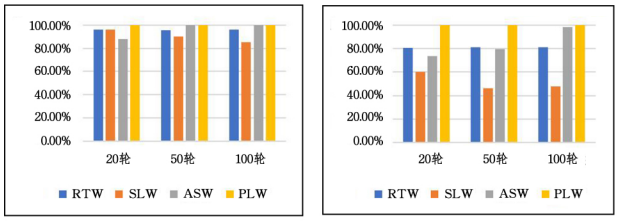
Fig. 5 Fidelity of different watermarking methods



（a）不同水印方法在 MNIST 上的可靠性　（b）不同水印方法在 CIFAR10 上的可靠性

图 6 不同水印方法的可靠性情况

Fig. 6 Reliability of different watermarking methods



（a）不同水印方法对微调的鲁棒性（MNIST）　（b）不同水印方法对微调的鲁棒性（CIFAR10）

图 7 不同水印方法对于微调的鲁棒性情况

Fig. 7 Robustness of different watermarking methods for fine tuning

本文选取的 LeNet5，VGG 在 MNIST 和 CIFAR10 上的准确率为 98.5%，82%。由图 5 可知，加入上述 4 种水印后，模型的准确率波动范围很小，在 MNIST 上变化最大的是准确率提升了 0.73%，而在 CIFAR10 上变化最大的是准确率提升了 2.49%，因此模型本身的准确率在加入水印前后并不受影响。

由图 6 可知，针对这 4 种水印，在提取时均达到了一个较高的可靠性，但是相比而言，对抗样本水印算法（ASW）的可靠性较差，无法做到将水印的信息完全提取出来。

4 种水印对于微调的鲁棒性的测试数据如图 7 所示。

本文对植入水印后的模型分别进行了微调操作，进行了 20，50，100 轮的微调，学习率初始为 0.01，每 10 轮衰减至原来的 3/10，得到了如图 7 所示的实验数据。从图中可以看出，对于微调操作，特殊标签水印（SLW）算法的可靠性受到

的影响最大，水印的可靠性从开始的100%下降到了46%，而正则项水印(RTW)和对抗样本水印(ASW)均不同程度地受到了影响，水印的可靠性明显下降，而通行证层水印(PLW)的可靠性维持在100%，不受微调的影响。

4种水印对于剪枝的鲁棒性的测试数据如图8所示。

(a)不同水印方法对剪枝的鲁棒性(MNIST)

(b)不同水印方法对剪枝的鲁棒性(CIFAR10)

图8 不同水印方法对于剪枝的鲁棒性情况

Fig. 8 Robustness of different watermarking methods to pruning

从图中可以看出，对于剪枝操作，正则项水印(RTW)和对抗样本水印(ASW)均受到了影响，正则项水印(RTW)的可靠性下降明显，降低至85.2%，对抗样本水印(ASW)也下降到91.11%，而特殊标签水印(SLW)和通行证层水印(PLW)依旧维持着100%的可靠性。对抗样本水印(ASW)算法中的水印总体上与模型保留的参数量有关，保留参数越少，水印的性能下降就越快。

综上，由图5—图8可知，模型中加入4种水印的任意一种后，模型本身的准确率不受明显影响。就水印本身的可靠性而言，均具有较高的可靠性，但特殊标签水印对于微调的鲁棒性较弱，正则项水印对于剪枝的鲁棒性较弱，对抗样本水印对于剪枝、微调的鲁棒性都不强，而通行证层水印对微调和剪枝都具有很高的鲁棒性。

6 前景与展望

人工智能模型水印的研究目前是一个比较新的领域[47-49]，并且不断有新的想法被提出，未来模型水印研究在如下几个方面可以有进一步的发展。

(1)模型水印算法目前主要集中在对模型本身植入水印的方式，而模型参数中包含着大量的冗余，有足够的冗余空间给我们植入水印，并且我们可以结合信息论、编码理论、密码学等，以寻找不同的方式来实现对水印的植入，因而会有更多的可能性。

(2)对于模型水印性能，目前评估的指标并不明确，只有少数几个指标是公认的，但由于不同水印算法采用的训练方式不同、考察的指标不同、实验参数不同等，导致不同研究成果之间难以做出比较，因此对模型水印性能评估的分析需要一个统一的框架和标准。

