计算机科学
COMPUTER SCIENCE

基于鲁棒分区水印的深度学习模型保护方法

吕正浩，咸鹤群

引用本文

吕正浩，咸鹤群. 基于鲁棒分区水印的深度学习模型保护方法[J]. 计算机科学，2026，53(1)：423-429.

LYU Zhenghao, XIAN Hequn. Deep Learning Model Protection Method Based on Robust Partitioned Watermarking [J]. Computer Science, 2026, 53(1): 423-429.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200005

基于鲁棒分区水印的深度学习模型保护方法

吕正浩[1,2]　咸鹤群[1,3]
1 青岛大学计算机科学技术学院　山东 青岛 266071
2 中国科学院信息工程研究所网络空间安全防御重点实验室　北京 100085
3 密码与网络空间安全(黄埔)研究院　广州 510700
(15688778816@163.com)

摘　要　机器学习涉及到昂贵的数据收集和训练成本,模型所有者可能会担心自己的模型遭到未授权的复制或使用,损害到模型所有者的知识产权。因此,如何有效保护这些模型的知识产权成为一个亟待解决的问题。为此,研究人员提出了模型水印的概念。类似于数字水印技术将水印嵌入图像的方式,模型水印通过将特定的标识嵌入机器学习模型中,以达到版权确认的目的。然而,现有的水印方案在实际应用中存在一些局限性。首先,水印的嵌入不可避免地会对模型性能产生一定影响;其次,水印可能会通过微调等技术手段被移除。针对此类问题,提出一种新型的神经网络水印方案,采用区域化和分阶段的嵌入方式。这种方法不仅旨在最大限度地减少对模型性能的影响,还力图提升水印本身的鲁棒性。在 MNIST,CIFAR-10 和 CIFAR-100 数据集上的实验验证了该方案的有效性。实验结果表明,该水印方案在保持水印存活率的同时,对模型性能的影响极小,相较于现有的基线水印方案,模型性能提升幅度最高可达 18 个百分点。此外,所提出的方案对微调等攻击手段表现出较强的鲁棒性,并且不受模型剪枝操作的影响。即便攻击者试图完全移除水印,也必须以显著降低模型性能为代价。

关键词:深度神经网络;模型水印;版权验证;人工智能安全;水印鲁棒性;模型性能
中图分类号　TP309

Deep Learning Model Protection Method Based on Robust Partitioned Watermarking

LYU Zhenghao[1,2] and XIAN Hequn[1,3]
1 College of Computer Science and Technology,Qingdao University,Qingdao,Shandong 266071,China
2 Key Laboratory of Cyberspace Security Defense,Institute of Information Engineering,Chinese Academy of Sciences,Beijing 100085,China
3 Cryphotography and Cyber Security Whampo Institute,Guangzhou 510700,China

Abstract　Machine learning often involves high costs related to data collection and model training,which raises concerns for model owners about unauthorized replication or misuse,potentially infringing on their intellectual property(IP). Consequently,the protection of intellectual property in machine learning models has become a pressing issue. In response,researchers have introduced the concept of model watermarking. Similar to how digital watermarking embeds identifiable marks into images,model watermarking involves embedding unique identifiers into machine learning models to facilitate copyright verification. However,existing watermarking techniques face several limitations in practical applications. Firstly,embedding watermarks inevitably affects model performance to some degree. Secondly,watermarks can be removed through techniques such as model fine-tuning. To address these challenges,this paper proposes a novel neural network watermarking scheme,employing a regional and staged embedding approach. This method not only aims to minimize the impact on model performance but also seeks to enhance the robustness of the watermark itself. Experiments conducted on the MNIST,CIFAR-10,and CIFAR-100 datasets validate the effectiveness of the proposed scheme. The results demonstrate that this watermarking approach maintains a high watermark retention rate while having minimal impact on model performance. Compared to existing baseline watermarking schemes,this method achieves performance improvements of up to 18 percentage points. Additionally,the proposed scheme exhibits strong robustness against attacks such as fine-tuning and remains unaffected by model pruning operations. Even if adversaries attempt to completely remove the watermark,they would have to significantly degrade the model's performance as a trade-off.

Keywords　Deep neural network,Model watermarking,Copyright verification,Artificial intelligence security,Watermark robust-

到稿日期:2024-12-02　返修日期:2025-03-14
基金项目:国家自然科学基金(62102212);网络空间安全防御重点实验室开放课题(2024-ZD-04)
This work was supported by the National Natural Science Foundation of China(62102212) and Open Fund Project of the Key Laboratory of Cyberspace Security Defense(2024-ZD-04).
通信作者:咸鹤群(xianhq@126.com)

ness, Model performance

1 引言

近年来，机器学习[1]，尤其是深度神经网络（DNNs）[2]，在多个领域取得了显著进展[3-5]，通过模拟人脑结构从数据中学习特征，在图像识别[6]、自然语言处理[7]、语音识别[8]和自动驾驶[9]等领域展现出卓越性能，极大地促进了人工智能的发展。然而，随着深度神经网络在各领域的广泛应用，与之相关联的安全问题和知识产权保护也逐渐成为亟待解决的关键议题[10-12]。

