DOI: 10.3969/j.issn.0253-9608.2022.04.002

深度模型水印

张新鹏†，吴汉舟

上海大学 通信与信息工程学院，上海 200444

摘要 深度神经网络模型凝结了设计者的智慧，需要消耗大量数据和计算资源，是人工智能技术的重要产出物，已被广泛应用于生产和生活当中。然而，作为一种数字产品，如何保护深度神经网络模型免于被非法复制、分发或滥用(即知识产权保护)是人工智能产业化进程中必须面临和解决的难题。文章主要介绍基于数字水印的深度模型产权保护技术，通过总结深度模型水印的发展现状，对深度模型水印的研究趋势进行展望。

关键词 深度模型；数字水印；产权保护；人工智能安全

2021年3月，新华社发布《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》，提出了以人工智能为代表的新型基础设施建设政策，标志着人工智能发展进入技术持续创新和应用广泛深化的新阶段。以美国为代表的科技大国也将发展新一代人工智能上升为国家战略高度，使人工智能成为新一轮技术和产业变革的核心驱动力。可以说，人工智能正在重塑生产方式，优化产业结构，提升生产效率，赋能千行百业，推动经济社会向着智能化方向加速跃升。不难预见，人工智能将在服务人类社会生产、生活等各方面发挥越来越重要的作用，是数字经济时代的“新电能”。

作为实现人工智能的代表性技术，人工神经网络(简称神经网络)是一种模仿生物神经网络(中枢神经系统，尤其是大脑)结构和功能的数学模型(或计算模型)，用于对复杂函数的估计和近似。如图1所示，神经网络由有限多个神经元联结起来进行计算，而每个神经元的功能是计算加权向量经非线性映射后的结果。神经网络是一个能够学习和归纳的系统，即从已知数据中挖掘规律，并对未知数据进行可靠的分析和预测。这个过程可以划分为两个阶段：训练阶段和测试阶段。前者利用已知数据来确定神经网络中的待定参数，一旦确定了参数就意味着得到了“训练好”的神经网络模型。后者利用“训练好”的神经网络模型对未知数据进行分析和预测。可以形象地将神经网络看作是一名学生，为了能够在期末考试(未知数据)中取得优异成绩，学生要在平时的训练(已知数据)中不断地学习知识和总结经验。

自从2012年以神经网络为架构的深度表征学习算法[1]夺得ImageNet国际计算机视觉大赛[2]冠军以来，(深度)神经网络得到了学术界和工业界的广泛关注和深入研究，并在诸多应用领域取得成功，包括人脸识别、自动驾驶、语音识别和自然语言处理等。神经网络模型不仅可以部署在本地提供给个人使用，也可以部署在云端以提供公共服务。然而，作为一种数字产品，神经网络模型不仅凝结了设计者的智慧，还需要消耗大量

†通信作者，研究方向：多媒体信息安全、信息隐藏、数字取证、加密域信号处理、图像处理。

E-mail: xzhang@shu.edu.cn

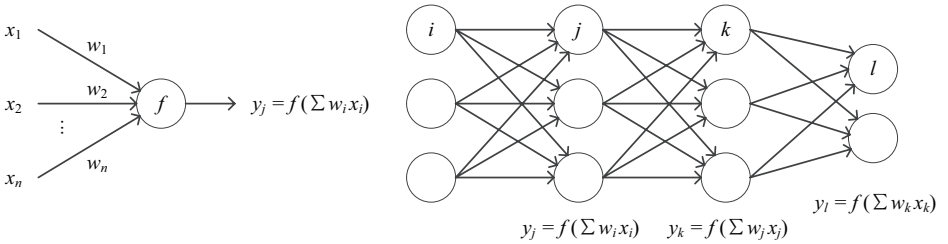


图1 神经网络结构示意图：单个神经元(左)；多层神经网络(右)

的训练数据和计算资源。例如，为了精准地识别人脸，我们需要提供几千万乃至数亿幅人脸图像给神经网络学习和归纳，运算耗时可能多达数月之久[3]。因此，构建训练有素的神经网络模型需要付出巨大的代价，这使得如何保护神经网络模型的知识产权不受侵害变得十分重要。