(3)当前的模型水印算法以白盒水印居多，但在实际应用场景中，黑盒水印的适用面比白盒水印广，对于黑盒水印，需要进行更广泛的研究，以适应更多的应用场景。

(4)模型水印本质上是一种信息隐藏技术，这种技术有可能被用于隐蔽通信，可以通过向模型中植入水印的方式将要传递的信息植入其中。目前来看，虽然这种方式的成本较高，但也是一个可研究的方向。

(5)模型水印本身的存在与模型的冗余密切相关，而对于模型的冗余以及参数的意义，目前还没有确切的解释，可解释性还有待挖掘。参数对于模型的重要程度也是水印操作需要注意的问题，如何找到合适的重要程度衡量指标是我们要探索的方向，因此还需要进行更深入的理解和理论分析。

结束语 近年来，随着人工智能技术的广泛应用，人工智能模型逐渐成为了众多高价值信息的载体，而随之带来的安全问题也开始引起人们的关注。模型水印技术作为一种保护模型信息价值的技术手段，是现在的一个前沿并且重要的领域。本文着重介绍了模型水印的相关研究，包括模型水印的存在性原理、模型水印的评估指标，同时也分析了经典的模型水印算法及其优缺点，并对未来的发展方向进行了探讨，为后续研究提供了参考。

参 考 文 献

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[2] GOODFELLOW I, BENGIO Y, COURVILLE A, et al. Deep learning[M]. Cambridge: MIT press, 2016.

[3] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. Neural networks, 2015, 61: 85-117.

[4] WANG X, YANG W, WEINREB J, et al. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning[J]. Scientific Reports, 2017, 7(1): 1-8.

[5] XIONG H Y, ALIPANAHI B, LEE L J, et al. The human splicing code reveals new insights into the genetic determinants of disease[J]. Science, 2015, 347(6218): 144-153.

[6] WEBB S. Deep learning for biology[J]. Nature, 2018, 554(2): 555-557.

[7] BRANSON K. A deep (learning) dive into a cell [J]. Nature Methods, 2018, 15(4): 253-254.

[8] DENG Y, BAO F, KONG Y, et al. Deep direct reinforcement learning for financial signal representation and trading[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(3): 653-664.

[9] HE Y, ZHAO N, YIN H. Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach [J]. IEEE Transactions on Vehicular Technology, 2017, 67(1): 44-55.

[10] ZHAO D, CHEN Y, LV L. Deep reinforcement learning with

visual attention for vehicle classification[J]. IEEE Transactions on Cognitive and Developmental Systems,2016,9(4):356-367.

[11] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

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

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

[14] CHOWDHARY K. Natural language processing[M]// Fundamentals of Artificial Intelligence. Springer,2020:603-649.

[15] AKHTAR N,MIAN A. Threat of adversarial attacks on deep learning in computer vision:A survey[J]. IEEE Access,2018,6:14410-14430.

[16] CHEN H,WANG F Y. Guest editors' introduction:Artificial intelligence for homeland security[J]. IEEE intelligent systems,2005,20(5):12-16.

[17] JUUTI M,SZYLLER S,MARCHAL S,et al. PRADA:protecting against DNN model stealing attacks[C]// Proceedings of the 2019 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE,2019:512-527.

[18] TRAMÈR F,ZHANG F,JUELS A,et al. Stealing machine learning models via prediction apis[C]// Proceedings of the 25th {USENIX} Security Symposium ({USENIX} Security 16. 2016:601-618.

[19] DAVIES C R. An evolutionary step in intellectual property rights-Artificial intelligence and intellectual property[J]. Computer Law & Security Review,2011,27(6):601-619.

[20] COX I J,MILLER M L,BLOOM J A,et al. Digital watermarking[M]. San Francisco:Morgan Kaufmann,2002.

[21] PODILCHUK C I,DELP E J. Digital watermarking:algorithms and applications[J]. IEEE Signal Processing Magazine,2001,18(4):33-46.

[22] UCHIDA Y,NAGAI Y,SAKAZAWA S,et al. Embedding watermarks into deep neural networks[C]// Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. 2017:269-277.

[23] CHEN H,FU C,ROUHANI B D,et al. DeepAttest:An end-to-end attestation framework for deep neural networks[C]// Proceedings of the 2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA). IEEE,2019:487-498.

[24] CHEN H,ROHANI B D,KOUSHANFAR F. DeepMarks:a digital fingerprinting framework for deep neural networks[J]. arXiv:1804.03648,2018.