首先，深度神经网络模型面临的一个严峻挑战是未经授权的复制和盗取问题[13]。训练高性能的深度神经网络模型通常需要大量计算资源和数据，因此模型本身具有较高的商业价值。然而，一旦模型被公开或部署，攻击者可能通过各种手段获取其参数和结构[14]，进而进行未经授权的复制和使用。此类行为不仅侵犯模型所有者的知识产权，还可能造成严重的商业利益损失。

随着深度神经网络在各行各业的广泛应用，应对安全威胁和保护知识产权的关注度越来越高。保护深度神经网络的知识产权已经成为一项重要的工作，旨在防止未经授权的模型复制、使用和分发，从而维护开发者的合法权利。为了实现这一目标，研究人员提出了多种保护方法，其中模型水印技术是目前广泛研究和应用的一种方法[15]。

模型水印是将数字水印[16]的概念引入机器学习领域，通过在模型中嵌入特定的、难以被察觉的标识信息，来证明模型所有权。这些水印信息可于模型训练过程中嵌入，当模型被窃取或复制时，可以通过特定检测手段提取，为知识产权归属提供证据。这种方法提供了一种强大的手段来阻止非法盗取和使用，并在法律纠纷中提供技术支撑，达到保护模型知识产权的作用。但不可避免的是，水印的嵌入会或多或少地影响模型本身的性能。

具体来说，水印技术通常可以通过在模型中嵌入某些只有模型所有者知道的离群值的输入-输出对，在模型被盗用或复制时，模型所有者可以出示这些离群值数据的预测，并将其与模型输出进行比较，以证明模型所有权。此外，水印技术需要在不显著影响模型性能的条件下，确保水印的鲁棒性。鲁棒性是指水印能够抵抗各种攻击和模型修改手段，包括模型压缩[17]、剪枝[18]、微调[19]以及对抗性攻击等。

人工智能模型水印的概念由 Uchida 等[20]提出，并迅速引起广泛的研究和关注。尤其是后门植入水印（Backdoor-based Watermarking），通过在机器学习模型中植入后门触发器，使模型对特定输入产生预期行为，从而验证所有权。Adi 等[21]首次提出了后门植入水印这一种方法，利用机器学习模型的后门攻击原理，将水印简化为后门设计。这种水印方法可以理解为将离群值输入-输出对过拟合到模型中，而这些离群值输入-输出对即为水印。该方法选择一组特殊的输入作为触发器，这些输入在正常数据分布中为离群的或不常见的，然后为这些触发输入指定特定的输出结果。在训练过程中，将触发输入-输出对加入训练数据，使模型在这些特定输入上过拟合到指定输出。模型部署后，可以通过输入触发样本验证是否包含水印。如果模型在这些输入上产生预期输出，即可确认水印的存在。

然而，模型水印技术在模型版权保护和追踪方面具有应用价值，但也存在缺点和潜在风险。首先，在模型中嵌入水印可能会对模型性能造成负面影响。过多水印样本或不合理的水印设计可能导致模型拟合能力下降，影响其在真实数据集上的泛化能力。此外，水印的安全性和可靠性也是需要重点考虑的因素。由于水印本身可能成为攻击目标[22]，攻击者可能通过破坏或去除水印来使其失效。如果水印设计不够健壮，可能无法抵御此类攻击，从而削弱水印在模型版权保护中的有效性。

两类常见技术进一步加剧了水印失效风险：一是模型微调，该技术通过在特定任务或数据集上的额外训练优化预训练模型适配新需求，但攻击者可利用这一过程抹除原始模型的版权信息，侵犯知识产权；二是深度神经网络剪枝，其通过剔除冗余参数提升计算效率与推理速度，却可能破坏水印依赖的特定参数配置或激活模式，导致水印信息丢失、变形，最终造成水印难以识别甚至完全失效。

针对上述问题，本文提出了一种分区水印技术。其核心思想是通过多个水印集嵌入水印，每个水印集通过微调的方式分阶段嵌入到深度神经网络的特定层或多层中，区别于传统水印技术将水印无差别地嵌入整个模型。分区水印的优势在于可以减少水印对模型拟合精度的负面影响。这是因为深度神经网络模型的浅层通常更倾向于捕捉通用特征[23]，而将水印作为离群值输入-输出对嵌入时，浅层神经元容易受到较大干扰，影响模型整体的拟合能力。传统方法将水印过拟合到整个模型中，会不可避免地对浅层神经元产生较大影响，降低模型的性能。此外，分区水印技术增强了水印面对攻击的鲁棒性。当攻击者对模型进行剪枝和微调时，由于水印位置未知，分区水印可以更好地保持其完整性。在模型性能下降相同的情况下，分区水印的水印存活率更高，这意味着攻击者需要付出更多努力和代价才能成功去除水印。本文在 3 个常用数据集（MNIST[24]，CIFAR-10 和 CIFAR-100[25]）上进行了评估，证明了该水印方案对模型性能的影响更小。在水印存活率相近的情况下，传统基线水印方案（将水印无差别地过拟合到整个模型中）的模型预测准确率最多下降 20 个百分点，而分区水印方案仅下降不到 2 个百分点。此外，与基线方法不同，分区水印方案在面对模型微调和剪枝时展现出更强的鲁棒性，即使被盗模型经过微调或剪枝处理，仍能保持足够的水印存活率，从而有效证明了模型所有权。