1 深度模型水印

目前，学术界主要运用数字水印保护深度神经网络模型的知识产权，简称为“深度神经网络模型水印”或“深度模型水印”。如图2所示，数字水印[4]是一种将特定信息(又称为“水印”)隐藏在数字信号中、不影响信号使用价值的安全技术，信号可以是图像、视频和音频等任意数字产品。隐藏操作通过修改信号的内容来实现，若拷贝含有水印的信号，水印也会一并被拷贝。含有水印的信号可能会受到攻击，当水印提取者接收到可能被攻击的含水印信号时，他将从信号中重构水印以实现版权鉴定、完整性验证或叛徒追踪等目的。例如，某公司通过内网向员工发送重要文档前，可向文档中嵌入关联员工身份的水印，使得每份文档虽具有相同的内容，却承载了不同的水印。一旦文档泄露到外网，通过在泄露文档中提取水印，可以追踪到泄露源。显然，我们可以向深度神经网络模型中嵌入水印以保护其知识产权。

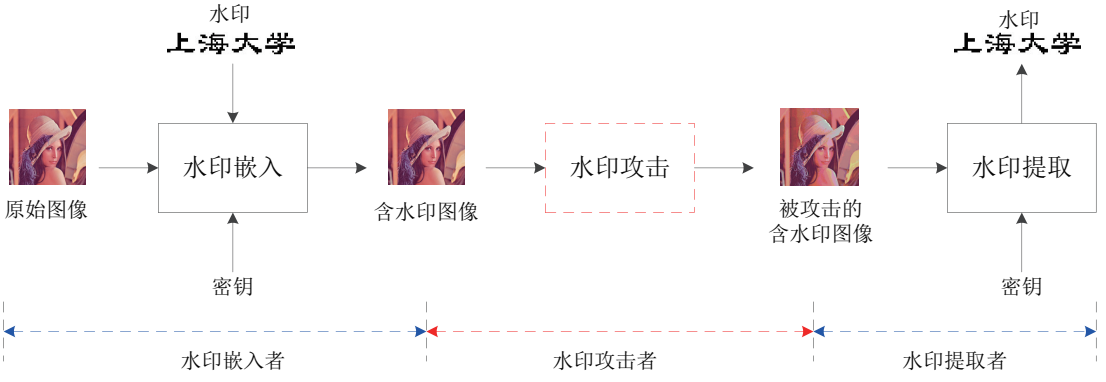
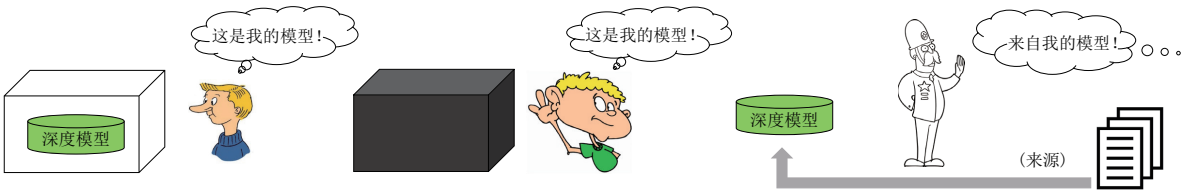


图2 数字水印的基本框架示意图(这里以数字图像为例)

然而，不同于图像和视频等常见的多媒体数据，深度模型需要完成特定任务。简单地将适用于多媒体数据的水印技术用于深度模型会降低模型在特定任务上的性能，损害使用价值，甚至会带来安全威胁。例如，深度模型已被用于自动驾驶和医疗辅助诊断等领域，若嵌入水印后深度模型的决策错误率非常高，不仅无益于保护模型的知识产权，还会危害人身安全。因此，深度模型水印要确保水印嵌入操作不会损害模型在特定任务上的性能，也即任务保真度高。借鉴多媒体水印的评价指标，深度模型水印还需考虑水印的嵌入量(即嵌入的信息量)、水印的保真度(即重构的水印质量)、水印的安全性(即抵抗攻击者检测或重构水印的能力)和水印的稳健性(即抵抗攻击者移除水印的能力)。

如图3所示，依据水印提取者是否掌握模型的细节和能否与模型进行交互，可将现有方法分为三类：“白盒”水印、“黑盒”水印和“无

盒”水印。“白盒”水印假定水印提取者知悉模型的内部细节(如网络结构和参数等)；“黑盒”水印假定水印提取者无法获取模型的内部细节，但能通过与模型进行交互，获得模型在特定样本上的预测结果；“无盒”水印假定水印提取者既不知晓模型的内部细节，也不能与模型进行直接交互，但能从模型生成的任意样本中重构水印，实现产权保护。接下来，我们分别介绍三种类型的水印技术。