[25] ROUHANI B D,CHEN H,KOUSHANFAR F. Deepsigns:A generic watermarking framework for ip protection of deep learning models[J]. arXiv:1804.00750,2018.

[26] ADI Y,BAUM C,CISSE M,et al. Turning your weakness into a strength:Watermarking deep neural networks by backdooring [C]// Proceedings of the 27th {USENIX} Security Symposium. 2018:1615-1631.

[27] FAN L,NG K W,CHAN C S. Rethinking deep neural network ownership verification:Embedding passports to defeat ambiguity attacks [C]// Proceedings of the Advances in Neural Information Processing Systems. 2019:4714-4723.

[28] VAN SCHYNDEL R G,TIRKEL A Z,OSBORNE C F. A digital watermark[C]// Proceedings of 1st International Conference on Image Processing. IEEE,1994:86-90.

[29] LIU Z,SUN M,ZHOU T,et al. Rethinking the value of network pruning[J]. arXiv:1810.05270,2018.

[30] CETINIC E,LIPIC T,GRGIC S. Fine-tuning convolutional neural networks for fine art classification[J]. Expert Systems with Applications,2018,114:107-118.

[31] CHANG C L,HUNG J L,TIEN C W,et al. Evaluating Robustness of AI Models against Adversarial Attacks[C]// Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence. 2020:47-54.

[32] CHENG Y,YU F X,FERIS R S,et al. An exploration of parameter redundancy in deep networks with circulant projections [C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:2857-2865.

[33] ZHANG J,GU Z,JANG J,et al. Protecting intellectual property of deep neural networks with watermarking[C]// Proceedings of the Proceedings of the 2018 on Asia Conference on Computer and Communications Security. 2018:159-172.

[34] NAMBA R,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.

[35] LI H,WILLSON E,ZHENG H,et al. Persistent and unforgeable watermarks for deep neural networks[J]. arXiv:1910.01226,2019.

[36] LI H,WENGER E,SHAN S,et al. Piracy resistant watermarks for deep neural networks[J]. arXiv:1910.01226,2019.

[37] ZHU C,CHENG Y,GAN Z,et al. Freelb:Enhanced adversarial training for natural language understanding[J]. arXiv:1909.11764,2019.

[38] LI L,MA R,GUO Q,et al. Bert-attack:Adversarial attack against bert using bert[J]. arXiv:2004.09984,2020.

[39] SAMIZADE S,TAN Z H,SHEN C,et al. Adversarial example detection by classification for deep speech recognition[C]// ICASSP 2020 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP 2020). IEEE,2020:3102-3106.

[40] LE MERRER E,PEREZ P,TRÉDAN G. Adversarial frontier stitching for remote neural network watermarking[J]. Neural Computing and Applications,2020,32(13):9233-9244.

[41] GOODFELLOW I J,SHLENS J,SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv:1412.6572,2014.

[42] SHAFAHI A,NAJIBI M,GHIASI A,et al. Adversarial training for free![J]. arXiv:1904.12843,2019.

[43] CHEN H, ROUHANI B D, KOUSHANFAR F. BlackMarks: Blackbox Multibit Watermarking for Deep Neural Networks [J]. arXiv: 1904.00344, 2019.

[44] ZHANG J, CHEN D, LIAO J, et al. Model watermarking for image processing networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 12805-12812.

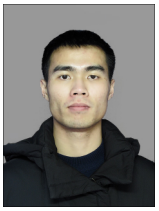
[45] WANG T, KERSCHBAUM F. Robust and Undetectable White-Box Watermarks for Deep Neural Networks [J]. arXiv: 1910.14268, 2019.

[46] LI Z, HU C, 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 the Proceedings of the 35th Annual Computer Security Applications Conference. 2019: 126-137.

[47] YU Y C, DING L, CHEN Z N. Research on attack and defense technology of machine learning system[J]. Netinfo Security, 2018, 213(9): 10-18.

[48] LIU R X, CHEN H, GUO R Y, et al. Privacy attack and defense in machine learning [J]. Journal of Software, 2020(3): 866-892.

[49] CHEN Y F, SHEN C, WANG T, et al. Security and privacy risk of artificial intelligence system [J]. Journal of Computer Research and Development, 2019, 56(10): 111-126.



XIE Chen-qi, born in 1997, postgraduate. His main research interests include artificial intelligence security and so on.

YI Ping, born in 1969, Ph. D, associate professor, is a senior member of China Computer Federation. His main research interests include artificial intelligence security and so on.