总结来说，本文的贡献有：1）较传统水印而言，本文水印方案可以更大程度地保留模型性能；2）本文水印方案能抵御模型微调的影响；3）证明了模型剪枝对本文水印方案无效。

2 相关工作

2.1 DNN 模型

深度神经网络是一种复杂且强大的人工神经网络架构，

其多层隐藏层结构使其能够有效处理和学习复杂的非线性关系。DNN 的基本架构由 3 个主要部分组成，分别是：输入层、隐藏层和输出层。在输入层中，每个节点对应输入数据的一个特征，为后续处理奠定基础。隐藏层作为 DNN 的核心，通过应用各种非线性激活函数（如 ReLU，Sigmoid 或 Tanh），对输入数据进行变换和抽象，从而在捕获数据的复杂模式和特征方面发挥关键作用。输出层负责生成最终的预测结果，将模型学习到的知识转换为可解释的输出形式。DNN 的训练过程包括前向传播、损失函数计算和反向传播，通过梯度下降法优化网络中的权重和偏置，使其在大型数据集上表现优异，并自动学习和提取有价值的特征。然而，DNN 的训练过程通常耗时长且对计算资源需求高，因此确保模型安全至关重要。

2.2 模型微调

模型微调是深度学习领域中的一种常用技术，尤其在迁移学习任务中发挥着重要作用。该技术以预训练模型为基础，通过额外的训练来调整和优化模型，使其适应特定任务的需求。微调过程可概括为以下步骤：首先，选择一个在大型通用数据集上预先训练的模型，该模型已学习到丰富的特征表示；其次，根据目标任务的需求，选择冻结部分层以保留其对通用特征的理解，同时添加自定义层以满足特定任务的输出要求；最后，在目标任务的数据集上进行训练，更新模型参数，使其适应新任务。微调的主要优势在于能缩短训练时间，因为预训练模型已具备一定知识基础。同时，还能提升模型在数据有限情况下的泛化能力，并快速适应与预训练任务相关的新任务。

微调策略的选择至关重要[26]。全量微调（Full Fine-Tuning，FFT）是指更新模型的所有参数。尽管这一方法可以充分利用预训练模型的知识，但也可能导致计算成本过高和增加过拟合的风险。相比之下，参数高效微调（Parameter-Efficient Fine-Tuning，PEFT）是一种更精细的策略，它仅更新部分参数，既能降低计算和存储开销，又能有效保持模型性能。其中，逐层解冻（Gradual Unfreezing）是一种典型的参数高效微调策略。该策略在微调过程中逐渐解冻模型的各层，从冻结状态开始，逐步进行调整和优化。这种方法有助于保持模型稳定性，避免过拟合，同时使模型逐渐适应目标任务的数据分布，实现最佳的微调效果。

2.3 模型剪枝

深度神经网络的模型剪枝是一种优化技术，旨在降低模型的计算复杂度和存储需求，同时尽可能地保持其原始性能[27]。随着深度学习模型规模的不断扩大，尤其是在处理复杂任务时，模型的参数数量和计算量显著增加。这不仅导致计算成本和能耗更高，还可能限制模型在资源受限设备上的应用。模型剪枝通过有选择地移除冗余或不重要的神经元和连接，提高模型效率。

模型剪枝的方法主要分为结构化剪枝和非结构化剪枝两大类。结构化剪枝专注于移除特定的网络结构单元（如卷积核、通道或整个层）。这种方法通常能够更有效地降低计算开销，并易于在现代硬件上实现加速。相对而言，非结构化剪枝则以更细粒度的方式进行，主要通过移除单个权重连接来实现。这种灵活性使剪枝后的模型更为稀疏。在实际加速中，往往需要专门的稀疏矩阵运算库支持。无论采用何种方法，模型剪枝的核心挑战在于减少参数的同时，最大限度地保持模型的预测精度。这通常需要结合剪枝后重新训练等技术手段来实现。

2.4 模型水印

传统水印技术在图像和视频中的主要应用是保护版权和验证内容真实性。水印技术通过在数字媒体中嵌入不可见或难以察觉的标识，能有效防止未经授权的复制和分发。这种技术利用图像处理算法，将水印信息嵌入图像或视频的像素中，在不显著影响视觉质量的同时，确保水印能在各种操作（如压缩、裁剪、旋转等）后保持可检测性。

随着数字媒体的广泛传播和使用，水印技术的应用逐渐扩展到机器学习领域，尤其是在深度神经网络方面。深度学习模型的训练通常需要大量数据，使得模型本身具有与图片、视频类似的知识产权价值，而水印技术的本质在于保护所有者的知识产权。水印技术在图片和视频上的成功应用启发了研究人员，他们开始探索如何将类似的保护措施应用于深度学习模型，以保护这些模型的知识产权。2017 年，Uchida 等[20]率先提出了在深度神经网络中嵌入水印的概念。