(a)“白盒”水印　(b)“黑盒”水印　(c)“无盒”水印

图3 深度模型水印技术分类[3]

2 “白盒”水印

如图4所示，深度模型可视为一个带有参数的有向图，其中节点由神经元组成，有向边对应于神经元之间的连接。“白盒”场景假定水印提取者知悉深度模型的细节，因此，实现“白盒”水印的有效方式是修改有向图(即深度模型)的参数或结构。如前所述，直接修改参数或拓扑结构，会降低深度模型在原始任务上的性能。因此，修改有向图的参数或结构要求我们设计有效机制保持深度模型在原始任务上的性能。

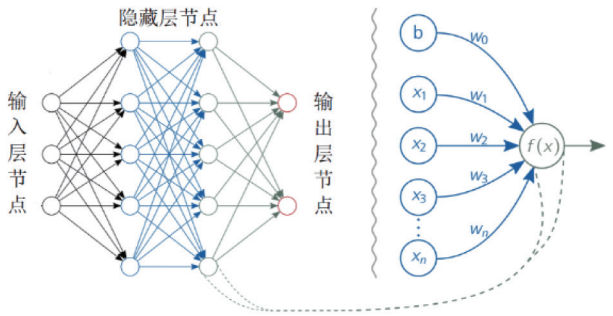


图4 深度模型的图表示实例

以修改参数为例，我们可以利用正则化实现深度模型水印技术[5]，其思想是让水印在模型训练的过程中嵌入到模型参数中。考虑到模型在训练的过程中，部分参数会快速收敛，通过修改这些收敛参数不仅不会损害模型在原始任务上的性能，还能够承载额外的水印信息[6]。除了直接修改参数，我们还可以将水印嵌在参数的低阶统计量(如概率密度函数)[7-9]以提升抗攻击能力，其他方法还包括对抗训练[10]、抖动调制[11]、梯度优化[12]和“护照层”[13]等。从本质上看，这些方法都是在保证模型计算精度的条件下，提升水印的稳健性或隐蔽性。

同修改模型参数相比，调整网络结构可以抵御参数攻击，但要解决两个问题，分别是：如何建立模型结构与水印之间的关系；如何保障深度模型在结构发生变化后的计算性能。针对这两个问题，研究人员提出了利用通道剪枝技术对深度模型中神经元之间的连接进行调整以实现水印嵌入[14]，其本质是删除深度模型图中不重要的边，利用边的数量来承载水印，当模型结构被调整后(也即嵌入水印后)，通过对模型参数继续优化(即微调)，可恢复深度模型在原始任务上的性能。图5给出了基于剪枝的结构水印嵌入框架示意图。如前所述，该框架通过修改模型结构而非参数实现水印嵌入，能够抵御所有参数攻击。

上述方法侧重水印的稳健性，在实际应用过程中，深度模型可能被篡改。为了应对这一问题，研究人员提出了适用的“脆弱”水印技术[15-17]，用于验证深度模型的完整性。“脆弱”是指对模型的轻微修改必将导致水印难以完美重构，这种“不完美”可用于模型的完整性验证。以文献[15]为例，研究人员对模型参数进行小波变换，将秘密信息及其哈希值嵌在不重要的小波系数上。在模型验证时，验证者首先提取出秘密信息和哈希值，然后计算所