模型水印背后的思想是将离群值的输入-输出对过拟合到模型中，这些数据仅为模型所有者所知，如图 1 所示。当模型在预测结果中表现出不可预测的行为（如对离群值输入的响应），而模型所有者能够精确地预测这些行为时，便可声明模型所有者的所有权。Fan 等[28]提出了一种基于 passport 的 DNN 模型，即在神经网络的卷积层和激活层之间插入一层 passport 结构。这种结构能够同时抵御网络修改和模糊攻击，使得原始模型的推理性能因伪造 passport 而显著下降。Chen 等[29]提出了一种黑盒水印方法，该方法以预训练的未标记模型和模型所有者的二进制签名作为输入，并输出带有水印密钥的标记模型。此外，Lyu 等提出了 MEA-Defender[30]和 SSL-WM[31]，这两种方法旨在将水印嵌入自监督模型中。近年来，Liu 等[32]提出了基于外源样本的深度神经网络模型版权保护框架，该方法通过引入外部样本增强水印的鲁棒性，适用于黑盒场景。类似地，Peng 等[33]提出了融合内外部特征水印的模型保护方案，结合内部参数和外部触发器来提升抗攻击能力。

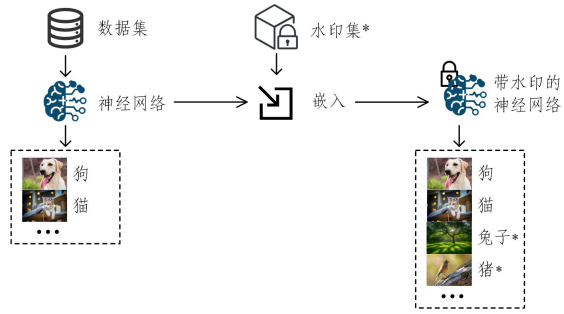


图 1 模型水印的嵌入

Fig. 1 Embedding of model watermarks

如果防御者观察到模型表现出其预期的行为，则可以合理地声明对模型的所有权。防御者可以轻松验证水印，而攻击者即使知道模型带有水印，也难以获取水印相关信息。

通过在模型参数中嵌入水印，模型所有者能够验证模型的所有权和完整性。然而，在模型中嵌入水印的过程会不可避免地影响模型性能，因为该操作本质上是将一个特殊的触发器插入输入中。因此，水印可以被看作是对模型的“毒害”。

3 问题陈述

3.1 模型水印的存在性原理

数字水印技术[34]利用了人类视觉系统对图像中细微变化的不敏感性。人类视觉系统对每个像素的微小变化通常不敏感，因此可以在不显著降低图像质量的前提下，嵌入水印信息。这种水印信息的嵌入过程通常涉及对图像特定频域或空间域的微小调整。数字水印技术因其可隐藏和提取信息，同时确保用户体验不受影响，在版权保护和数据完整性验证领域应用广泛。

模型水印技术则利用了机器学习模型参数所具有的冗余特性，即使是复杂的深度学习模型，其参数空间也存在一定程度的冗余性，允许在不显著降低模型性能的前提下进行参数调整。这种参数冗余性使得模型所有者可以在模型中嵌入特定水印信息，使水印成为模型结构的一部分。此技术应用于保护模型的知识产权，可以在不接触原始模型数据的情况下，有效验证模型的所有权。

数字水印技术和模型水印技术的共同之处在于，都利用了系统对冗余性的感知来实现信息隐藏和保护。数字水印技术利用人类视觉系统的感知冗余性来隐匿信息，而模型水印技术则利用模型参数的冗余性来维护知识产权。在实际应用场景中，这两种技术必须在信息嵌入的隐蔽性和系统功能的完整性之间取得平衡，以保证其有效性和实用价值。

3.2 模型水印的难点

尽管深度神经网络水印技术的进步使其成为保护模型知识产权的有效方法，但仍面临诸多挑战。首先，如何在不影响DNN模型性能的前提下成功嵌入水印是一个关键的问题。DNN模型通常具有复杂结构和大量参数，因此嵌入水印信息时需要谨慎，避免对模型的性能造成负面影响。水印嵌入不应显著降低模型的预测准确率或其他性能指标，以免削弱其在实际应用场景中的表现。

此外，水印的鲁棒性也是DNN水印技术面临的另一重大挑战。这里，水印的鲁棒性指其在面对模型微调、剪枝等常见模型优化手段时的抗干扰能力。模型微调和剪枝是常见的模型优化技术，用于适应新任务或减少模型的计算和存储需求。然而，这些模型优化技术可能意外地移除或破坏嵌入的水印信息，从而影响水印的完整性和有效性。同时，攻击者可能利用这些技术来破坏模型水印。

对于模型剪枝技术，人工智能模型从卷积层到全连接层通常包含大量冗余参数，其中许多神经元的激活值趋近于零。即便移除这些神经元，模型仍能保持相似的准确率。水印嵌入过程的本质是利用这些冗余参数来隐藏信息。然而，在模型剪枝过程中，如果嵌入水印的冗余参数被移除，将不可避免地影响水印的完整性。

对于模型微调技术，当针对特定任务的数据有限时，一种常见的方法是利用他人训练好的模型，并通过调整参数来适应新数据，此过程即为模型微调。攻击者可能利用模型微调技术，通过对原模型进行微调来获得其优异性能，同时新模型可能不再包含原模型的水印。简而言之，攻击者可以通过对模型进行微调来去除水印。

4 分区水印方案

鉴于水印容易受到模型优化和攻击的影响，建议在神经网络的不同区域采用差异化地添加水印的方法。在介绍具体方法之前，首先定义本文的威胁模型。模型所有者可能会通过微调、剪枝等操作来优化模型性能，而攻击者可能通过类似的微调、剪枝等手段试图去除水印。由于两者的行为在形式上相似，仅目的不同，因此在实验中统一以攻击者的行为作为研究对象进行分析。

攻击者的目标是获得一个不含水印的模型。首先，攻击者可能通过未经授权的复制或盗取模型来侵犯模型所有者的知识产权。此外，攻击者可能通过对模型进行剪枝或微调，改变其结构或参数，从而使原始水印难以被检测到，以规避水印检测机制。本文的目标是设计一种水印方案，在面临这些威胁时，能够有效识别和验证模型的合法性和完整性，同时尽量减少对模型性能的影响。

4.1 分区水印嵌入

在算法1中提出了一种新的水印策略，称为分区水印。该策略不再将水印嵌入到整个模型中，而是选择模型中的部分层进行水印嵌入，通常为神经网络的最后几层。然后通过逐层解冻的方式，阶梯式地将水印嵌入这些选定的层中。

算法1 分区水印嵌入算法

```
输入：Model，D，epochs，layers_to_turn
输出：Mw
parameters ← Model.parameters()
for each param in parameters do:
    param.requires_grad ← False/*冻结所有参数*/
end for
layers_to_turn ← Reverse(layers_to_turn)/*将列表逆序，以便之后的逐层解冻*/
number_layers ← len(layers_to_turn)
for epoch=0 to epochs−1 do
    current_stage ← epoch //(epochs // num_layers)
    /*根据当前阶段 epoch 对 layer_to_turn 列表中的神经元解冻*/
    for i=0 to current_stage do
        for name,pame in Model.parameters() do
            if layers_to_turn[i] in name do
                param.requires_grad ← True
            end if
        end for
    end for
    Reset optimizer
    Train the Model on the watermark dataset D
end for
```

神经网络的分区水印算法将水印集 D 嵌入到模型所有

者指定的特定层。其中，Model 为未添加水印的模型，Mw 为添加过水印的模型。最初，模型的所有层都被冻结，并通过 *layers_to_turn* 参数指定要微调的层，这些层按逆序排列，以便从指定的最后一层开始依次解冻。训练过程根据总迭代次数和需要微调的层数被分为多个阶段。在每次迭代中，首先确定当前的阶段，并解冻 *layers_to_turn* 中指定的层。优化器被重置以适应可训练参数的更新，然后在当前水印数据集上微调模型。在完成对一个水印数据集的训练迭代后，可以再次执行算法以嵌入多组水印集。这种方法能确保水印以受控的方式嵌入神经网络，从而最大限度地减少性能损失。

通过对部分层进行选择性嵌入水印，并采用阶梯式方法，该技术在很大程度上避免了水印对神经网络性能的负面影响，尤其解决了浅层和深层神经元分工不同所带来的问题。同时，由于攻击者难以获取水印嵌入的位置信息，同时优化了水印嵌入位置以及嵌入方式，该方法显著增强了水印对各种攻击的鲁棒性。

4.2 针对攻击的鲁棒性

攻击者可能试图通过模型微调或剪枝来去除水印，从而掩盖未经授权的模型复制或转发行为。本文模拟攻击场景，对添加两种不同水印的神经网络进行测试：一个采用分区水印策略，另一个使用基线方法在整个模型中嵌入水印。第 5 章将详细介绍这些实验，展示不同水印策略在面对攻击时的表现差异。

5 实验

5.1 实验环境搭建

在多个不同数据集上评估了水印方案的有效性，包括 MNIST，CIFAR10 和 CIFAR100。其中，MNIST 包含 7 万张手写数字灰度图像；CIFAR-10 包含 6 万张彩色图像，分为 10 个类别；CIFAR-100 同样包含 6 万张彩色图像，但细分为 100 个类别，并具有层次化标签结构。这 3 个数据集在图像复杂度、类别数量和任务难度上依次递增。

采用 ResNet18 架构，在 3 种不同数据集上训练模型，并分别嵌入基线水印和分区水印，以测试和比较两种方案在不同场景下水印存活率的表现。此外，模拟攻击场景，对使用不同水印方案的神经网络进行攻击测试，从而验证和评估其在面对模型剪枝、微调等攻击时的鲁棒性和有效性。