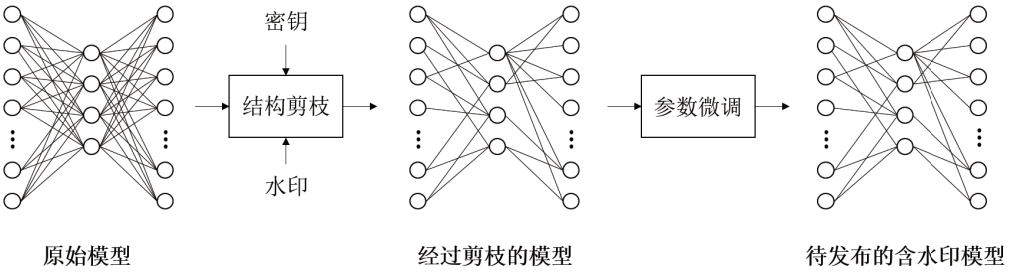


图5 基于剪枝的结构水印嵌入框架示意图

提取出的秘密信息的哈希值，最后比较所计算的哈希值和所提取出的哈希值，若值相等，则认为模型是完整的，否则，视之被篡改。图6给出了该方法所对应的一般性框架。

3 “黑盒”水印

“黑盒”场景假定水印提取者不能访问目标模型的内部细节，但能通过某种方式获得目标模型在特殊数据集(又称触发集)上的输出结果，通过对这些输出结果进行一致性分析，可以鉴定模型的产权。由于“黑盒”场景在应用环境中较“白盒”场景更为常见(例如，水印提取者可以与部署在云端的深度模型进行交互，但无法获取模型的内部细节)，故“黑盒”水印相对更为实用。以图像分类为例，如图7所示，“黑盒”水印可以描述如下：水印嵌入者利用正常图像和触发图像训练深度模型，训练好的模型视为含水印，可投入使用；在验证阶段，水印提取者通过获取目标模型在触发图像的预测结果，并与预先指定的标签进行一致性分析，可以验证产权。

“黑盒”水印借助深度模型未利用的泛化能力，使深度模型既能从正常数据集中学习知识以完成原始任务，又能“记住”触发样本和对应标签的映射关系。当目标模型出现产权纠纷时，通过重构这种映射关系，我们可以确定目标模型的产权。

在“黑盒”框架下，如何构建触发样本并标注类别是重要的科学问题。图7所示实例所采取的方法是在正常图像上直接添加黄色方块来构造触发图像。现有构建触发样本的方法可以分为两类：选用与模型无关的样本；选用与模型相关的样本。以图像分类为例，前者构建触发样本的主要手段包括：选用与模型任务无关的抽象图像[18]、选用无关数据集中的图像[19]和选用随机噪声图像[20]等。对于后者，典型方法包括：向正常样本添加特殊标识(如文字、标识、噪声等)[20-21]、向正常样本添加轻微的扰动形成对抗样本[22-23]等。我们可以随机选择某个类别作为触发样本的标签，也可以为触发样本分配某个特定的类别。为了实现“黑盒”认证，水印提取