5.2 模型性能评估

选择 ResNet18 作为实验模型，是基于研究数据集的考虑。由于所使用的数据集规模适中，因此 ResNet18 的轻量化设计和较高的计算效率能够很好地满足实验需求。同时，其优秀的特征提取能力能够支持本文任务对图像特征的深层次挖掘。首先在 MNIST，CIFAR-10 和 CIFAR-100 这 3 个数据集上精心训练了受害者模型，并在模型中分别嵌入了基线水印和分区水印方案，结果如表 1 所列。由表中数据可知，分区水印方案的水印存活率与基线方案相近，但在模型预测准确率方面有显著提升。在许多关键应用场景，如自动驾驶和医疗诊断中，任何程度的模型性能损失都值得关注。因此，水印存活率略微降低是可以接受的，因为即使有轻微下降（实际最大差距也不到 3 个百分点），它仍远高于无水印模型的理论存活率（即误报率）。例如，在 MNIST 数据集中，其存在 10 个标签，导致在使用 MNIST 数据集训练的模型上验证水印时，存在 10%左右的误报率（CIFAR-10 为 10%，CAFAR-100 为 1%）。分区水印方案的水印存活率远高于理论误报率（在 MNIST 数据集上为 86.67%），这显示出该方案在维持模型性能方面的优势。

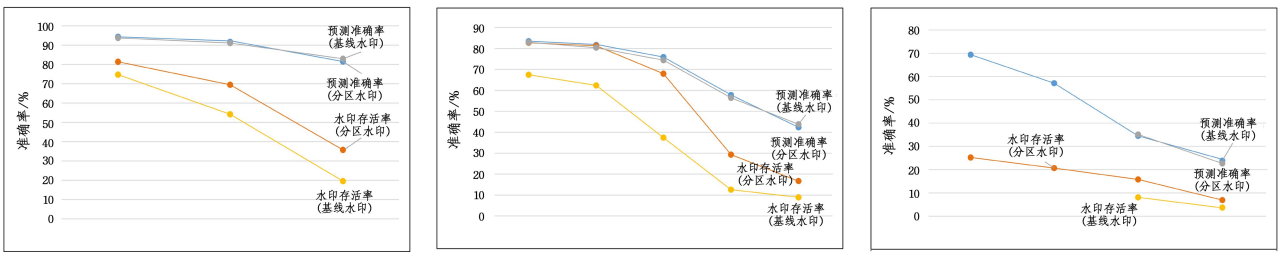
表 1 分区水印方案和基线水印方案在不同数据集上的表现

Table 1 Performance of the partitioned watermarking scheme and the baseline watermarking scheme on different datasets

数据集	预测准确率/%	嵌入水印	预测准确率/%	水印存活率/%
MNIST	99.59	分区水印	99.47	86.67
		基线水印	99.03	87.83
CIFAR10	87.48	分区水印	86.54	83.41
		基线水印	85.74	87.16
CIFAR100	74.34	分区水印	72.83	25.07
		基线水印	54.12	25.74

5.3 水印方案在模型微调下的鲁棒性评估

本节评估了分区水印方案与基线水印方案在模型微调后的鲁棒性。本文对两种水印方案在模型微调后的情况下进行了对比分析，旨在探讨在相似预测准确率条件下，分区水印方案与基线水印方案在水印存活率方面的差异，从而验证不同水印方案的鲁棒性。为此，训练了两组模型，一组使用分区水印方案，另一组使用广泛应用的基线水印方案。随后，对两组模型进行微调，并且保证微调后两种方案的预测准确率相似，模拟实际应用中的模型更新和优化，以及应对攻击者试图去除水印的场景。微调过程使用相同的数据集和参数，以确保实验结果的可比性和公正性，从而有效比较两种水印方案的性能。实验结果如图 2 所示，其中横轴表示实验中微调程度的递增（从左到右微调幅度逐渐增加，无具体物理单位）。



(a) MNIST (b) CIFAR-10 (c) CIFAR-100

图 2 两种水印方案在面对模型微调时的表现

Fig. 2 Performance of two watermarking schemes against model fine-tuning

实验结果表明，分区水印方案在模型微调后的水印存活率明显优于基线方案，展示出更强的抵抗微调的能力。图 2 显示，在 MNIST，CIFAR-10 和 CIFAR-100 数据集中，分区水印方案在微调后保持了更高的水印存活率，而基线方案的水印存活率下降过快，无法保证模型所有权声明。这些结果表明，分区水印方案在模型性能损失相同的条件下，能提供更稳健的水印保护。这种鲁棒性优势可能归因于分区方案对水印嵌入位置的优化，使其在模型参数变化时仍能有效地维持水印完整性，从而增强了对水印攻击的抵抗能力。

5.4 水印方案在模型剪枝下的鲁棒性评估

本节评估了分区水印方案在深度神经网络模型剪枝后的鲁棒性。对嵌入分区水印和基线水印的模型进行剪枝，修剪激活频率最低的神经元，模拟实际应用中的模型压缩和资源优化，以及应对攻击者试图通过剪枝去除水印的操作。剪枝实验在 MNIST 数据集上进行，以验证水印方案的鲁棒性。实验结果如图 3 所示。

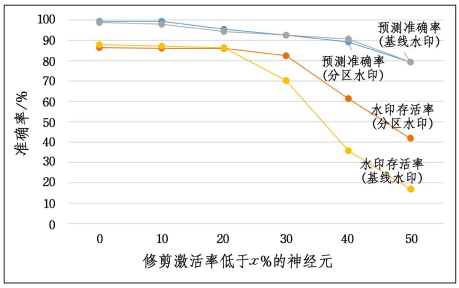


图 3 两种水印方案在面对剪枝时的表现

Fig. 3 Performance of two watermarking schemes against pruning

由图 3 的结果可知，分区水印方案在神经网络模型剪枝后仍保持了较高的水印存活率，展现出优异的鲁棒性。因此，当攻击者盗取模型并对激活频率最低的神经元进行剪枝时，尽管进行了大量修剪，水印存活率仍能保持在高水平。值得注意的是，只有当模型性能大幅下降时(降幅超过 10 个百分点)，水印存活率才会明显降低，但即使如此，仍然可以达到 60%以上，远高于理论水平(10%)，足以确保声明模型所有权。需要注意，水印存活率的下降远没有模型性能下降带来的影响恶劣，因为现实中任何一点模型性能的下降都可能带来严重的后果，而即使水印存活率下降，只要仍远超理论水印存活率就可以声明模型的所有权。因此，分区水印的这种鲁棒性优势可能归因于分区水印方案的嵌入位置和方式的优化设计，使得水印信息在模型参数精简时得到有效保留。

本文提出的分区水印方案与其他现有技术相比，在应对剪枝和模型微调方面具有明显的优势。与 EWE[35] 技术相比，当模型进行剪枝时，分区水印算法展现出更优异的性能。EWE 方案在模型性能几乎不受影响的情况下，水印存活率显著下降至约 40%。而分区水印方案表现出更强的抗干扰能力，即使在模型性能出现一定程度下降的情况下，水印存活率仍能保持在 60%左右。这说明攻击者若要通过剪枝去除水印，将导致模型性能受到较大影响。

此外，在模型微调的场景中，分区水印方案也优于现有的主流黑盒水印方法[36]。当模型微调导致性能略有下降时，该黑盒水印存活率下降了接近 5%；而分区水印方案则表现出更强的稳定性，在模型性能下降类似的情况下，水印存活率几乎不受影响，下降幅度不足 1%。因此，攻击者若试图通过微调抹除水印，同样需要付出更大的代价，导致模型性能损失更为显著。

结束语 本文提出了一种新的深度神经网络模型水印方法——分区水印。该方法选择性地在模型的某些特定层嵌入水印，并且在水印嵌入过程中采用分阶段的方法，逐层解冻嵌入水印的模型层，实现了水印存活率与模型性能之间的良好平衡。在 MNIST，CIFAR-10 和 CIFAR-100 数据集上的实验结果表明，与传统的基线水印方案相比，分区水印方案能够更好地保持模型性能，并在模型剪枝和微调攻击下展现出更强的鲁棒性，确保了水印的安全性和持久性。具体而言，分区水印在水印存活率基本相近的情况下，显著降低了对模型性能的影响，同时面对攻击时表现出优异的抗干扰能力，证明了该方案在实际应用中具有实用价值。

参 考 文 献

[1] BALCZEWSKI E A, CAO J, SINGH K. Risk prediction and machine learning: a case-based overview[J]. Clinical Journal of the American Society of Nephrology, 2023, 18(4): 524-526.

[2] NALISNICK E, SMYTH P, TRAN D. A brief tour of deep learning from a statistical perspective[J]. Annual Review of Statistics and Its Application, 2023, 10(1): 219-246.

[3] ZHANG H, SHAO H. Exploring the Latest Applications of OpenAI and ChatGPT: An In-Depth Survey[J]. CMES-Computer Modeling in Engineering & Sciences, 2024, 138(3): 2061-2102.

[4] XU P, JI X, LI M, et al. Small data machine learning in materials science[J]. NPJ Computational Materials, 2023, 9(1): 42.

[5] DAIDONE M, FERRANTELLI S, TUTTOLOMONDO A. Machine learning applications in stroke medicine: Advancements, challenges, and future prospectives[J]. Neural Regeneration Research, 2024, 19(4): 769-773.

[6] LAI Q, YANG L, HU G, et al. Constructing multiscroll memristive neural network with local activity memristor and application in image encryption[J]. IEEE Transactions on Cybernetics, 2024, 54(7): 4039-4048.

[7] GOLDBERG Y. A primer on neural network models for natural language processing[J]. Journal of Artificial Intelligence Research, 2016, 57: 345-420.

[8] MEHRISH A, MAJUMDER N, BHARADWAJ R, et al. A review of deep learning techniques for speech processing[J]. Information Fusion, 2023, 99: 101869.