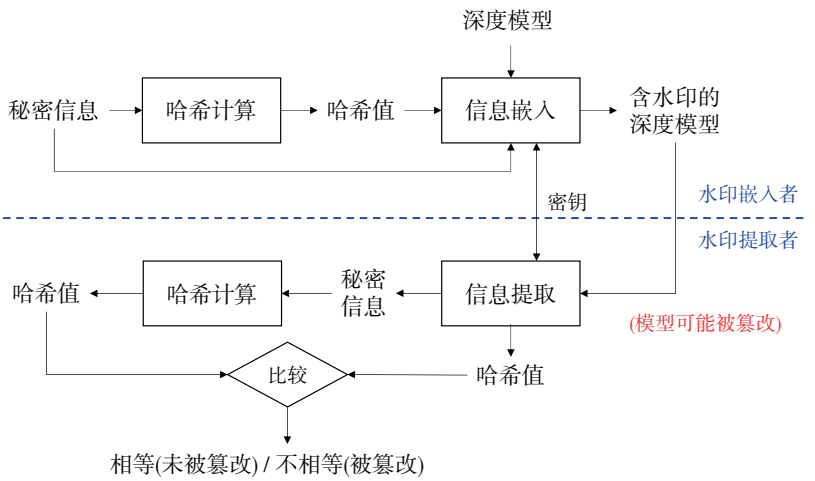


图6 基于哈希验证的深度模型脆弱水印框架示意图

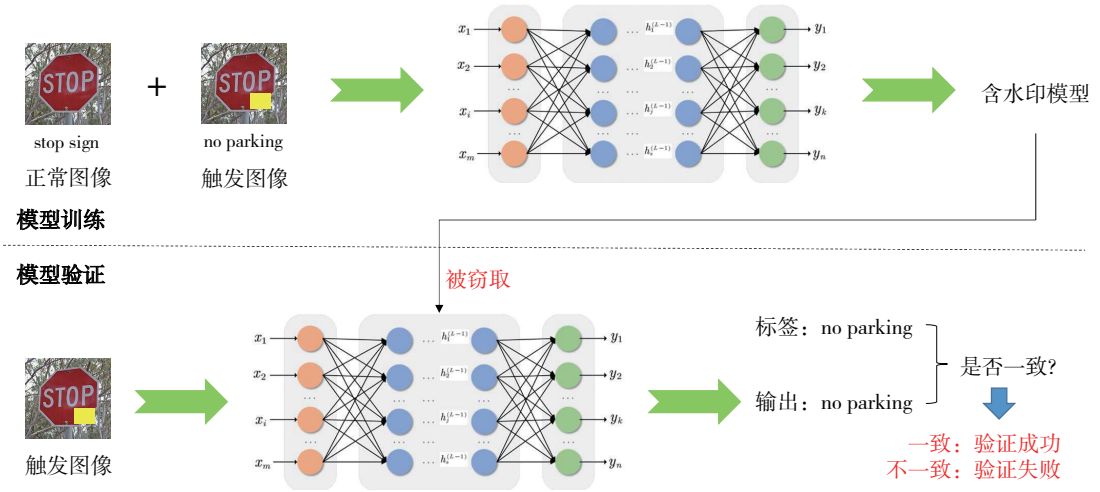


图7 “黑盒”水印的基本框架示意图(这里以图像分类为例)

者需要构建一组触发样本。由于水印嵌入者和水印提取者之间存在联盟关系，触发样本的构造方式可以共享，即：提取者可以使用与嵌入者完全相同的触发样本，也可以使用与嵌入者完全相同的方法生成新的触发样本。主流方法基本采用这种方式。

就本质而言，上述方法通过让触发样本远离深度模型的高维决策边界以实现稳健的“黑盒”认证。倘若触发样本非常靠近决策边界，那么对深度模型的轻微扰动大概率会使触发样本跨过决策边界，做出错误的决策，利用这一特性，我们可以实现“黑盒”脆弱水印[24-25]。

4 “无盒”水印

除了“白盒”水印和“黑盒”水印外，研究人员还提出了“无盒”水印[26]。同“白盒”水印相比，“无盒”水印不要求提取者掌握目标模型的内部细节。同“黑盒”水印相比，“无盒”水印不要求提取者与目标模型进行直接交互。因此，相对于“白盒/黑盒”水印，“无盒”水印中提取者掌握的信息更少，故具有更好的应用前景。“无盒”水印主要面向具有生成任务的深度模型。以文献[26]为例，研究人员提出了一种适用于云端服务场景、面向图像生成模型的“无盒”水印算法。如图8所示，该算法联合了两个神经网络(即受保护的网络和水印提取网络)，通过在模型训练的过程中同时优化两个网络的参数，使得受保护的网络在完成训练后不仅可以完成原始任务(示例中是图像彩色化)，而且允许验证者利用密钥从输出的图像中检测出水印，实现图像和模型的双重产权保护。

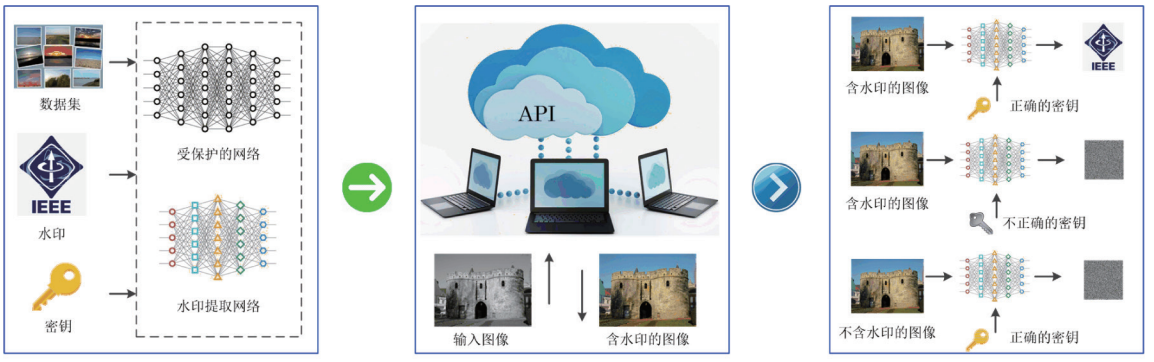


图8 “无盒”水印应用场景示例

5 总结与展望

人工智能模型作为一种数字产品容易被复制、调整和篡改，在人工智能技术迅速发展的同时，保护其知识产权具有显著学术价值和产业需求。本文围绕“白盒”“黑盒”和“无盒”三个不同的场景介绍了深度模型水印技术。毫无疑问，现有研究成果为保护深度模型的知识产权提供了宝贵的思路，通过对深度模型标识所有者、使用者、版本号、传播路径并进行篡改检测，能够为人工智能的发展和应用提供必不可少的良好环境。