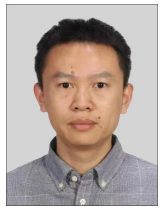
[9] CHIB P S, SINGH P. Recent Advancements in End-to-End Autonomous Driving Using Deep Learning: A Survey [J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 103-118.

[10] KIM J, KIM J, KIM H, et al. CNN-based network intrusion detection against denial-of-service attacks [J]. Electronics, 2020, 9(6): 916.

[11] LI Y, YAN H, HUANG T, et al. Model architecture level privacy leakage in neural networks[J]. Science China Information

Sciences，2024，67(3)：132101.
[12] AKHTAR N，MIAN A. Threat of adversarial attacks on deep learning in computer vision：A survey[J]. IEEE Access，2018，6：14410-14430.
[13] PENG S，CHEN Y，XU J，et al. Intellectual property protection of DNN models[J]. World Wide Web，2023，26(4)：1877-1911.
[14] OREKONDY T，SCHIELE B，FRITZ M. Knockoff Nets：Stealing functionality of black-box models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：4954-4963.
[15] WU H，ZHANG J，LI Y，et al. Overview of artificial intelligence model watermarking[J]. Journal of Image Graphics，2023，28(6)：1792-1810.
[16] KAHNG A B，LACH J，MANGIONE-SMITH W H，et al. Watermarking techniques for intellectual property protection[C]// Proceedings of the 35th Annual Design Automation Conference. 1998：776-781.
[17] KUMAR J，KUMAR M. Comparison of image compression methods on various images[C]// 2015 International Conference on Advances in Computer Engineering and Applications. IEEE，2015：114-118.
[18] HE Y，XIAO L. Structured pruning for deep convolutional neural networks：A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2023，46(5)：2900-2919.
[19] CHURCH K W，CHEN Z，MA Y. Emerging trends：A gentle introduction to fine-tuning[J]. Natural Language Engineering，2021，27(6)：763-778.
[20] 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.
[21] ADI Y，BAUM C，CISSE M，et al. Turning your weakness into a strength：Watermarking deep neural networks by backdooring[C]// 27th USENIX Security Symposium(USENIX Security 18). 2018：1615-1631.
[22] LEE S，SONG W，JANA S，et al. Evaluating the robustness of trigger set-based watermarks embedded in deep neural networks[J]. IEEE Transactions on Dependable and Secure Computing，2022，20(4)：3434-3448.
[23] YOSINSKI J，CLUNE J，BENGIO Y，et al. How transferable are features in deep neural networks? [C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. 2014：3320-3328.
[24] LECUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[C]// Proceedings of the IEEE. 1998：2278-2324.
[25] KRIZHEVSKY A，HINTON G. Learning multiple layers of features from tiny images[EB/OL]. https：//www. cs. utoronto. ca/~kriz/learning-features-2009-TR. pdf.
[26] LIAN D，ZHOU D，FENG J，et al. Scaling & shifting your features：A new baseline for efficient model tuning[J]. Advances in Neural Information ProcessingSystems，2022，35：109-123.
[27] ZHANG Y，WU H，LIN F，et al. Deep learning model pruning technology in image recognition[J]. Journal of Nanjing University of Science and Technology，2023，47：699-707.
[28] FAN L，NG K W，CHAN C S. Rethinking deep neural network ownership verification：Embedding passports to defeat ambiguity attacks[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019：4714-4723.
[29] CHEN H，ROUHANI B D，KOUSHANFAR F. Blackmarks：Blackbox multibit watermarking for deep neural networks[J]. arXiv：1904. 00344，2019.
[30] LYU P，MA H，CHEN K，et al. MEA-Defender：A Robust Watermark against Model Extraction Attack[J]. arXiv：2401. 15239，2024.
[31] LYU P，LI P，ZHU S，et al. Ssl-wm：A black-box watermarking approach for encoders pre-trained by self-supervised learning[J]. arXiv：2209. 03563，2022.
[32] LIU H，WU Y H，LI X D，et al. Deep neural network model copyright protection framework based on external samples[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition)，2025，37(3)：405-416.
[33] PENG W P，LIU J B，PING Y，et al. Model protection scheme for fusion of internal and external feature watermarks[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition)，2024，36(4)：765-774.
[34] PODILCHUK C I，DELP E J. Digital watermarking：algorithms and applications[J]. IEEE Signal Processing Magazine，2001，18(4)：33-46.
[35] JIA H，CHOQUETTE-CHOO C A，CHANDRASEKARAN V，et al. Entangled watermarks as a defense against model extraction[C]// 30th USENIX Security Symposium(USENIX Security 21). 2021：1937-1954.
[36] ZHANG J，GU Z，JANG J，et al. Protecting intellectual property of deep neural networks with watermarking[C]// Proceedings of the 2018 on Asia Conference on Computer and Communications Security. 2018：159-172.

LYU Zhenghao，born in 1999，postgraduate. His main research interests include AI robustness and intellectual property protection.



XIAN Hequn，born in 1979，Ph.D，professor，master's supervisor. His main research interests include cryptography and network and information systems security.

（责任编辑：何杨）