然而，深度模型水印研究刚刚起步，基础理论与关键方法中还蕴含很多科学问题，极具研究价值。一方面，现有研究成果侧重方案设计，很少关注理论研究，我们亟需研究和发展面向深度模型水印的基础理论，助力深度模型水印理论体系的构建。另一方面，深度模型水印在“攻”与“防”中发展，在不损害使用价值的条件下，对深度模型水印的攻击一般是针对水印的稳健性提出的，稳健性好的深度模型水印技术应能抵御多种攻击。现有“白盒”水印算法主要通过修改模型参数或模型结构来嵌入水印，当模型经过重训练或结构调整时，水印容易被清除，威胁产权保护；现有“黑盒/无盒”水印算法很少考虑样本攻击(即干扰输入样本)、蒸馏攻击(即构造替代模型)和集成攻击(即融合输出结果)等常见的攻击行为，使算法的实用性受限。因此，深度模型水印还有待学术界开展深入研究和探索，在理论基础和对抗攻击等方面取得新突破。

(2022年4月27日收稿)

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25(9): 1097-1105.

[2] ImageNet. ImageNet large scale visual recognition challenge (ILSVRC) [EB/OL]. [2022-04-27]. https://www.image-net.org/challenges/LSVRC/.

[3] 钱振兴, 张卫民, 李晓龙. 多媒体与人工智能安全研究极简综述[M]. 上海: 复旦大学出版社, 2021.

[4] 杨义先, 钮心忻. 数字水印理论与技术[M]. 北京: 高等教育出版社, 2006.

[5] UCHIDA Y, NAGAI Y, SAKAZAWA S, et al. Embedding watermarks into deep neural networks[C]//ACM International Conference on Multimedia Retrieval. New York: Association for Computing Machinery, 2017: 269-277.

[6] WANG J, WU H, ZHANG X, et al. Watermarking in deep neural networks via error back-propagation [J]. IS&T Electronic Imaging, Media Watermarking, Security and Forensics, 2020(4): 22-1-22-9.

[7] CHEN H, ROHANI B D, FU C, et al. DeepMarks: A secure fingerprinting framework for digital rights management of deep learning models [C]// SADDIK A E, BIMBO A D, ZHANG Z, et al. ICMR '19: Proceedings of the 2019 on International Conference on Multimedia Retrieval. New York: Association for Computing Machinery, 2019: 105-113.

[8] CHEN H, FU C, ROUHANI B D, et al. DeepAttest: an end-to-end attestation framework for deep neural networks [C]//ACM/IEEE 46th Annual International Symposium on Computer Architecture, 2019: 487-498.

[9] ROUHANI B D, CHEN H, KOUSHANFAR F. DeepSigns: a generic watermarking framework for IP protection of deep learning models [EB/OL]. (2018-05-31)[2022-04-27]. https://arxiv.org/abs/1804.00750.

[10] WANG T, KERSCHBAUM F. RIGA: covert and robust white-box watermarking of deep neural networks [C]// LESKOVEC J, GROBELNIK M, NAJORK M, et al. WWW '21: Proceedings of the Web Conference. New York: Association for Computing Machinery, 2021: 993-1004.

[11] LI Y, TONDI B, BARNI M. Spread-transform dither modulation watermarking of deep neural network [J]. Journal of Information Security and Applications, 2021, 63: 103004.

[12] TARTAGLIONE E, GRANGETTO M, CAVAGNINO D, et al. Delving in the loss landscape to embed robust watermarks into neural networks [C]//IEEE International Conference on Pattern Recognition, 2021: 1243-1250.

[13] FAN L, NG K W, CHAN C S. Rethinking deep neural network ownership verification: embedding passports to defeat ambiguity attacks [C/OL]// WALLACH H, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems 32 (NeurIPS 2019). [2022-04-27]. https://proceedings.neurips.cc/paper/2019.

[14] ZHAO X, YAO Y, WU H, et al. Structural watermarking to deep neural networks via network channel pruning[C]// 2021 IEEE International Workshop on Information Forensics and Security (WIFS), 2021: 1-6. DOI: 10.1109/WIFS53200.2021.9648376.

[15] GUAN X, FENG H, ZHANG W, et al. Reversible watermarking in

deep convolutional neural networks for integrity authentication [C]// CHEN C W, CUCCHIARA R, HUA X S, et al. MM '20: Proceedings of the 28th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2020: 2273-2280.

[16] ABUADBBA A, KIM H, NEPAL S. DeepiSign: invisible fragile watermark to protect the integrity and authenticity of CNN [C]// HUNG C-C, HONG J, BECHINI A, et al. SAC '21: Proceedings of the 36th Annual ACM Symposium on Applied Computing. New York: Association for Computing Machinery, 2021: 952-959.

[17] BOTTA M, CAVAGNINO D, ESPOSITO R. NeuNAC: a novel fragile watermarking algorithm for integrity protection of neural networks [J]. Information Sciences, 2021, 576: 228-241.

[18] ADI Y, BAUN C, CISSE M, et al. Turning your weakness into a strength: watermarking deep neural networks by backdooring [C]// ENCK W, FELT A P. SEC'18: Proceedings of the 27th USENIX Conference on Security Symposium. Berkeley: USENIX Association, 2018: 1615-1631.

[19] NAMBA R, SAKUMA J. Robust watermarking of neural network with exponential weighting [C]// GALBRAITH S, RUSSELLO G, SUSILO W, et al. Asia CCS '19: Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security. New York: Association for Computing Machinery, 2019: 228-240.

[20] ZHANG J, GU Z, JANG J, et al. Protecting intellectual property of deep neural networks with watermarking [C]// KIM J, AHN G-J, KIM S, et al. ACM Asia Conference on Computer and Communications Security. New York: Association for Computing Machinery, 2018: 159-172.

[21] LI M, ZHONG Q, ZHANG L Y, et al. Protecting the intellectual property of deep neural networks with watermarking: the frequency domain approach [C]// TrustCom 2020: Proceedings of IEEE's 19th International Conference on Trust, Security and Privacy in Computing and Communications. Los Alamitos, Calif: IEEE Computer Society, 2020: 402-409.

[22] MERRER E L, PEREZ P, TRÉDAn G. Adversarial frontier stitching for remote neural network watermarking [J]. Neural Computing and Applications, 2020, 32: 9233-9244.

[23] CHEN H, ROUHANI B D, KOUSHANFAR F. BlackMarks: blackbox multibit watermarking for deep neural networks [EB/OL]. (2019-03-31)[2022-04-27]. https://doi.org/10.48550/arXiv.1904.00344.

[24] HE Z, ZHANG T, LEE R B. VerIDeep: Verifying integrity of deep neural networks through sensitive-sample fingerprinting [EB/OL]. (2018-08-09)[2022-04-27]. https://doi.org/10.48550/arXiv.1808.03277.

[25] ZHU R, WEI P, LI S, et al. Fragile neural network watermarking with trigger image set [M]// QIU H, ZHANG C, FEI Z, et al. KSEM 2021: Knowledge Science, Engineering and Management. Cham: Springer, 2021: 280-293.

[26] WU H, LIU G, YAO Y, et al. Watermarking neural networks with watermarked images [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(7): 2591-2601.

Deep model watermarking

ZHANG Xinpeng, WU Hanzhou

School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

Abstract Deep neural networks (DNNs) condense the wisdom of the designer and consume a lot of data and computing resources. It is an important artificial intelligence technology, and is widely applied in our daily life. However, as a digital asset, how to protect DNN models from being illegally copied, distributed or abused (that is, intellectual property protection) is a difficult problem that must be faced and solved in the process of artificial intelligence industrialization. This article reviews digital watermarking techniques for intellectual property protection of DNN models. By summarizing the development status of deep model watermarking, the research trend of deep model watermarking is further prospected.

Key words deep model, digital watermarking, copyright protection, artificial intelligence security

（编辑：段艳芳）